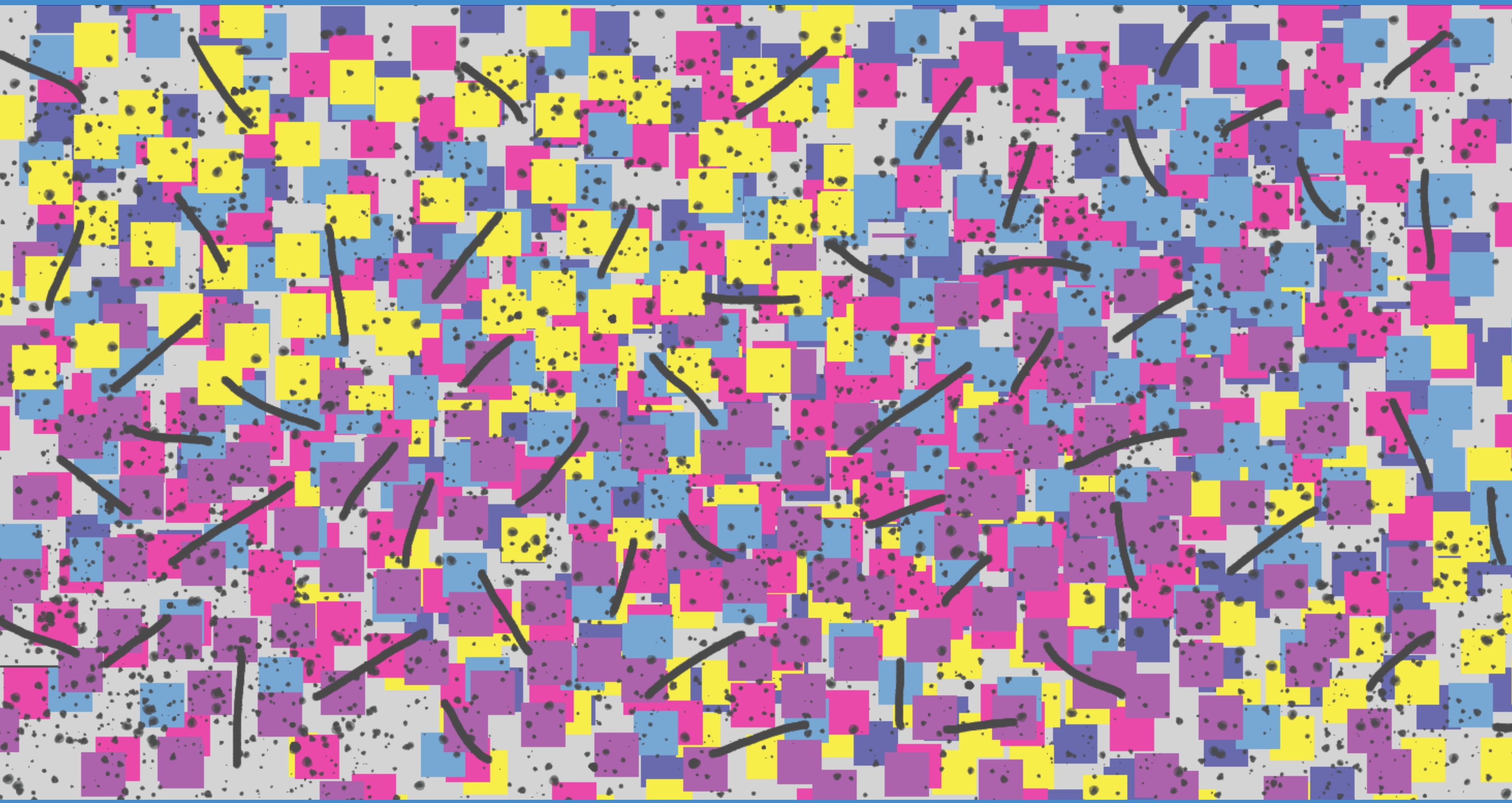
Selecting an Instrument for Measuring Infant and Toddler Language and Preliteracy

Shira Mattera
Jane Choi



mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

October 2019

Selecting an Instrument for Measuring Infant and Toddler Language and Preliteracy

Shira Mattera
Jane Choi

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

OCTOBER 2019

FUNDERS

This report was funded by the William Penn Foundation. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the William Penn Foundation.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Arnold Ventures, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, and Sandler Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

For information about MDRC and copies of our publications,
see our website: www.mdrc.org.

ABSTRACT

Language and preliteracy skills from birth to age 3 are important predictors of the reading gap that may emerge in the elementary school years and beyond. This research brief surveys instruments for assessing infants and toddlers' language and preliteracy skills to inform selection of appropriate measures for different uses by practitioners, researchers, program evaluators, funders, and policymakers. The scan draws from existing reviews and compendia of infant and toddler measures in the language and literacy domains. It evaluates them using a set of logistical and psychometric criteria informed by discussions with infant-toddler experts and practitioners for three purposes: identifying delays, measuring individual differences and change, and informing teaching and learning.

CONTENTS

ABSTRACT iii

LIST OF EXHIBITS vii

ACKNOWLEDGMENTS ix

REPORT

Introduction 1

Methodology 1

Uses of Measurement 2

Identifying Delays 2

Measuring Individual Differences and Change 4

Informing Teaching and Learning 5

Recommended Measures 6

Conclusion 9

APPENDIX

A Methodology 11

B Subset of 40 Measures that Meet Scan Criteria, Based on Availability, Domain, and Age Range 17

C Detailed Review of Selected Measures 25

D Additional Domains and Constructs Predictive of Kindergarten and Third-Grade Outcomes 39

REFERENCES 43

ABOUT MDRC 55

LIST OF EXHIBITS

TABLE

1	Review of Top Measures by Logistical and Psychometric Criteria	7
B.1	Subset of 40 Measures, by Primary Use, Domain, and Other Criteria	20

BOX

1	Considerations for Measurement Selection	3

ACKNOWLEDGMENTS

We thank the early childhood experts and providers who were interviewed and provided valuable perspectives that shaped this research brief and the recommendations that emerged.

At MDRC, Sharon Huang, Michelle Maier, Meghan McCormick, and Ximena Portilla supplied expertise on early childhood measurement and gave ongoing advice and comments on the selected criteria and measures' psychometric properties. JoAnn Hsueh generously provided feedback at various stages of the scan and writing process.

Margaret Bald, Jill Kirschenbaum, Ann Kottner, and Ali Tufel in MDRC's Publications department advised the team on the project timeline and carefully revised, edited, and formatted this research brief.

The Authors

INTRODUCTION

Third-grade students with poor reading skills are likely to remain poor readers throughout the rest of their education and are more likely than their reading-proficient peers to drop out of high school (The Annie E. Casey Foundation, 2010; Miles & Stipek, 2006; National Research Council, 1998). Skills measured in infants and toddlers from birth to age 3, although less stable due to rapid development, are considered to be the critical starting point for the disparity in reading ability and other academic outcomes that unfold later in life (Bornstein et al., 2014; Fenson et al., 2000; National Research Council, 1998; Tamis-LeMonda et al., 2006). This research brief surveys existing infant-toddler assessments and screeners to inform selection of appropriate measures in the areas of language and preliteracy for different uses by practitioners, researchers, program evaluators, funders, and policymakers. The review assesses the measures on a set of logistical and psychometric criteria relevant for three purposes: identifying delays; measuring individual differences and change; and informing teaching and learning. Final recommendations for promising infant-toddler measures of language and preliteracy outcomes build on:

- A review of existing research-based and state-derived measures
- Conversations with experts on state policies and measurement
- A detailed evaluation of a small number of measures across a range of logistical and psychometric criteria
- Discussions with practitioners about the strengths and weaknesses of selected measures

METHODOLOGY

MDRC employed a multistep process for a targeted review of infant-toddler measures, starting with a scan of outcome measures in infancy and toddlerhood and interviews with policy and measurement experts. Next, a subset of 40 measures was identified for more systematic review, prioritizing those that were: (1) publicly available for use; (2) focused on the areas of language and preliteracy; and (3) were validated for use with infants and toddlers from birth to age 3. Please see Appendix B for a review of this subset of measures.

Interviews were conducted with policymakers, practitioners, researchers, and program evaluators to better understand their needs for early childhood assessment as key stakeholders. Based on those interviews, each measure in the subset was categorized by its suitability for three distinct uses: (1) identifying delay; (2) measuring individual differences in skills and identifying changes in those skills across a year; and (3) informing teaching and learning.

Within each category, measures were prioritized that were: (1) validated and available in additional languages for use with diverse populations; (2) considered logistically feasible because they were short, did not require a highly trained assessor, and were identified by interviewees as already in wide use by infant-toddler service providers; and (3) had adequate reliability and validity. Recommended instruments were identified in each category after applying additional criteria that emerged from the interviews about needs for assessments in that category. See Appendix A for more details about methodology.

USES OF MEASUREMENT

Measuring young children's skills in a meaningful and consistent way can present unique challenges. Box 1 describes issues relevant to measuring infant-toddler outcomes. Given how complex it is to assess young children's abilities reliably and validly, it is important to identify the purpose for measurement when selecting an assessment tool. Three categories of how infant-toddler assessments are primarily used were identified through the stakeholder interviews: identifying delays in children, measuring individual differences in skills and identifying change in those skills over a year; and informing curriculum and professional development. Measures may be used for different intended purposes and may come with different logistical constraints. For example, measures that are used to evaluate programs or for making policy may be held to a high standard for reliably and validly assessing changes in children's skills, even if this makes the measure more logistically challenging and resource-intensive to collect, because evidence from these measures can contribute to high-stakes decision-making. When selecting measures that inform teachers' day-to-day instructional adaptations, stakeholders may lean toward those measures that can be easily administered multiple times during the year to guide short-term instructional decisions, leading to shorter and less detailed measures of children's abilities. The rest of the section briefly discusses applicable considerations for each category.

Identifying Delays

Pediatricians, teachers, and caregivers are often required by state or local policies to screen children for delay. A screener is a diagnostic tool that tends to be used to identify children at risk for a delay or in need of additional intervention or support (National Center for Systemic Improvement, 2018). In interviews, practitioners also described using parent-reported screeners to build rapport and strengthen communication with families about a child's skills and needs. Screeners can be used in medical settings (for example, pediatric practices) and childcare settings. They are often intended to be brief measures that capture quick, high-level snapshots of children's skills to identify whether a referral is needed for a further, more detailed assessment of children's abilities.

Because screeners often take place in real-world settings and are not intended to provide a detailed measure of a child's skills and abilities, they are generally inexpensive, quick, and do not require extensive training to complete; they are often filled out by a child's caregivers. These caregiver reports include scales, checklists, or interviews provided by parents, teachers, and others who see a child in various contexts. They are completed by an adult who sees the child consistently and can assess that child's skills or abilities overall, not only for the assessment's duration (Miller et al., 2017). Caregiver reports provide a holistic view of a child that may vary in different contexts, situations, and interactions with others — for example, with strangers as compared with familiar caregivers (Miller et al., 2017).

BOX 1

Considerations for Measurement Selection

Infant-toddler assessment comes with a unique set of challenges. Selection of measures in this age range can be guided by reviewing the needs of stakeholders and the purpose for measurement. Based on the purpose, each measurement selection should balance the logistical constraints of collecting the data with psychometric considerations. First, measures can be assessed for logistical constraints. After narrowing the range based on those considerations, instruments can be reviewed to ensure that they meet the appropriate bar for psychometric strength, depending on their intended purpose.

Deciding on the purpose for collecting the measure helps guide decision-making about how to balance the logistical constraints of a measure with the psychometric needs of the intended purpose. Measures that inform large-scale policies may require a level of confidence in their reliability and validity that calls for a substantial investment of resources and effort. Meanwhile, it may be most important for measures that are used for frequent, less high-stakes decisions to be easily administrable multiple times during the year—a logistical consideration that may lead to shorter and less detailed assessments of children's abilities.

Logistical considerations can influence the ability of an organization or stakeholder to use a specific measure. In a formal school setting, assessments occur during the school day and can be administered to children directly as tests or other written work. However, young children also receive services in less formal settings, including pediatric practices, family- or home-based child care, and center-based programs. Each setting has different logistical considerations, including how extensively personnel can be trained, how long assessments can be, and how much space is available for assessing children. Other considerations may include the frequency and length of administration, the need for a certified or trained assessor, requirements for retraining, administration restrictions, the possible age range for the assessment, and cost. In addition, young children often have not been exposed to formal schooling and may speak only their home language or a mix of English and their home language. Caregivers who fill out reports on children may also be more comfortable speaking their own home language. Measures to assess children that are available and, if possible, validated in multiple languages, are most flexible for use across a diverse set of children and caregivers alike.

Psychometric considerations refer to the ability of an instrument to measure a child's skills reliably and validly. Reliability refers to an assessment's consistency in measuring the same skill or outcome over time or across assessors. Children's behaviors and skills can vary in different contexts, particularly at young ages. The first words an infant speaks at home may not be repeated in a child care or pediatric setting until much later (Fenson et al., 2000). Additionally, children's growth in the first three years of life is nonlinear and only moderately stable, with low to moderate correlations between a child's language at 12 months and 24 months (Bornstein, Tamis-LeMonda, & Haynes, 1999; Fenson et al., 2000). For these reasons, observing young children's skills in different contexts in a reliable and consistent way can be more challenging than when assessing older children. Validity refers to how well an assessment captures the intended skill or outcome. For example, measures used to screen for delays need to be validated to ensure that they accurately identify those children at risk and don't miss children who may need further support. In a typical testing setting, infants and toddlers are not able to

BOX 1 (continued)

report for themselves or demonstrate their skills on demand. Assessing whether an instrument is validly measuring a child's skills and whether those skills matter for future development requires innovative research paradigms or measures and studies with extensive longitudinal data, which are rare. Achieving the highest levels of confidence in reliability and validity often requires substantial resources and effort.

Screening tools for child outcomes typically derive a score that describes a child's risk for or likelihood of a delay based on previous research. While continuous scores may be derived from the assessment of a child's skills, the ultimate validated score tends to be binary, flagging whether or not the child is at risk for a delay and may need further evaluation to determine the extent or substance of that delay (National Center for Systemic Improvement, 2018). Information from screeners can be used to flag needed services and provide referrals. Along with logistical constraints, the need to discuss the child's functioning in multiple domains is why many screeners assess all domains in one instrument briefly, instead of measuring a single domain in depth.

Measuring Individual Differences and Change

Researchers and program evaluators may use infant-toddler measures to assess the effect of a program on children's outcomes, examine change in outcomes over time, or identify individual differences in children's abilities. Stakeholders in this group emphasized that they look for measures that align with and are predictive of the outcomes that ultimately matter to the program or evaluation, because policymakers often use infant-toddler outcome measurement for high-stakes decision-making and for guiding policy. Interviewees described policymakers' need for a measure that assesses how children are improving throughout the year. They also said that measures that work across age groups (in this case, from birth to age 5 or beyond) are strategically useful, as a variety of systems and programs continue to align to serve children across this age range.

Because both researchers and policymakers rely on the consistent collection of data across groups of interest, standardized and reliable measures are crucial. These measures maximize the collected data's utility in producing conclusions that can be applied to a larger population. Detailed, standardized assessments of children's development measure specific skills and abilities in this way, making them appropriate for psychoeducational testing and research purposes. Their administration is often standardized to maximize reliable data collection across assessors, and they are studied to ensure strong psychometric properties (Wortham & Hardin, 2016). They generally produce continuous scores that provide information about individual differences in children's specific abilities within a domain and are often standardized against a meaningful comparison group so that children's skills or changes in skills can be compared with their peers (National Research Council, 2000). Information from these assessments may be used for monitoring changes in children's skills and for program evaluation.

Standardized assessments for measuring individual differences and changes in skills can be collected through caregiver reports (described above) or through direct assessments. Direct assessments are conducted by a trained assessor sitting with a child in a quiet, separate space and presenting the child with novel stimuli in a standardized manner to elicit a response. For example, children's language can be directly assessed by asking them to point to or name specific items in a picture. Direct assessments generally require a high level of training to achieve reliable administration and can be resource-intensive, since each child needs to be assessed one-on-one. Most direct assessments are age-normed and have been studied extensively for reliability and validity in both general and population-specific contexts. Due to their strong reliability, these scores can be used to compare or rank children's skills against a larger population, identify individual differences between children or groups of children, and guide high-stakes decision-making (National Research Council, 2000). In measuring outcomes at young ages, the mode of assessment — caregiver report or direct assessment — influences the logistical complexity of gathering information and the range of information that can be collected.[1]

Population-level measures are another class of assessments designed to measure change throughout the year, producing summary scores of overall child health or development assessed at a state or national level. These measures can include a short form, intended for population-level assessment and producing an overall functioning score; and a long form, designed to produce more detailed, domain-specific scores for children that would be suited to the purposes of program evaluation or research. However, validation work on the psychometric properties, translations, and suitability to different contexts of these measures is preliminary and still ongoing (McCoy & Fink, 2018; Richter et al., 2019; Fernandes et al., 2014). As such, while available population-level measures were incorporated into and reviewed in this scan, they were ultimately not recommended.

Informing Teaching and Learning

Practitioners and policymakers also use child assessments throughout the year, on an ongoing basis. Teachers and caregivers may use formative assessments, which are conducted several times during the year and are typically less formal than standardized assessments, to monitor and evaluate a child's progress (Wortham & Hardin, 2016). Formative assessments tend to be embedded within a program's activities or instruction and can take on many forms (for example, a child's oral reflection on a book that has been read) that do not necessarily produce a quantitative score. Information from formative assessments is often used to shape instruction, inform professional development, or refine implementation of a program (Wortham & Hardin, 2016). Formative assessments are rarely used in program evaluation. In interviews, experts recommended that they be used solely for their intended purpose of supporting instruction. Many formative assessments do not have established psychometric properties (Meisels, Wen, & Beachy-Quick, 2010). Practitioners and policymakers, meanwhile, noted that such measures should be low-cost, frequent, and capable of being implemented quickly, so they can be used for day-to-day adaptation. Additionally, both groups suggested that selection of assessments for this purpose is often driven by two considerations: policies and funding requirements, and alignment with a specific curriculum.

[1] In this review, we included measures that could be assessed in early child care centers, homes, or pediatric practices. We did not review lab-based observational protocols or measures, which are not generally feasible in less-controlled settings.

RECOMMENDED MEASURES

This next section lays out instruments for each purpose of measurement. It focuses on the first two categories: “identifying delay” and “assessing individual differences and change.” Recommendations are not offered in the third category, “informing teaching and learning,” because the requirements for selection of these measures are constrained by specific program and curriculum characteristics and are therefore highly context-dependent.

Instruments were selected through a targeted review that balanced the logistical and scalability concerns described above for each purpose. To address logistical considerations for data collection across real-world contexts, priority was given to: (1) measures that were validated in multiple languages and could be used with a wide range of populations; (2) relied on caregiver reports of child skills; and (3) would not require trained assessors, take too much time, or include resource-intensive assessment models. To further support scalable data collection and to streamline logistical challenges, the review favored measures that were described by stakeholders as widely in use — a characteristic indicative of the feasibility of administration and usability of the data. Within each category, unique criteria were applied if they were raised during the interviews (described below) to address logistical and scalability considerations. The final measures identified for feasibly assessing infant-toddler language and preliteracy skills were reviewed for psychometric strength to ensure they met the psychometric bar for their intended use. See Table 1 for the review of the final measures.

In the “identifying delay” category only, measures that assessed all areas of child development (that is, whole child assessments) were prioritized, since teachers and pediatricians may not have time to assess each area individually. This led to two recommended measures (Ages and Stages Questionnaire-3rd Edition [ASQ-3] and Parents’ Evaluation of Developmental Status [PEDS]). One additional measure that has been rapidly taken up by cities and pediatric practices (Survey of Well-being of Young Children [SWYC]) was included as an additional recommendation, but psychometric work is still underway for this measure.

In the “assessing individual differences and change” category, measures that specifically focused on language and preliteracy were prioritized to provide the most detailed assessment of children’s skills within a limited time frame. Two caregiver reports met that criteria (MacArthur-Bates Communicative Development Inventories [MB-CDI] and Early Childhood Longitudinal Study-Birth Cohort [ECLS-B] Parent Report Preliteracy Questions). In addition, given the extensive use of more resource-intensive direct assessments to inform high-stakes decision-making by researchers and policymakers due to their psychometric strength, direct assessments were also reviewed in this category, but only as a secondary recommendation for stakeholders with additional resources. Three measures were included as secondary recommendations: One was short enough to be logistically feasible and has been used extensively for program evaluation (Woodcock-Johnson IV Tests of Oral Language [W-J IV OL]: Picture Vocabulary); two longer measures provided more detailed, psychometrically strong assessments of children’s skills in both English and Spanish, but were also lengthy and required extensive training (Preschool Language Scales-5th Edition [PLS-5] and Expressive One-Word Picture Vocabulary Test-4th Edition [EOWPVT-4]). It is important to note that the age range for these three measures, while allowing for the assessment of children past age 3, only begins at age 2.

TABLE 1
Review of Top Measures by Logistical and Psychometric Criteria

Logistic Criteria					Psychometric criteria			
Measure[a]	**Use**	**Validation in other languages / Availability[b]**	**Mode**	**Time and price**	**Internal consistency[c]**	**Inter-rater reliability[d]**	**Concurrent validity[e]**	**Predictive validity[f]**
ASQ-3	ID delay	Validated in 5 other languages. Available in many others.	CR	10-15 min.; $55.00 manual, $11.50 per child[g]	✓	✓	✓	✓✓
PEDS	ID delay	Validated in English, Spanish, Vietnamese. Available in many others.	CR	20-30 min.; $89.95 manual, $0.84 per child	✓	✓	✓	✓✓
SWYC	*ID delay*	*Validated in English and Spanish. Available in many others.*	*CR*	*10-15 min.; free*	✓	—[h]	✓	—
MB-CDI	ID change	Validated in English and Spanish. Available in many others.	CR	20 min.; $59.95 manual, $1.20 per child	✓	✓	✓	✓✓
ECLS-B Parent Report Preliteracy Questions	ID change	Validated in English and Spanish.	CR	10-15 min.; free	—	—	✓	✓✓

(continued)

TABLE 1 (continued)

Logistic Criteria					Psychometric criteria			
Measure[a]	**Use**	**Validation in other languages / Availability[b]**	**Mode**	**Time and price**	**Internal consistency[c]**	**Inter-rater reliability[d]**	**Concurrent validity[e]**	**Predictive validity[f]**
W-J IV OL: Picture Vocabulary	*ID change*	*Validated in English and Spanish.*	*DA*	*5 min. per test.; $702.00 for full kit[i]*	✓	—	✓	✓✓
PLS-5	*ID change*	*Validated in English and Spanish.*	*DA*	*30-60 min., $406.75 for full kit*	✓	✓	✓	✓✓
EOWPVT-4	*ID change*	*Validated in English and Spanish.*	*DA*	*20 min., $185.00 for full kit*	✓	✓	✓	✓✓

NOTES:

[a] Bold indicates primary recommendations; italic indicates secondary recommendations; no shading indicates whole child measures that include the domain of language and literacy; blue shading indicates language- and literacy-specific measures.

[b] If a measure has been translated into another language and additional validation work has been conducted on that version, it is considered validated in that language. If a measure has been translated into another language, it is considered available in that language.

[c] Internal consistency: ✓ = Internal consistency was established ($\alpha > 0.7$).

[d] Inter-rater reliability: ✓ = Inter-rater reliability was established (> 0.8).

[e] Concurrent validity: ✓ = Evidence of concurrent validity with meaningful or appropriate infant/toddler measures.

[f] Predictive validity: ✓✓ = Significantly predictive of later language & literacy skills; ✓ = Significantly predictive of any later skill. Predictive validity nformation is reported for all versions of the measure in question.

[g] Costs of assessment per child are approximate calculations.

[h] Dashes (—) indicate that this research is not available.

[i] For direct assessments, costs of the full assessment kit — including scoring sheets, manuals, training materials, and any manipulatives — are listed.

As described earlier, practitioners and policymakers said in interviews that measures of child outcomes used in the "informing teaching and learning" category were generally selected by programs based on two main criteria: assessments that were mandated by federal, state, or local policies or funding requirements; and those that were aligned with the curriculum used in their child care settings. Because these criteria are highly specific to each program, no specific assessment was recommended in this category. However, any selection of formative assessments should also consider their psychometric characteristics and how those properties inform the information derived from the measure.

Appendix C provides a more detailed review of each of these measures, based on the full list of criteria, which can be found in Appendix A.

CONCLUSION

Infant-toddler measurement is complex and the reasons for its use by different stakeholders vary widely. In the measurement of infant-toddler outcomes, there is no perfect single instrument to test whether a child is on the path to third-grade literacy; the particular instrument selected will be directly tied to the purpose of the measurement and the context in which it is used. How widely the measure will be used, logistics such as how often the assessment needs to be conducted, and resource availability will also contribute to measure selection. This brief lays out several suggestions for considerations to review when selecting an instrument:

- First, consider how and for what purpose the measure will be used. The type of information or score needed and how that information will be used should inform the logistical and psychometric considerations for measurement selection.

- Review the logistical constraints implied by the measure's proposed use. Measures that need to be conducted in real-world care settings (for example, pediatric practices or classrooms) and by varied reporters may require time limits, translations in multiple languages for diverse respondents, and a low bar for training in how to administer them.

- These logistical constraints may pose a trade-off with the types of information that can be gathered and how that information can be used. Quick, parent-reported measures designed to assess multiple domains of functioning in a pediatric office to flag potential delay, for example, may not provide the nuanced, precise information about changes in a child's language development that longer, domain-specific direct assessments of child skills would.

- Consider the psychometric properties of any potential measures. Decision-making should be based on the most reliable and valid measure that is possible given the logistical constraints of data collection. The purpose for which the information will be used, the logistical constraints on how frequently a measure is administered, and how high stakes the decisions derived from this information will be may suggest different psychometric "bars." For example, data used to inform funding decisions may need to clear a higher bar for stability and validity than information used by a teacher to make small adaptations to daily practice.

While the brief identifies promising measures available to date, additional work is needed:

- The targeted scan conducted in this brief focused on early language and preliteracy as direct predictors of third-grade reading ability. Research suggests several additional areas to consider that are also strong indicators of later reading outcomes, including the home environment, parent-child-specific preliteracy activities, and other or combined categories of infant-toddler ability and environmental factors (Evans, Li, & Whipple, 2013; Bornstein, 2012; Duncan et al., 2007). See Appendix D for additional domains and constructs from birth to age 3 that may be important indicators of later outcomes.

- Many of the infant-toddler measures reviewed in this brief could benefit from further development and refinement. For example, screeners that yield delay or risk scores, such as the Ages & Stages Questionnaire-3rd Edition (ASQ-3), are some of the most widely used measures of children from birth to age 3. Developing a way to use those data to create a continuous ability score would allow policymakers to capitalize on its widespread use in a more nuanced way. Alternatively, further validation and development of additional parent reports similar to the MacArthur-Bates Communicative Development Inventories (MB-CDI) could make standardized assessments more logistically feasible. Validation work for population-level measures is ongoing.

Taking on any new measure brings with it challenges and opportunities. Aligning a research agenda with the introduction of a measure in a program or policy could help build the knowledge base on the measurement of infant-toddler outcomes, create promising programs that support young children, and identify useful predictors of later outcomes.

APPENDIX A

Methodology

GENERAL SCAN

First, a targeted scan was conducted of available infant-toddler measures. These were drawn from sources ranging from existing compilations in the applied development research literature to internal reviews of infant-toddler instruments that were conducted for prior large-scale impact studies at MDRC.[1] In selecting a set of literature, the scan began with internal documents and continued through existing compendia until measures within any new documents became redundant. The following sources were consulted:

1. MDRC measurement scans conducted for relevant age groups

2. Office of Planning, Research and Evaluation (OPRE), *Early Childhood Developmental Screening: A Compendium of Measures for Children Ages Birth to Five* (Moodie et al., 2014)

3. National Center for Systemic Improvement, *Measuring Social and Emotional Development in Children Birth to Age 3* (2018)

4. Child Trends, *Early Childhood Measures Profiles* (Bridges et al., 2004)

5. James Bell Associates, *Design Options for Home Visiting Evaluation Compendium of Measurement Tools for MIECHV Grantees* (2016)

6. OPRE, *Early Head Start Research and Evaluation Project (EHSREP): 1996-2010 Measures Compendium* (Kopack Klein, Kemmerer, West, & Lim, 2016)

7. OPRE, *Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition* (Halle, Vick Whittaker, & Anderson, 2010)

From the preliminary scan, a total of 135 measures were identified.

Selection of a Subset of Measures

The scan applied a set of considerations primary to assessing progress in infants and toddlers toward kindergarten readiness and third-grade reading. These were used to narrow the breadth of potential areas and key constructs of infant-toddler development by prioritizing those that (1) were available for use, (2) focused on the domains of language and preliteracy, and (3) were validated for use with infants and toddlers from birth to age 3. Sorting the full list of measures using these three criteria resulted in a smaller subset of 40 infant-toddler measures that were reviewed in more depth. A review of these 40 measures can be found in Appendix B.

[1] For example, Head Start CARES assessed the effects of three social-emotional, preschool curricula on 3- and 4-year-olds around the country. Mother and Infant Home Visiting Program Evaluation (MIHOPE) was a national evaluation of four Maternal, Infant and Early Childhood Home Visiting (MIECHV) program-funded home visiting models and their two-generational outcomes.

Interviews

To narrow in on measures that met the needs of key stakeholders, interviews were conducted with practitioners, researchers, and policymakers. These included experts on state infant-toddler policies, experts on infant-toddler measurement, MDRC experts on early childhood measurement, and home- and center-based child care providers.

All addressed the relative strengths and weaknesses of various instruments and modes of administration based on their experience. In particular, they were asked to reflect on what their field was looking for, their experience with particular assessments, and their perspective on which ones could be practically applied. Experts were also asked about additional instruments that may not have been on the list, to ensure that all of the most recent and relevant research-based and locally derived measures were included in the review.

Three distinct uses for assessment arose from these conversations: identifying delay; measuring individual differences in skills and identifying change in those skills across a year; and informing teaching and learning.

FINAL RECOMMENDATIONS

Each measure within the subset of 40 was categorized by primary use. Eleven were categorized as identifying delay, 18 as identifying change across a year, and 11 as informing teaching and learning.

Measures were reviewed using the full set of criteria, detailed at the end of this appendix. In all categories, measures were catalogued by domain (specific to assessing language and literacy, assessing all domains of child functioning), mode of administration (caregiver report, direct assessment), availability or validation in languages other than English, and how widely the measure was in use. Wideness of use was a criterion derived from interviews with stakeholders, who identified it as an important contributor to whether an instrument would be selected and utilized.

The purpose of the scan was to identify logistically feasible measures for use by a wide range of stakeholders and to provide useful data for stakeholder needs. To this end, measures that were given priority were (1) available and validated in additional languages for use with diverse populations; (2) considered feasible because they were short, did not require a highly trained assessor, and were identified by interviewees as already in wide use by infant-toddler service providers; and (3) had adequate reliability and validity. To bolster the practicality of the second criteria, caregiver reports were favored because they did not require highly trained assessors or resource-intensive data collection. Recommended instruments were identified for each category after applying these criteria and any additional unique criteria that emerged from interviews about needs for assessments in that category (see below for category-specific criteria).

Within the "identifying delay" category, nine measures were validated for use in other languages. Of those, seven were caregiver reports. In the "identifying delay" category only, measures that assessed all domains of child development (that is, whole child assessments) were prioritized, since teachers and pediatricians might not have time to assess each domain individually; this led to five measures.

Finally, discussions with stakeholders identified two measures as the most widely used currently in pediatric and child care settings (Ages & Stages Questionnaire-3rd Edition [ASQ-3] and Parents' Evaluation of Developmental Status [PEDS]). One additional measure was raised by stakeholders because it has been rapidly taken up by cities and pediatric practices (Survey of Wellbeing of Young Children [SWYC]) and was included as a secondary recommendation. However, additional psychometric work is needed for this measure to be recommended.

For the "assessing individual differences and change" category, ten measures were available in other languages, and six of those were validated for use in other languages. Of the six remaining measures, two were caregiver reports. The two caregiver reports were the MacArthur-Bates Communicative Development Inventories (MB-CDI) and Early Childhood Longitudinal Study-Birth Cohort (ECLS-B) Parent Report Preliteracy Questions.

Given researcher and policymakers' extensive use of more resource-intensive direct assessments for high-stakes decision-making, direct assessments were also reviewed in this category only as a secondary recommendation for those with additional resources. Of the six measures validated in other languages, four were direct assessments. Measures that specifically focused on the domain of language and literacy were prioritized in this category to provide the most detailed assessment of individual differences in children's skills within the limited time for assessment, leading to three measures. This led to a secondary recommendation of one measure that was short enough to be logistically feasible (Woodcock-Johnson IV Tests of Oral Language [W-J IV OL]: Picture Vocabulary) and two longer measures that provide more detailed assessments of children's skills but are also lengthy and require extensive training (Preschool Language Scales-5th Edition [PLS-5] and Expressive One Word Picture Vocabulary Test-4th Edition [EOWPVT-4]). It is important to note that the age range for these three measures, while meeting the needs identified by policymakers to allow for assessment into the preschool years, only begins at age 2.

Finally, within the "informing teaching and learning" category, nine measures were validated for use in other languages. Of those, seven were caregiver reports. As described above, practitioners and policymakers identified in interviews that measures of child outcomes used in the "teaching and learning" category were selected by programs based on two main criteria. Practitioners reported selecting assessments that were mandated by federal, state, or local policies or funding requirements, and that were aligned with the curriculum used in their child care settings. Because these criteria are highly specific to each program, no specific assessment was recommended in this category.

Table 1 in the brief provides an overview of these top measures on the full list of criteria, and Appendix C offers a more detailed review of each of these measures.

ADMINISTRATIVE/LOGISTICAL CRITERIA

Availability: The measure is available for purchase or publicly available for use.

Age range: The measure is designed for use with infants and toddlers from birth to age 3.

Domain(s): The measure either exclusively assesses language and literacy, or assesses the whole child, including language and literacy.

Primary use: The measure is used to identify delay, to identify change across a year, or to inform teaching and learning.

Accessibility in multiple languages: Versions of the measure are validated or available in other languages through the official publisher.

Mode of collection: The measure is either a direct assessment of the child's ability that is completed by a trained administrator, or items reported by the caregiver (including parent and teacher).

Length of assessment: The amount of time the measure takes to administer per child.

Cost: The price of purchasing the measure.

Standardized and accessible training: The training for those who wish to administer the measure is available for purchase.

In wide use: States or locales have formally recommended or standardized the use of this measure in their own contexts. Experts and practitioners report that this measure is widely used.

PSYCHOMETRIC CRITERIA

Internal consistency is a measure of reliability that tells how well the items in the screener/assessment address the same construct. A Cronbach's alpha coefficient greater than 0.7 indicates adequate internal consistency (Ponterotto & Ruckdeschel, 2007).

Inter-rater reliability is the degree of consensus among different raters using the same measure. It can be assessed using Cohen's kappa and deemed adequate when this score is greater than 0.8 (McHugh, 2012).

Concurrent validity shows that the measure correlates well with similar measures that have been validated when administered at approximately the same time.

Predictive validity demonstrates that the measure correlates with later skills in the same or in other domains.

The measure is **validated for use in minority subgroups**, meaning prior studies have assessed the measure's validity for racially and ethnically diverse groups, low-income groups, and/or language minority groups.

APPENDIX

B

Subset of 40 Measures that Meet Scan Criteria, Based on Availability, Domain, and Age Range

From a preliminary scan of existing literature, 135 measures were identified, drawing from existing compilations in the applied development research and interviews with policy and measurement experts. Next, a subset of 40 measures was identified for more systematic review by prioritizing those that: (1) were available for use, (2) focused on the domains of language and preliteracy, and (3) were validated for use with infants and toddlers from birth to age 3. See Table B.1 for the subset of measures. This appendix summarizes those measures based on the age range they assessed, if they were available in other languages, the mode of administration, whether they were reported in interviews to be widely in use currently, and their primary use.

TABLE B.1
Subset of 40 Measures, by Primary Use, Domain, and Other Criteria

Measure[a]	Age	Other languages[b]	Mode	Reported to be in wide use[c]	Uses[d]		
					Identify delay	Identify change	Teaching, learning
Communication and Symbolic Behavior Scales, Developmental Profile: Infant/ Toddler Checklist	0:6-2	✓✓	Caregiver		✓✓	✓	✓
Child Behavior Checklist 1 ½ -5, Language Development Survey	1:6-2:11	✓✓	Caregiver		✓✓	✓	
Ages & Stages Questionnaire (ASQ-3)	**0:1-5:6**	✓✓	**Caregiver**	✓	✓✓		✓
Survey of Well-being of Young Children (SWYC)	*0:1-5:5*	✓✓	Caregiver	✓	✓✓		✓
Parents' Evaluation of Developmental Status (PEDS)	**0-8**	✓✓	Caregiver	✓	✓✓	✓	
Developmental Profile (DP-3)	0-12:11	✓✓	Caregiver		✓✓	✓	
Infant Development Inventory (IDI)	0-1:6	✓✓	Caregiver		✓✓		
Early Coping Inventory	0:4-3		Caregiver		✓✓		
Infant-Toddler Developmental Assessment (IDA-2)	0-3:6	✓✓	Direct		✓✓		✓
Vineland Adaptive Behavior Scales-III	0-90	✓✓	Direct		✓✓	✓	✓
Early Screening Profiles (ESP)	2-6:11		Direct		✓✓		

(continued)

TABLE B.1 (continued)

Measure[a]	Age	Other languages[b]	Mode	Reported to be in wide use[c]	Uses[d]		
					Identify delay	Identify change	Teaching, learning
MacArthur-Bates Communicative Development Inventories (CDI)	**0:8-2:6**	✓✓	**Caregiver**	**N/A**	✓	✓✓	
Early Childhood Longitudinal Study-Birth Cohort (ECLS-B) Parent Report Preliteracy Questions	**0:9-4**	✓✓	**Caregiver**	**N/A**		✓✓	
Woodcock-Johnson Tests of Oral Language (W-J IV OL): Picture Vocabulary	*2-80+*	✓✓	*Direct*	*N/A*	✓	✓✓	
Preschool Language Scales (PLS-5)	*0-6:11*	✓✓	*Direct*	*N/A*	✓	✓✓	
Expressive One-Word Picture Vocabulary Test (EOWPVT-4)	*2-80+*	✓✓	*Direct*	*N/A*	✓	✓✓	
New Reynell Developmental Language Scales (NRDLS)	2-7		*Direct*	N/A	✓	✓✓	
Test of Early Language Development (TELD-4)	2-7		Direct	N/A	✓	✓✓	
Mullen Scales of Early Learning (MSEL)	0-5:8		Direct	N/A	✓	✓✓	
Sequenced Inventory of Communication Development-Revised (SICD-R)	0:4-4		Direct	N/A		✓✓	
Child Development Inventory (CDI)	1:3-6		Caregiver	N/A	✓	✓✓	
Caregiver-Reported Early Development Instruments (CREDI)*	0-3	✓✓*	Caregiver	N/A		✓✓*	

(continued)

TABLE B.1 (continued)

					Uses[d]		
Measure[a]	**Age**	**Other languages[b]**	**Mode**	**Reported to be in wide use[c]**	**Identify delay**	**Identify change**	**Teaching, learning**
Global Scale for Early Development (GSED)*	0-3	✓✓*	Caregiver	N/A		✓✓*	
Intergrowth-21st Neurodevelopmental Assessment (INTER-NDA)*	1:10-2:2	✓✓*	Caregiver	N/A		✓✓*	
Battelle Development Inventory (BDI-2): Normative Update	0-7	✓✓	Direct	N/A	✓	✓✓	
Griffiths Mental Development Scales-3	0-6		Direct	N/A	✓	✓✓	
Stanford-Binet Intelligence Scales for Early Childhood (SB-V)	2-5:11		Direct	N/A		✓✓	✓
Bayley Scales of Infant Development (BSID-III)	0:1-3:6		Direct	N/A	✓	✓✓	
Assessment Technology Incorporated: Galileo Pre-K	0-5	✓✓	Caregiver	✓		✓	✓✓
Brigance Inventory of Early Development (IED-III)	0-2:11	✓✓	Caregiver	✓	✓	✓	✓✓
Desired Results Developmental Profile (DRDP): Infant/Toddler Comprehensive View	0-3	✓✓	Caregiver	✓		✓	✓✓
Teaching Strategies GOLD	0-8	✓✓	Caregiver	✓		✓	✓✓
The Ounce Scale	0-3:6	✓✓	Caregiver	✓	✓	✓	✓✓
HighScope Child Observation Record (COR) Advantage	0-6	✓✓	Caregiver	✓			✓✓
The Vine Assessment	0-3		Caregiver	✓			✓✓

(continued)

TABLE B.1 (continued)

					Uses[d]		
Measure[a]	Age	Other languages[b]	Mode	Reported to be in wide use[c]	Identify delay	Identify change	Teaching, learning
Hawaii Early Learning Profile (HELP)	0-3	✓✓	Caregiver		✓		✓✓
Assessment, Evaluation, and Programming System (AEPS-2) for Birth to Three Years	1:10-2:2	✓✓	Direct		✓	✓	✓✓
Early Learning Accomplishment Profile (E-LAP)	0-3	✓✓	Direct			✓	✓✓
Carolina Curriculum for Infants and Toddlers with Special Needs (CCITSN-3)	0-3		Direct		✓		✓✓

NOTES:

[a] Bold indicates primary recommendations; italic indicates secondary recommendations; no shading indicates whole child measures that include the domain of language and literacy; blue shading indicates language and literacy-specific measures.

[b] Other languages: ✓✓ = Measure has been validated in at least one language other than English; ✓ = Measure is available in at least one language other than English.

[c] Reported to be in wide use: ✓ = Measure has been reported to be in wide use by experts across states or locales; N/A = Measures used to identify change are used for program evaluation or research and therefore are not typically in wide use across a state or locale.

[d] Uses: ✓✓ = Measure is primarily designed and used for this purpose; ✓ = Measure has been used for this purpose previously.

* = These measures are designed to target populations at a national level, meaning they produce summary scores of overall development per child that are intended to be assessed at a higher level than at the state or local level. However, each includes a longer form that produces more detailed, domain-specific scores per child. CREDI was developed in 2018, GSED was developed in 2019, and INTER-NDA was developed in 2014. As these measures are recent, validation work on their psychometric properties, translations, and suitability to different contexts is still ongoing (McCoy & Fink, 2018; Richter et al., 2019; Fernandes et al., 2014).

APPENDIX

C

Detailed Review of Selected Measures

MEASURES THAT IDENTIFY DELAYS

The **Ages & Stages Questionnaire-3rd Edition (ASQ-3)** (Squires & Bricker, 2009) is a caregiver-report screening tool that is widely used by early childhood education programs and child care centers, pediatric practices, and state/local organizations such as Early Head Start, Child Find, California's First 5 County Commissions, and Nurse-Family Partnership. It is designed for ease of use by caregivers, pediatricians, and educators to identify delay across the domains of communication, gross motor, fine motor, problem solving, and personal-social skills at 17 time points between birth and 36 months, with additional versions up to age 5.5. The measure has adequate reliability and validity, correctly identifying 86 percent of children ages 2 to 3 at risk for developmental delay. The screener is logistically feasible for collection on a large scale: It takes 10 to 15 minutes to collect, can be scored in less than 5 minutes, does not require extensive training, and is available in 15 languages (Squires & Bricker, 2009; Bridges et al., 2004). While the measure is designed to assess risk and identifies children below and above a risk cutoff for developmental delay, children are assessed on individual items and a summative score is created from those items. It may be possible to use this continuous score from the underlying items to assess individual differences for research and policymaking purposes, but further research is needed to assess the psychometric properties of such a score.

Description: 21 questionnaires (30 items each) and scoring sheets at 2, 4, 6, 8, 9, 10, 12, 13, 16, 18, 10, 22, 24, 27, 30, 33, 36, 42, 48, 54, and 60 months of age.

Age range: 0:1-5:6.

Domain(s): Whole child — communication, gross motor, fine motor, problem solving, and personal-social development.

Primary use: Identify delay.

Accessibility in multiple languages: Validated in English, Arabic, Chinese, French, Spanish, Vietnamese; also available in Persian, Korean, Portuguese, Hindi, Dutch, Thai, Norwegian, Turkish, and Afrikaans.

Mode of collection: Caregiver report, scored by trained professionals.

Length of assessment: 10-15 minutes to collect, 1-3 minutes to score.

Cost: $55.00 for the user's manual, and approximately $11.50 per child.

Standardized and accessible training: Yes; digital training with DVDs and on-site training are both offered.

In wide use: Yes; widely used in early childhood education programs and organizations like Early Head Start, Child Find, and California's First 5 Country Commissions.

Internal consistency: Cronbach's alpha = .51 to .87 for age intervals 0:2-5 across 5 domains.

Inter-rater reliability: 93 percent agreement between parents and trained examiners (Rothstein et al., 2017).

Concurrent validity: ASQ-3 showed moderate to high agreement with classifications from the Battelle Developmental Inventory (86 percent agreement of classification between measures) (Squires & Bricker, 2009).

Predictive validity: Using the ASQ, vocabulary size at 20 months of age was shown to predict semantic processing ability for newly learned words at 24 months of age (Borgström, von Koss Torkildsen, & Lindgren, 2015).

Validated for use in minority subgroups: The psychometric performance of the Spanish-language version of ASQ-3 was tested on a randomized cohort of Hispanic children in Spanish-speaking families in Philadelphia, Pennsylvania. The study found that the instrument's sensitivity to identifying severe delay ranged from .40 to .71 for children 9 to 41 months old and was strongest at .71 for children 31 to 41 months old (Gerdes et al., 2016).

The **Parents' Evaluation of Developmental Status (PEDS)** (Glascoe, 2013) is a caregiver-reported screener that assesses children's risk of delay across multiple domains. The measure covers a broad age range from 0 to 8 years old and is widely used in pediatric settings to identify a need for more intensive evaluation. The measure is created to be easy to use, with 10 items for parents to complete at each time point and availability in over 18 languages. The screener has strong inter-rater reliability between parents and trained assessors (*r* = .95) and adequate internal consistency (Cronbach alpha = .81). The measure is highly correlated with other assessments of children's outcomes, such as the Bayley Scales of Infant Development, the Vineland Adaptive Behavior Scale, and the Childhood Autism Rating Scale (Glascoe, 2003), and is predictive of later academic concerns (Wake et al., 2005).

Description: 10 items, with response and score forms available at 12 points from birth to age 8.

Age range: 0-7:11.

Domain(s): Whole child — language, motor, self-help, early academic skills, behavior, social-emotional development.

Primary use: Identify delay.

Accessibility in multiple languages: Validated in English, Spanish, and Vietnamese; also available in Somali, Hmong, Malaysian, Arabic, Chinese, Swahili, and others.

Mode of collection: Caregiver report, scored by trained administrator.

Length of assessment: 20-30 minutes.

Cost: $89.95 for the user's manual, and approximately $0.84 per child.

Standardized and accessible training: Yes; online training and certification modules are available via the official PEDStest.com website.

In wide use: Yes; in pediatric settings like TennCare (Tennessee Medicaid), in educational settings like Head Start, and in program evaluation, such as of Bright Futures and Healthy Steps.

Internal consistency: Cronbach alpha = .81 (Glascoe, 2003).

Inter-rater reliability: 95 percent agreement between caregivers and trained raters (Moodie et al., 2014).

Concurrent validity: Pearson correlations of .70 or higher with the Bayley Scales of Infant Development, the Vineland Adaptive Behavior Scale, and the Childhood Autism Rating Scale (Glascoe, 2003).

Predictive validity: Using PEDS, parent-reported concerns about self-help and school skills moderately predicted low language and academic scores two years later. Teacher concerns about early school skills moderately predicted low academic scores two years later (Wake et al., 2005).

Validated for use in minority subgroups: In 2012, PEDS was re-standardized on a nationally representative sample of families across the United States and Canada, which represented white non-Hispanic, black, American Indian, Asian, Hawaiian/Pacific Islander, Hispanic, and children of other ethnicities at proportions reflective of both 2010 U.S. Census indicators and 2020 projections (Glascoe, 2013). Most recently, the psychometric performance of the Mandarin adaptation of PEDS was assessed on a group of Mandarin-speaking caregivers and children. The screener identified children at risk of severe developmental delay with a sensitivity of .80, and caregivers found the tool easy to administer and useful (Toh et al., 2017).

The **Survey of Well-being of Young Children (SWYC)** (Tufts Medical Center, 2010) is a free and comprehensive caregiver-report screening tool consisting of six questionnaires: Developmental Milestones, Parent's Observations of Social Interactions (POSI), Baby Pediatric Symptom Checklist (BPSC), Preschool Pediatric Symptom Checklist (PPSC), parent concerns, and family context questions. Each questionnaire includes approximately 40 items across the age ranges assessed; however, at each time point only a subset of items are administered as a two-page screener for parents to complete. This measure scores delay across the domains of cognitive, language, and motor development; social-emotional functioning; and family risk factors (Tufts Medical Center, 2010). While the measure is relatively new, it is notable in that it is being rapidly taken up in clinical settings, with formal integration into electronic health record systems like EPIC. In Philadelphia, the Department of Public Health officially recommends the use of SWYC to screen children for general developmental progress, autism, and Early Intervention referral at 10 points between birth and 36 months (City of Philadelphia Public Health, 2019). As with ASQ-3, SWYC is logistically promising: It takes only 10 to 15 minutes to administer, is available in 10 languages, and is free. Early psychometric testing suggests that the measure has acceptable internal consistency and concurrent validity, and correctly identifies severe developmental delay in 69 percent of toddlers ages 9 to 41 months (Gerdes et al., 2018). Further research is needed to assess the predictive validity and inter-rater reliability of this tool (Moodie et al., 2014).

Description: 6 questionnaires with approximately 40 items total (Developmental Milestones, Parent's Observations of Social Interactions, Baby Pediatric Symptom Checklist, Preschool Pediatric Symptom Checklist, parent concerns, family context questions) and forms for 2, 4, 6, 9, 12, 15, 18, 24, 30, 36, 48, and 60 months of age.

Age range: 0:1-5:6.

Domain(s): Whole child — cognitive, language, and motor development; social-emotional functioning; autism risk; social determinants of health (such as family risk factors).

Primary use: Identify delay.

Accessibility in multiple languages: Validated in English and Spanish; also available in Burmese, Nepali, Portuguese, Haitian-Creole, Arabic, Somali, and Vietnamese.

Mode of collection: Caregiver report, scored by trained administrator.

Length of assessment: 10-15 minutes.

Cost: Free.

Standardized and accessible training: Yes; manuals and training resources are freely available on the official website (http://floatinghospital.org/The-Survey-of-Wellbeing-of-Young-Children).

In wide use: Yes; in pediatric settings and electronic health record systems such as the Philadelphia Department of Public Health and EPIC.

Internal consistency: Cronbach's alpha > .70 except on "irritability" subscale in BPSC; factor analysis conducted found loadings greater than .80 for the PPSC and POSI scales (Gerdes et al., 2018).

Inter-rater reliability: More research is needed (Moodie et al., 2014).

Concurrent validity: Low positive predictive values (.14-.49) and high negative predictive values (.89-.96) for children with severe delayed demonstrated overlap in identifying delay between SWYC and gold-standard clinical measures (BSID-III and DAS-II) (Gerdes et al., 2018).

Predictive validity: More research is needed.

Validated for use in minority subgroups: The psychometric performance of the Spanish-language version of SWYC Milestones was tested on a randomized cohort of Hispanic children in Spanish-speaking families in Philadelphia, Pennsylvania. The study found that the instrument's sensitivity to identifying severe delay ranged from .59 to .76 for children 9 to 41 months old and was strongest at .76 for children 31 to 41 months old (Gerdes et al., 2016).

MEASURES THAT IDENTIFY CHANGE ACROSS A YEAR

The **MacArthur-Bates Communicative Development Inventories (MB-CDI)** (Fenson et al., 1993) are a collection of two caregiver-report checklists that assess a child's proficiency in language and communication. The first, Words and Gestures, is an infant form for ages 8 to 16 months; the second, Words and Sentences, is a toddler form for ages 16 to 30 months. Both are validated in Spanish (see *Inventarios*) and adapted into a variety of other languages. Depending on the child's skill level, the CDI takes approximately 20 minutes to complete and 10 to 15 minutes to score (Bridges et al., 2004). Short forms are available for both infants (89 items) and toddlers (100 items) that are highly correlated (r = .74 - .93) with the long-form versions and are logistically feasible for parent use. Because of its wide availability and ease to administer, the CDI has become a popular tool for assessing infant and toddler language and communication proficiency by both caregivers and developmental psychologists. Three of the four components across the forms demonstrate strong internal consistency (Cronbach's alpha coefficient > .90), strong concurrent validity with the Expressive One-Word Picture Vocabulary Test (EOWPVT) (Brownell & Martin, 2011), and adequate predictive validity of 6-month-apart outcomes between the infant and toddler forms (r = .38 to .73, median .69) (Fenson et al., 1993). MB-CDI results have shown to be an accurate basis on which to estimate infant-toddler total receptive and expressive vocabulary sizes (Mayor & Plunkett, 2010). Additionally, MB-CDI vocabulary scores on Words and Sentences at age 2 were able to predict language skills a year later as assessed on the extension of the assessment designed for children ages 30 to 37 months (r = .70) (Feldman et al., 2005). Scores from the MB-CDI can be used to reliably assess individual differences and change over time in children's language skills.

Description: Infant form for ages 0:8-1:4 (Words and Gestures) and toddler form for ages 1:4-2:6 (Words and Sentences).

Age range: 0:8-2:6.

Domain(s): Language and literacy — sample constructs: verbal comprehension, verbal production, gestures, vocabulary production, use of grammatical suffixes.

Primary use: Identify change.

Accessibility in multiple languages: Validated in English and Spanish; also available in Afrikaans, American Sign Language (ASL), Arabic, Cantonese, Mandarin, Tagalog, and many others.

Mode of collection: Caregiver report, scored by trained administrator.

Length of assessment: 20 minutes to report, 10-15 minutes to score.

Cost: $59.95 for the user's manual, and approximately $1.20 per child.

Standardized and accessible training: Yes; training materials and example infant and toddler forms available online.

In wide use: Yes; widely used by caregivers in home settings and researchers in program evaluation settings alike.

Internal consistency: Cronbach's alpha .95 to .96 for infant and toddler form vocabulary scales; .39 for infant form gesture scale (.79 for subcategories 1-3; .69 for subcategories 4-5); .95 for toddler form word/sentence complexity scale (Fenson et al., 1993).

Inter-rater reliability: A sample of 55 toddlers, 28 of whom were minimally verbal, were assessed by both parents and teachers on Words and Gestures. Inter-rater reliability between parents' and teachers' ratings on the word production and word understanding scores were high for the total sample (.87). For the subsample of minimally verbal toddlers, inter-rater reliability was slightly lower (.72). (Nordahl-Hansen et al., 2013).

Concurrent validity: Pearson correlations of .73 to .85 between Words and Sentences and the EOWPVT (Brownell & Martin, 2011). Using laboratory methods, correlations with the MB-CDI gestural scale were high (Fenson et al., 1993).

Predictive validity: A subsample of the norming sample completed the MB-CDI at one point and then again six months later. Of this subsample, 288 children were assessed using the toddler form twice. Between the two time points, the correlation for vocabulary scores was .71 and the correlation for grammatical complexity was .62 (Bridges et al., 2004). Also, 217 children moved from the infant to the toddler form. Correlations between the two forms ranged from .38 to .73. Sixty-two children were assessed using the infant form twice. The correlation was .44 for vocabulary comprehension, .38 for vocabulary production, and .44 for total gestures (Fenson et al., 1993).

Validated for use in minority subgroups: To assess the vocabulary of a cohort of low-income, Spanish-English bilingual children 24 to 48 months old in the Northeastern U.S., the English-language MB-CDI with the accompanying Spanish Vocabulary extension was used. Psychometric results showed that these forms demonstrated adequate concurrent and discriminant validity as measures of productive vocabulary in both English and Spanish (Mancilla-Martinez Garrez, Vagh, & Lesaux., 2016).

The set of preliteracy questions in **Section CD: Child Development, Literacy, and School Readiness** in the Parent Interviews of the **Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)** (National Center for Education Statistics [NCES], 2017) is a 10- to 15-minute battery of caregiver-report questions used to assess both the child's language and preliteracy skills and the parent's preliteracy activities with the child. Example questions are: "Although [CHILD] doesn't yet read storybooks on [his/her] own, does [he/she] ever look at a book with pictures and pretend to read?" and "Can [CHILD] identify the colors red, yellow, blue, and green by name?" The original birth cohort in which these interviews were conducted consisted of a nationally representative sample of approximately 14,000 U.S. children born in 2001, and variations of these questions were administered to their caregivers at five intervals from birth through kindergarten in order to track the children's development in language and literacy, among many other domains addressed in the full study. Data from this study have been used in many studies to demonstrate predictive validity within and across domains. For caregiver reports focused on preliteracy questions, the frequency of shared reading between caregivers and children and children's ability to combine words have shown to be strong predictors of both problem behaviors and low academic scores by the time a child enters kindergarten (Nelson et al., 2016).

Description: Collected at 9 months, 2 years, and 4 years old (prekindergarten age) as part of the ECLS-Birth Cohort study, alongside direct developmental assessments, cognitive assessments, and birth certificate data.

Age range: 0:9-4.

Domain(s): Language and preliteracy — child's language, child's preliteracy, parent's preliteracy activities with child.

Primary use: Identify change.

Accessibility in multiple languages: Validated in English and Spanish.

Mode of collection: Caregiver report.

Length of assessment: 10-15 minutes.

Cost: Free.

Standardized and accessible training: Protocols and user's manuals freely available on official website (http://nces.ed.gov/ecls/birthinstruments.asp).

In wide use: Used in some large-scale studies (for example, MIHOPE).

Internal consistency: This information is not reported for the parent-report preliteracy questions.

Inter-rater reliability: Not available for interview items.

Concurrent validity: Not available for interview items. For preschool reading-related field test items, concurrent validity with items from the Bracken Basic Concept Scale-Revised was .82 (Najarian et al., 2010).

Predictive validity: As captured by the ECLS-B parent-reported preliteracy questions, the frequency of shared reading between caregivers and children and children's ability to combine words predicted problem behaviors and low academic scores by kindergarten entry (Nelson et al., 2016).

Validated for use in minority subgroups: ECLS-B assessments were selected because they were deemed appropriate for the target population of the study, which was a nationally representative sample of 14,000 U.S-born children in terms of both socioeconomic and racial/ethnic background (National Center for Education Statistics, 2017).

Woodcock-Johnson IV Tests of Oral Language: Picture Vocabulary (W-J IV OL: Picture Vocabulary) (Schrank, McGrew, & Mather, 2014) is the first of 12 individually administered tests in the Tests of Oral Language. Picture Vocabulary, formerly in the Tests of Achievement in W-J III, was reorganized into a specific battery for oral language and linguistic ability in the new edition, which emphasizes the importance of language skills to overall cognitive and academic outcomes (McGrew, LaForte, & Schrank, 2014; Miller, 2014). For providers who can invest in staff training to administer a direct assessment but are concerned about the time burden inherent in a lengthy research measure, W-J IV OL: Picture Vocabulary is a more concise alternative for testing expressive vocabulary and that is also available in Spanish. The design of this assessment for a wide age range (2 to 80+) means that, specifically for infants and toddlers, the score is somewhat limited. On the other hand, its design makes this measure one that could potentially scaffold across several age ranges, from the infant and toddler to the prekindergarten to the kindergarten phase of a child's life. This tool's internal consistency is .94 for ages 2 to 3 and retains a median of .88 throughout the 2 to 80+ age range (McGrew, LaForte, & Schrank, 2014). Scores on the Picture Vocabulary test are highly correlated with scores on other W-J IV tests across the three batteries, including Science, Social Studies, and Humanities (.60, .65, and .62, respectively) (McGrew, LaForte, & Schrank, 2014).

Description: One test in a range of 12 in W-J IV OL (Oral Comprehension, Segmentation, Rapid Picture Naming, Sentence Repetition, Understanding Directions, Sound Blending, Retrieval Fluency, Sound Awareness, Vocabulario sobre dibujos, Comprensión oral, Comprensión de indicaciones). The other W-J IV instruments are the W-J IV Tests of Achievement (W-J IV ACH) and the W-J IV Tests of Cognitive Abilities (W-J IV COG).

Age range: 2-80+.

Domain(s): Language and literacy — oral language, broad oral language, verbal ability, vocabulary.

Primary use: Identify change.

Accessibility in multiple languages: Validated in English and Spanish (see the Batería Woodcock-Muñoz IV).

Mode of collection: Direct assessment.

Length of assessment: 5 minutes per test.

Cost: $702.00 for the full kit of W-J IV OL.

Standardized and accessible training: Only officially trained administrators may conduct this assessment; publisher offers single-day workshops and on-site training and certification services on website (nelson.com/assessment/training.html).

In wide use: Yes; W-J (including previous versions) is a popular option for a research measure.

Internal consistency: Cronbach's alpha = .94 for ages 2-3; median .88 throughout the full age 2-80+ range.

Inter-rater reliability: Not available.

Concurrent validity: Pearson correlations of .62 and .68 with the DAS-II General Conceptual Ability and School Readiness clusters (McGrew, LaForte, & Schrank, 2014).

Predictive validity: The NICHD Study of Early Child Care and Youth Development (SECCYD) used the Picture Vocabulary test in the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R), a previous version of W-J IV, to measure the pathways from variables of early child care to adolescent outcomes at age 15. WJ-R Picture Vocabulary results, which were collected at eight time points between first grade and age 15, partially predicted the association between early child care quality and overall cognitive-academic skills at age 15 (Vandell et al., 2010). The Picture Vocabulary test in the Woodcock-Muñoz Language Proficiency Battery, a Spanish edition of W-J IV, was used as part of a latent variable for oral language to predict first- and second-grade reading outcomes for English language learners (Nakamoto, Lindsey, & Manis, 2007).

Validated for use in minority subgroups: The specific psychometric properties of the Picture Vocabulary test for infants and toddlers in racially, ethnically, and/or socioeconomically diverse contexts has yet to be explored. However, Ortiz, Ortiz, & Devine discuss the considerations involved in using W-J IV to assess culturally and linguistically diverse groups in Chapter 16 of *W-J IV Clinical Use and Interpretation: Scientist-Practitioner Perspectives* (2016).

The **Preschool Language Scales-5th Edition (PLS-5)** (Zimmermann, Steiner, & Pond, 2011), also available in Spanish, is a comprehensive direct assessment of the language skills of children from birth to age 7 and 11 months. It uses two core language subscales: Auditory Comprehension and Expressive Communication. It also uses three supplemental assessments: the Language Sample Checklist, the Articulation Screener, and the Caregiver Questionnaire. As a direct assessment, PLS-5 usually takes 30 to 60 minutes to complete and may only be administered by trained professionals, making it a relatively resource-intensive tool that would likely only be used within the context of research and program evaluation or clinical practice (Zimmerman, Steiner, & Pond, 2011). The PLS-5 is widely used in early childhood intervention studies, including MDRC's Mother and Infant Home Visiting Program Evaluation (MIHOPE), a national evaluation of four Maternal, Infant and Early Childhood Home Visiting Program (MIECHV)-funded models and their two-generational outcomes; and Child First, an impact study of a home visiting intervention that targets at-risk children and families in Connecticut and North Carolina. It is a strong assessment of children's specific language skills and is psychometrically strong in terms of internal consistency (Cronbach's alpha coefficient = .91 to .98), inter-rater reliability (kappa = .96 to .99), and concurrent validity with both PLS-4 and CELF-P2 (*r* = .80 to .85 and .70 to .82, respectively) (Leaders Project, 2013). The PLS-4 at age 2 has been found to predict language at age 3, and language at 2 and 3 years of age predicted executive function at age 4 (Kuhn, et al., 2014).

Description: 25 questionnaires and 15 record forms; 2 core language subscales (Auditory Comprehension and Expressive Communication) and supplemental assessments (Language Sample Checklist, Articulation Screener, Caregiver Questionnaire).

Age range: 0-6:11.

Domain(s): Language and literacy — receptive and expressive language skills, comprehension, fluency, vocabulary.

Primary use: Identify change.

Accessibility in multiple languages: Validated in English and Spanish.

Mode of collection: Direct assessment.

Length of assessment: 30-60 minutes.

Cost: $406.75 for the full kit.

Standardized and accessible training: Only trained professionals may conduct this assessment; informational webinars and training opportunities are available on the publisher's website (http://pearsonassessments.com).

In wide use: Yes; widely used as a research measure in studies of national or state-level early childhood interventions, such as MIHOPE and Child First.

Internal consistency: Cronbach's alpha = .91 to .98 (Leaders Project, 2013).

Inter-rater reliability: .96 across subtests for ages 0-3:11.

Concurrent validity: .80 to .85 with PLS-4; .70 to .82 with CELF-P2 (Leaders Project, 2013).

Predictive validity: The Spanish version of PLS-3, a previous edition of PLS-5, has been used in studies of bilingual children to demonstrate that early language and preliteracy skills predict first grade reading outcomes (Scheffner Hammer, Lawrence, & Miccio, 2007). For example, growth in either the English or Spanish receptive vocabularies of bilingual preschoolers, measured at two time points, positively predicted their reading outcomes in letter-word identification and passage comprehension in both English and Spanish by the end of first grade (Davison, Hammer, & Lawrence, 2011).

Validated for use in minority subgroups: PLS-4 was normed with a nationally representative group, with nearly 40 percent of the sample identified as racial or ethnic minorities. PLS-4 did not show statistical differences between English-speaking Hispanic and English-speaking white children (Qi & Marley, 2010).

The **Expressive One-Word Picture Vocabulary Test-4th Edition (EOWPVT-4)** (Brownell & Martin, 2011) is a direct assessment that measures expressive vocabulary. The EOWPVT-4 asks children to give a word that best describes the pictures they are shown and can be used to test from 2 to 80 years of age, allowing for measurement of individual differences and changes over time across a wide age range. The measure takes approximately 10 to 15 minutes to administer in the younger ages and is administered by extensively trained assessors, making it resource-intensive to administer outside of a research study. The EOWPVT-4 includes a Spanish-Bilingual Edition for use with Spanish-speaking populations and has been used in research studies (for example, Head Start CARES, Head Start REDI, FACES 2009, Reach Out and Read) to assess program impacts and individual differences in Spanish- and non-Spanish-speaking children. The measure has demonstrated strong internal consistency (Cronbach's alpha = 0.93 to 0.98, with a median of 0.96 across different age groups). Correlations between EOWPVT-4 and other tests of vocabulary range from 0.67 to 0.90, with a median of 0.79. A previous version of EOWPVT-4 demonstrated that measurements of preschool preliteracy and literacy skills predicted reading ability in kindergarten (Lonigan, Burgess, & Anthony, 2000).

Description: 190 items with full-color illustrations describing common actions, objects, or concepts.

Age range: 2-80+.

Domain(s): Language — expressive vocabulary.

Primary use: Identify change.

Accessibility in multiple languages: Validated in English and Spanish.

Mode of collection: Direct assessment.

Length of assessment: 20 minutes.

Cost: $185.00 for the full kit.

Standardized and accessible training: Only officially trained administrators may conduct this assessment; publisher offers some training and certification opportunities on website (http://proedinc.com).

In wide use: Yes; widely used in research studies such as Head Start CARES, Head Start REDI, FACES 2009, and Reach Out and Read.

Internal consistency: Cronbach's alpha = .93 to .98.

Inter-rater reliability: 100 percent agreement between manual and computer scorers, experienced and inexperienced.

Concurrent validity: .67 to .90 with a median of .79 with PPVT-R, PPVT-III, ROWPVT, TELD, WISC-III Vocabulary, CAT-5, MAT-7, and SAT-9 (Bridges et al., 2004).

Predictive validity: A past version of EOWPVT-4, the Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R), was used to demonstrate that preschool preliteracy skills, as measured by the variables of phonological sensitivity, oral language, and nonverbal cognitive skills, predicted kindergarten reading ability, as measured by the variables of phonological sensitivity, letter knowledge, environmental print, and concepts about print (Lonigan, Burgess, & Anthony, 2000).

Validated for use in minority subgroups: EOWPVT-4 has an English version and bilingual version that was normed on a sample of bilingual speakers to create comparable scores across instruments (Brownell & Martin, 2011). The EOWPVT demonstrates strong concurrent validity with MB-CDI across monolingual and bilingual groups (Hoff, Rumiche, Burridge, Ribot, & Welsh, 2014).

APPENDIX

D

Additional Domains and Constructs Predictive of Kindergarten and Third-Grade Outcomes

With a primary focus on evaluating progress toward and predicting outcomes in third-grade reading ability, the current review targeted early language and preliteracy outcomes from birth to age 3 as direct predictors of literacy at age 8. However, research suggests additional indicators in infancy and toddlerhood that are also strongly predictive of later reading outcomes. These measures, while not direct precursors to future language skills, can provide additional clues to whether interventions in the early years may be working to change children's experiences and skills. This appendix briefly describes some alternative predictors.

At 36 months of age, a significant portion of a child's daily life is grounded in the **family context and home environment** (Bronfenbrenner & Morris, 2006). Research shows that the home environment in the early years is a strong predictor of later outcomes for children (Evans, Li, & Whipple, 2013). Indicators such as the number of books in the home correlate with children's language and cognition (Gottfried [Ed.], 1984; Sanders et al., 2004). The home environment is typically measured using structured observations by external observers or home visitors (for example, the Infant-Toddler Home Observation for Measurement of the Environment [IT-HOME] Inventory; Caldwell & Bradley, 2001), or parent reports of home environment and preliteracy activities (for example, the Early Head Start Research and Evaluation Project [EHSREP] parent interview; Paulsell, Kisker, Love, & Raikes, 2000).

Parent-child interactions are also strongly predictive of a child's later outcomes. Certain components of these interactions have been found to be particularly important. Maternal sensitivity to a child's cues, responsiveness to a child's needs, and overall warmth are all highly correlated with positive outcomes (Landry, Smith, & Swank, 2006; Juffer, Bakermans-Kranenburg & van IJzendoorn, 2008). Conversely, more intrusive parenting and harsher discipline have been found to be detrimental to children's outcomes (Chang, Schwartz, Dodge & McBride-Chang, 2003; Pettit, Bates, & Dodge, 1997). Within an interaction, parents' use of cognitively stimulating activities and questions and rich language are correlated with children's cognition and language (Bornstein, 2012; Dodici, Draper, & Peterson, 2003; Evans, Shaw, & Bell, 2000; Tamis-LeMonda, Shannon, Cabrera, & Lamb, 2004). Parent-child interactions have been measured using parent surveys (for example, the Conflict Tactics Scale-Parent Child Version; Straus et al., 1996) or observations of parents and children during structured and unstructured tasks (for example, the Three-Bag Task; Love et al., 2005).

In addition to children's preliteracy and language skills, **other domains of children's well-being** have been found to be strong predictors of reading outcomes by kindergarten and the third grade. It is noteworthy that in a rigorous but noncausal analysis of which skills at school entry predict a child's third-grade achievement, early math ability was the strongest predictor of third-grade reading, beyond even preschool reading skills (Duncan et al., 2007). Although newer analyses have raised questions about whether math itself, or other underlying cognitive factors, are driving the relationship (Bailey et al., 2018), cross-domain effects such as these suggest that other domains of a child's ability that are predictive of later reading, such as cognition and executive function, may also be promising constructs to measure.

REFERENCES

Achenbach, T. M. & Rescorla, L. A. (2001). *Manual for ASEBA school-age forms and profiles.* Burlington, VT: University of Vermont. http://www.ASEBA.org.

Akshoomoff, N. (2006). Use of the Mullen Scales of Early Learning for the assessment of young children with autism spectrum disorders. *Child Neuropsychology, 12*(4-5): 269-277. http://doi.org/10.1080/09297040500473714.

Alpern, G. D. (2007). *Developmental Profile — 3.* Torrance, CA: WPS. https://www.wpspublish.com/store/p/2743/dp-3-developmental-profile-3.

The Annie E. Casey Foundation. (2010). *Early warning! Why reading by the end of third grade matters*. Retrieved from https://www.aecf.org/resources/early-warning-why-reading-by-the-end-of-third-grade-matters/.

The Annie E. Casey Foundation. (2018). *2018 KIDS COUNT data book*. Retrieved from https://www.aecf.org/resources/2018-kids-count-data-book/.

Assessment Technology, Inc. (2002). *Galileo online technical manual* [Web]. http://www.ati-online.com/galileoPreschool/indexPreschool.php.

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist, 73*(1), 81-94. http://dx.doi.org/10.1037/amp0000146.

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

Biasini, A., Monti, F., Gianstefani, I., Bertozzi, L., Agostini, F., & Neri, E. (2015). Griffiths Mental Development Scales as a tool for the screening of motor disability in premature infants: Is it worth it? *Journal of Clinical Neonatology, 4*(1): 22-25. http://doi.org/10.4103/2249-4847.151162.

Black, M. M. & Matula, K. (1999). *Essentials of Bayley Scales of Infant Development II assessment.* New York, NY: Wiley. https://www.wiley.com/en-us/Essentials+of+Bayley+Scales+of+Infant+Development+II+Assessment-p-9780471326519.

Bogue, E. L., DeThorne, L. S., Schaefer, B. A. (2014). A psychometric analysis of childhood vocabulary tests. *Contemporary Issues in Communication Science and Disorders, 41*(1), 55-69. http://doi.org/1092-5171/14/4101-0055.

Borgström, K., von Koss Torkildsen, J., & Lindgren, M. (2015). Substantial gains in word learning ability between 20 and 24 months: A longitudinal ERP study. *Brain and Language, 149*(1), 33-45. http://doi.org/10.1016/j.bandl.2015.07.002.

REFERENCES *(CONTINUED)*

Bornstein, M. H. (2012). Cultural approaches to parenting. *Parenting, Science and Practice*, *12*(2-3), 212–221. http://doi.org/10.1080/15295192.2012.683359.

Bornstein, M. H., Hahn, C. S., Putnick, D. L., & Suwalsky, J. T. (2014). Stability of core language skill from early childhood to adolescence: A latent variable approach. *Child Development, 85*(4), 1346-1356. http://doi.org/10.1111/cdev.12192.

Bornstein, M. H., Tamis-LeMonda, C. S., & Haynes, M. (1999). First words in the second year: Continuity, stability, and models of concurrent and lagged correspondence in vocabulary and verbal responsiveness across age and context. *Infant Behavior and Development*, *22*(1), 65-85. http://doi.org/10.1016/S0163-6383(99)80006-X.

Bridges, L. J., Barry, D. J., Johnson, R., Calkins, J., Margie, N. G., Cochran, S. W., & Ling, T. J. (2004). *Early childhood measures profiles.* Retrieved from https://www.childtrends.org/publications/early-childhood-measures-profiles/.

Brigance, A. H. & French, B. F. (2013). *Brigance inventory for early development III.* North Billerica, MA: Curriculum Associates. https://www.curriculumassociates.com/products/brigance/early-childhood.

Bronfenbrenner, U. & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Theoretical models of human development*. Hoboken, NJ: John Wiley. http://doi.org/10.1002/9780470147658.chpsy0114/.

Brownell, R. & Martin, N. A. (2011). *Expressive One-Word Picture Vocabulary Test 4 examiner's manual.* Novato, CA: Academic Therapy. https://www.proedinc.com/Products/13692/eowpvt4-expressive-oneword-picture-vocabulary-testfourth-edition.aspx.

Caldwell, B. M. & Bradley, R. H. (2001). *HOME inventory and administration manual (3rd ed.)* Little Rock, AR: University of Arkansas for Medical Sciences.

California Department of Education. (2016). *Desired Results Development Profile (DRDP) 2015: A developmental continuum from early infancy to kindergarten entry* [Web]. Sacramento, CA: California Department of Education. https://www.cde.ca.gov/sp/cd/ci/drdpforms.asp.

Canivez, G. L. (2017). *Test reviews: Woodcock-Johnson IV* [PDF file]. Lincoln, NE: Buros Center for Testing. Retrieved from http://www.ux1.eiu.edu/~glcanivez/Adobe%20pdf/Publications-Papers/Canivez%20(2017)%20WJ%20IV%20Review.pdf/.

Chang, L., Schwartz, D., Dodge, K. A., & McBride-Chang, C. (2003). Harsh parenting in relation to child emotion regulation and aggression. *Journal of Family Psychology, 17*(4), 598-606. http://dx.doi.org/10.1037/0893-3200.17.4.598/.

REFERENCES *(CONTINUED)*

City of Philadelphia Department of Public Health. (2019). *Check & connect: Recommendations to promote healthy childhood development* [PDF file]. Retrieved from http://runningstarthealth.phila.gov/wp-content/uploads/2019/05/Check-and-Connect-Brochure.pdf/.

Claessens, A. & Dowsett, C. (2014). Growth and change in attention problems, disruptive behavior, and achievement from kindergarten to fifth grade. *Psychological Science*, *25*(12), 2241–2251. https://doi.org/10.1177/0956797614554265/.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*(4), 415-427. https://doi.org/10.1016/j.econedurev.2008.09.003/.

Creighton, D. E. & Suave, R. S. (1988). Minnesota Infant Development Inventory in the developmental screening of infants at eight months. *Canadian Journal of Behavioural Science, 20*(4): 424-433. http://dx.doi.org/10.1037/h0079933.

Cripe, J., Slentz, K., & Bricker, D. (1993). *AEPS curriculum for birth to three years, volume 2.* Baltimore, MD: Brookes. https://brookespublishing.com/product/aeps/.

Davison, M. D., Hammer, C., & Lawrence, F. R. (2011). Associations between preschool language and first grade reading outcomes in bilingual children. *Journal of Communication Disorders, 44*(4), 444-458. http://doi.org/10.1016/j.jcomdis.2011.02.003.

Dodici, B. J., Draper, D. C., & Peterson, C. A. (2003). Early parent–child interactions and early literacy development. *Topics in Early Childhood Special Education*, *23*(3), 124-136. https://doi.org/10.1177/02711214030230030301/.

Doig, K. B., Macias, M. M., Saylor, C. F., Craver, J. R., & Ingram, P. E. (1999). The child development inventory: A developmental outcome measure for follow-up of the high-risk infant. *Journal of Pediatrics, 135*(3): 358-362. http://doi.org/10.1016/s0022-3476(99)70134-4/.

DRDP Collaborative Research Group. (2018*). Technical report for the Desired Results Developmental Profile (2015)* [PDF file]. Prepared for the California Department of Education. Sacramento, CA: California Department of Education. Retrieved from https://www.desiredresults.us/sites/default/files/docs/resources/research/DRDP2015_Technical%20Report_20180920_clean508.pdf/.

Duncan, G. J., Claessens, A., Huston, A. C., Pagani, L. S., Engel, M., Sexton, H., … Japel, C. (2007). School readiness and later achievement [PDF file]. *Developmental Psychology, 43*(6), 1428-1446. https://www.apa.org/pubs/journals/releases/dev-4361428.pdf.

Edwards, S., Letts, C., & Sinka, I. (2011). *The New Reynell Developmental Language Scales*. London, UK: GL Assessment. https://www.gl-education.com/products/new-reynell-developmental-language-scales-nrdls/.

REFERENCES *(CONTINUED)*

Evans, G. W., Li, D., & Whipple, S. S. (2013). Cumulative risk and child development. *Psychological Bulletin, 139*(6), 1342-1396. http://dx.doi.org/10.1037/a0031808/.

Evans, M. A., Shaw, D., & Bell, M. (2000). Home literacy activities and their influence on early literacy skills. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 54*(2), 65-75. http://dx.doi.org/10.1037/h0087330/.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*, *76*(4), 856–868. http://dx.doi.org/10.1111/j.1467-8624.2005.00882.x/.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000). Measuring variability in early child language: Don't shoot the messenger. *Child Development, 71*(2), 323-328. http://doi.org/10.1.1.140.3574.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., … Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual.* San Diego, CA: Singular. http://www.brookespublishing.com/cdi.

Fernandes, M., Stein, A., Newton, C. R., Cheikh-Ismail, L., Kihara, M., Wulff, K., …Villar, J. for the International Fetal and Newborn Growth Consortium for the 21st Century. (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. http://doi.org/10.1371/journal.pone.0113360.

Furuno, S., O'Reilly, K. A., Hosaka, C. M., Inatsuka, T. T., Zeisloft-Falbey, B., & Allman, T. (1988). *Hawaii Early Learning Profile checklist (HELP).* Palo Alto, CA: VORT. https://www.vort.com/product.php?productid=1.

Gerdes, M., Garcia-Espana, J. F., Webb, D., Friedman, K., Winston, S., & Culhane, J. (2018). Psychometric properties of two developmental screening instruments for Hispanic children in the Philadelphia region. *Academic Pediatrics, 19*(6), 638-645. https://doi.org/10.1016/j.acap.2018.10.002/.

Glascoe, F. P. (2003). Parents' Evaluation of Developmental Status: How well do parents' concerns identify children with behavioral and emotional problems? *Clinical Pediatrics, 42*(1): 133-138. http://doi.org/10.1177/000992280304200206.

Glascoe, F. P. (2013). *Collaborating with parents, 2nd edition*. Nolensville, TN: PEDSTest.

Glover, E. M., Preminger, J. L., & Sanford, A. R. (1995). *Early Learning Accomplishment Profile revised edition (E-LAP).* Lewisville, NC: Kaplan.

REFERENCES *(CONTINUED)*

Gottfried, A. W. (Ed.). (1984). *Home environment and early cognitive development: Longitudinal research*. New York, NY: Elsevier. https://doi.org/10.1016/C2013-0-10749-0/.

Gustawan, I. W. & Machfudz, S. (2010). Validity of Parents' Evaluation of Developmental Status (PEDS) in detecting developmental disorders in 3-12 month old infants. *Paediatrica Indonesiana, 50*(1): 6-10. http://doi.org/10.14238/pi50.1.2010.6-10.

Halle, T., Vick Whittaker, J. E., & Anderson, R. (2010). *Quality in early childhood care and education settings: A compendium of measures, second edition*. Washington, DC: Child Trends. Prepared for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Harrison, P., Kaufman, A., Kaufman, N., Bruininks, R., Rynders, J., Ilmer, S., … Cicchetti, D. (n.d.). *Early Screening Profiles (ESP).* San Antonio, TX: Pearson. http://www.pearsonclinical.com/childhood/products/100000089/earlyscreening-profiles-esp.html#tab-details.

Hedrick, D., Prather, E., & Tobin, A. (1984). *Sequenced Inventory of Communication Development — revised edition: Test manual.* Torrance, CA: WPS. https://www.wpspublish.com/store/p/2970/sicd-r-sequenced-inventory-of-communication-development-revised.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur-Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14*(1), 40-51. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15966111/.

HighScope. (2019). *Child Observation Record Advantage.* Ypsilanti, MI: HighScope. https://coradvantage.com/.

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014). Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years. *Early Child Research Quarterly, 29*(4), 433-444. http://doi.org/10.1016/j.ecresq.2014.04.012.

Hresko, W. P., Reid, D. K., & Hammill, D. (1999). *Test of Early Language Development — third edition: Examiner's manual.* Austin, TX: Pro-Ed. https://www.proedinc.com/Products/14645/teld4-test-of-early-language-developmentfourth-edition.aspx.

Huntley, M. (1996). *Griffiths Mental Development Scales — revised: Birth to 2 years.* Oxford, UK: Hogrefe. https://www.hogrefe.co.uk/shop/griffiths-scales-of-child-development-third-edition.html.

Ireton, H. R. (1992). *Child Development Inventory (CDI).* Minneapolis, MN: Behavior Science Systems. http://childdevrev.com/specialiststools/child-development-inventory.

REFERENCES *(CONTINUED)*

Ireton, H. R. (1994). *Infant Development Inventory*. Minneapolis, MN: Behavior Science Systems. http://childdevrev.com/specialiststools/infant-development-inventory.

Ireton, H. R. & Glascoe, F. P. (1995). Assessing children's development using parents' reports: The Child Development Inventory. *Clinical Pediatrics, 34*(5): 248-55. http://doi.org/10.1177/000992289503400504.

James Bell Associates. (2016). *Design options for home visiting evaluation compendium of measurement tools for MIECHV grantees*. Retrieved from https://www.jbassoc.com/resource/design-options-home-visiting-grantees/.

Johnson-Martin, N. M., Attermeier, S. M., & Hacker, B. (2004). *The Carolina Curriculum for Infants and Toddlers with Special Needs, third edition*. Baltimore, MD: Brookes. https://brookespublishing.com/product/the-carolina-curriculum/.

Juffer, F., Bakermans-Kranenburg, M. J., & Van IJzendoorn, M. H. (2008). *Promoting positive parenting: An attachment-based intervention*. New York, NY: Taylor & Francis.

Kells, S. (2018). Introducing screening for family risks in young children in primary care. Scholarworks @UMass Amherst, Doctor of Nursing Practice (DNP) Projects, 166. Retrieved from https://scholarworks.umass.edu/nursing_dnp_capstone/166/.

Kopack Klein, A., Kemmerer, C., West, J., & Lim, G. (2016). *Early Head Start Research and Evaluation Project (EHSREP): 1996-2010 measures compendium*. OPRE Report 2016-101. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U. S. Department of Health and Human Services.

Kuhn, L. J., Willoughby, M. T., Wilbourn, M. P., Vernon-Feagans, L., Blair, C. B., & Family Life Project Key Investigators. (2014). Early communicative gestures prospectively predict language development and executive function in early childhood. *Child Development, 85*(5), 1898-1914. http://doi.org/10.1111/cdev.12249.

Lam, J. (2015). A systematic review of measurement instruments to assess cognition and language development at 24 months of age, for use in effectiveness trials of nurse-home visitation programs [PDF file]. Burnaby, BC, Canada: Simon Fraser University. Retrieved from https://pdfs.semanticscholar.org/74fe/9016a3c40cbddea11e4639a22126049fcf97.pdf/.

Landry, S. H., Smith, K. E., & Swank, P. R. (2006). Responsive parenting: Establishing early foundations for social, communication, and independent problem-solving skills. *Developmental Psychology, 42*(4), 627-642. http://dx.doi.org/10.1037/0012-1649.42.4.627/.

REFERENCES *(CONTINUED)*

Leaders Project. (2013). *Test review: PLS-5 English*. Web. Retrieved from https://www.leadersproject.org/2013/11/25/test-review-pls-5-english/.

Lenkarski, S., Singer, M., Peters, M., & McIntosh, D. (2001). Utility of the early screening profiles in identifying preschoolers at risk for cognitive delays. *Psychology in the Schools, 38*(1): 17-24. https://doi.org/10.1002/1520-6807(200101)38:1<17::AID-PITS3>3.0.CO;2-G.

LifeCubby. (2010). *The Vine Assessment.* Columbus, OH: LifeCubby. https://www.lifecubby.me/index.php.

Lonigan, C. J., Burgess, S. R., Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. Developmental Psychology, 36(5), 596-613. http://doi.org/10.1037//0012-1649.36.5.596.

Love, J. M., Kisker, E. E., Ross, C., Constantine, J., Boller, K., Chazan-Cohen, R., … Vogel, C. (2005). The effectiveness of early Head Start for 3-year-old children and their parents: Lessons for policy and programs [PDF file]. *Developmental Psychology, 41*(6), 885-901. https://www.apa.org/pubs/journals/releases/dev-416885.pdf.

Luiz, D. M., Foxcroft, C. D., & Povey, J. L. (2006). The Griffiths Scales of Mental Development: A factorial validity study. *South African Journal of Psychology, 36*(1): 192-214. https://doi.org/10.1177/008124630603600111.

Mancilla-Martinez, J., Gamez, P. B., Vagh, S. B., & Lesaux, N. K. (2016). Parent reports of young Spanish-English bilingual children's productive vocabulary: A development and validation study. *Language, Speech, and Hearing Services in Schools, 47*(1), 1-15. http://doi.org/10.1044/2015_LSHSS.15.0013.

Marchman, V. A. & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science, 11*(3), F9–F16. http://doi.org/10.1111/j.1467-7687.2008.00671.x.

Mayor, J. & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, *14*(1), 769-785. http://dx.doi.org/10.1111/j.1467-7687.2010.01024.x/.

McCoy, D. & Fink, G. (2018). *Caregiver Reported Early Childhood Development Instruments (CREDI).* Boston, MA: Harvard T. H. Chan School of Public Health. https://sites.sph.harvard.edu/credi/.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual. Woodcock Johnson IV* [PDF file]. Rolling Meadows, IL: Riverside. Retrieved from https://www.wjscore.com/Files/WJIVTechnicalManual.PDF/.

REFERENCES *(CONTINUED)*

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282. http://doi.org/10.11613/BM.2012.031.

Meisels, S. J., Marsden, D. B., & Dombro, A. L. (2003). *The Ounce Scale.* Bloomington, MN: Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Developmental-Early-Childhood/The-Ounce-Scale/p/100000403.html.

Meisels, S. J., Wen, X., & Beachy-Quick, K. (2010). Authentic assessment for infants and toddlers: Exploring the reliability and validity of the Ounce Scale. *Applied Developmental Science, 14*(2), 55-71. http://doi.org/10.1080/10888691003697911.

Miles, S. B. & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development*, *77*(1), 103-117. http://dx.doi.org/10.1111/j.1467-8624.2006.00859.x/.

Miller, D. C. (2014). *Using the WJ IV Cognitive, Oral Language, and Achievement Tests in research* [PDF file]. Denton, TX: Texas Woman's University Woodcock Institute. Retrieved from https://twu.edu/media/documents/woodcock-institute/Using-the-WJ-IV-Cognitive,-Oral-Language,-and-Achievement-Tests-in-Research.pdf/.

Miller, L. E., Perkins, K. A., Dai, Y. G., & Fein, D. A. (2017). Comparison of parent report and direct assessment of child skills in toddlers. *Research in Autism Spectrum Disorders, 41-42,* 57-65. http://doi.org/10.1016/j.rasd.2017.08.002.

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.

Mullen, E. M. (1989). *Infant Mullen Scales of Early Learning.* Torrance, CA: WPS. https://www.wpspublish.com/mullen-scales-of-early-learning.

Murphey, D., Epstein, D., Shaw, S., McDaniel, T., & Steber, K. (2018). *The status of infants and toddlers in Philadelphia*. Philadelphia, PA: William Penn Foundation. Retrieved from https://www.williampennfoundation.org/what-we-are-learning/status-infants-and-toddlers-philadelphia/.

Najarian, M., Snow, K., Lennon, J., Kinsey, S., & Mulligan, G. (2010). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B): Preschool-kindergarten 2007 psychometric report* [PDF file]. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2010/2010009.pdf/.

REFERENCES *(CONTINUED)*

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing, 20*(1), 691-719. http://doi.org/10.1007/s11145-006-9045-7.

National Center for Education Statistics. (2017). Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) [PDF file]. *NCES Handbook of Survey Methods*. Washington, DC: Institute of Education Sciences. Retrieved from https://nces.ed.gov/statprog/handbook/pdf/ecls_b.pdf/.

National Center for Systemic Improvement. (2018). *Measuring social and emotional development in children birth to age 3*. San Francisco, CA: WestEd.

National Research Council. (1998). Predictors of success and failure in reading. In C. E. Snow, S. M. Burns, & P. Griffin. (Eds.), *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

National Research Council. (2000). Assessment in early childhood education. In *Eager to learn: Educating our preschoolers*. Washington, DC: National Academies. http://doi.org/10.17226/9745.

Nelson, B. B., Dudovitz, R. N., Coker, T. R., Barnert, E. S., Biely, C., Li, N., … Chung, P. J. (2016). Predictors of poor school readiness in children without developmental delay at age 2. *Pediatrics, 138*(2). https://pediatrics.aappublications.org/content/138/2/e20154477.

Newborg, J. (2004). *Battelle Developmental Inventory — Second edition normative update.* Itasca, IL: Riverside Insights. https://www.riversideinsights.com/solutions/battelle-developmental-inventory?tab=0.

Nordahl-Hansen, A., Kaale, A., & Ulvund, S. E. (2013). Inter-rater reliability of parent and preschool teacher ratings of language in children with autism. *Research in Autism Spectrum Disorders, 7*(11), 1391-1396. https://doi.org/10.1016/j.rasd.2013.08.006/.

Ortiz, S. O., Ortiz, J. A., & Devine, R. I. (2016). Assessment of culturally and linguistically diverse individuals with the Woodcock-Johnson IV. In D. Flanagan and V. Alfonso, (Eds.) *WJ IV clinical use and interpretation: Scientist-practitioner perspectives*. New York, NY: Academic. https://www.elsevier.com/books/wj-iv-clinical-use-and-interpretation/flanagan/978-0-12-802076-0.

Paulsell, D., Kisker, E. E., Love, J. M., & Raikes, H. (2000). *Leading the way: Characteristics and early experiences of selected early Head Start programs. Volume III: Program implementation*. Washington, DC: Commissioner's Office of Research and Evaluation, Administration on Children, Youth, and Families, U.S. Department of Health and Human Services.

PEDStest.com. (2013). *Parents' Evaluation of Developmental Status (PEDS)*. Nolensville, TN: PEDStest.com. https://pedstest.com/.

REFERENCES *(CONTINUED)*

The Pennsylvania Key. (2018). Early learning outcomes reporting [Web]. Retrieved from https://www.pakeys.org/getting-started/ocdel-programs/early-learning-outcomes-reporting/.

Pettit, G. S., Bates, J. E., & Dodge, K. A. (1997). Supportive parenting, ecological context, and children's adjustment: A seven-year longitudinal study. *Child Development, 68*(5), 908-923. http://dx.doi.org/10.2307/1132041/.

Perrin, E. C., Sheldrick, C., Visco, Z., & Mattern, K. (2016). *The Survey of Well-being of Young Children (SWYC) user's manual*. Boston, MA: Tufts Medical Center. Retrieved from https://www.floatinghospital.org/-/media/Brochures/Floating-Hospital/SWYC/SWYC-Manual-v101-Web-Format-33016.ashx?la=en&hash=E0C2802F003ED312E9D5268374C540A112151FB3/.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105*(3), 997-1014. https://doi.org/10.2466/pms.105.3.997-1014.

Provence, S., Erikson, J., Vater, S., & Palmeri, S. (1995). *Infant-Toddler Developmental Assessment (IDA) administration manual.* Itaska, IL: Riverside. https://www.riversideinsights.com/p/infant-toddler-developmental-assessment-second-edition-ida-2-administration-manual/.

Prizant, B. M. & Wetherby, A. M. (2002). *Communication and Symbolic Behavior Scales Developmental Profile manual*. Baltimore, MD: Brookes. http://www.brookespublishing.com/resource-center/screening-and-assessment/csbs/csbs-dp/.

Putnam, S. P., Helbig, A. L., Gartstein, M. A., Rothbart, M. K., & Leerkes, E. (2014). Development and assessment of short and very short forms of the Infant Behavior Questionnaire — Revised. *Journal of Personality Assessment, 96*(4): 445-458. http://doi.org/10.1080/00223891.2013.841171.

Qi, C. H. & Marley, S. C. (2010). Validity study of the Preschool Language Scale-4 with English-speaking Hispanic and European American children in Head Start programs. *Topics in Early Childhood Education, 31*(2), 89-98. http://doi.org/10.1177/0271121410391108

Reardon, S. F. & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*. https://doi.org/10.1177/2332858416657343/.

Richter, L., Black, M., Britto, P., Daelmans, B., Desmond, C., Devercelli, A., …Vargas-Barón, E. (2019). Early childhood development: An imperative for action and measurement at scale. *BMJ Global Health, 4*(1): i54-i160. http://doi.org/10.1136/bmjgh-2018-001302.

Roid, G. A. (2003). *Stanford-Binet Intelligence Scales, fifth edition.* Torrance, CA: WPS. https://www.wpspublish.com/store/p/2951/sb-5-stanford-binet-intelligence-scales-fifth-edition.

REFERENCES *(CONTINUED)*

Rothbart, M. K. (1981). *The Infant Behavior Questionnaire (IBQ and IBQ-R)* [Web]. Brunswick, ME: Bowdoin. https://research.bowdoin.edu/rothbart-temperament-questionnaires/instrument-descriptions/the-infant-behavior-questionnaire/.

Rothstein, A., Miskovic, A., & Nitsch, K. (2017). Brief review of psychometric properties and clinical utility of the Ages and Stages Questionnaires, Third Edition for evaluating pediatric development. *Archives of Physical Medicine and Rehabilitation*, *98*(4), 809-810.

Sanders, L. M., Zacur, G., Haecker, T., & Klass, P. (2001). Number of children's books in the home: An indicator of parent health literacy. *Ambulatory Pediatrics, 4*(5), 424-428. https://doi.org/10.1367/A04-003R.1/.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV. Rolling Meadows, IL: Riverside. https://www.riversideinsights.com/solutions/woodcock-johnson-iv?tab=0.

Scheffner Hammer, C., Lawrence, F. R., & Miccio, A. W. (2007). Bilingual children's language abilities and early reading outcomes in Head Start and kindergarten. *Language, Speech, and Hearing Services in Schools, 38*(3), 237-248. http://doi.org/10.1044/0161-1461(2007/025).

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales Interview edition expanded form manual.* Circle Pines, MN: American Guidance Service. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Behavior/Adaptive/Vineland-Adaptive-Behavior-Scales-%7C-Third-Edition/p/100001622.html.

Squires, J. & Bricker, D. (2009). Ages and Stages Questionnaires — Third edition starter kit. Baltimore, MD: Brookes. https://brookespublishing.com/product/asq-3/.

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The Revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, *17*(3), 283-316. https://doi.org/10.1177/019251396017003001/.

Tamis-Lemonda, C. S. & Bornstein, M. H. (1990). Language, play, and attention at one year. Infant Behavior and Development, 13(1), 85-98. https://doi.org/10.1016/0163-6383(90)90007-U.

Tamis-Lemonda, C. S., Cristofaro, T., Rodriguez, E., & Bornstein, M. H. (2006). Early language development: Social influences in the first years of life. In L. Balter & C. Tamis-LeMonda (Eds.), *Child psychology: A handbook of contemporary issues.* Philadelphia, PA: Psychology Press.

Tamis-LeMonda, C. S., Shannon, J. D., Cabrera, N. J., & Lamb, M. E. (2004). Fathers and mothers at play with their 2- and 3-year-olds: Contributions to language and cognitive development. *Child Development, 75*(6), 1806-1820. http://dx.doi.org/10.1111/j.1467-8624.2004.00818.x/.

REFERENCES *(CONTINUED)*

Teaching Strategies. (2013). *Teaching Strategies GOLD® assessment system: Concurrent validity* [Web]. Washington, DC: Teaching Strategies. https://teachingstrategies.com/solutions/assess/gold/.

Toh, T. H., Lim, B. C., Bin Bujang, M. A., Haniff, J., Wong, S. C., Abdullah, M. R. (2017). Mandarin parents' evaluation of developmental status in the detection of delays. *Pediatrics International, 59*(8), 861-868. http://doi.org/10.1111/ped.13325.

Tufts Medical Center. (2010). *The Survey of Well-Being of Young Children*. Boston, MA: Tufts Medical Center. https://www.floatinghospital.org/The-Survey-of-Wellbeing-of-Young-Children/Overview.

Vandell, D. L., et al. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD Study of Early Child Care and Youth Development. *Child Development, 81*(3), 737-756. http://doi.org/10.1111/j.1467-8624.2010.01431.x.

Wake, M., Gerner, B., & Gallagher, S. (2005). Does Parents' Evaluation of Developmental Status at school entry predict language, achievement, and quality of life 2 years later? *Ambulatory Pediatrics, 5*(3), 143-149. http://doi.org/10.1367/A04-162R.1.

Wortham, S. C. & Hardin, B. J. (2016). How infants and young children should be assessed. In S. C. Wortham and B. J. Hardin (Eds.), *Assessment in early childhood education, eighth edition*. Hoboken, NJ: Pearson. http://www.mypearsonstore.com/bookstore/assessment-in-early-childhood-education-0135206529.

Zeitlin, S., Williamson, G. G., & Szczepanski, M. (1988). *Early Coping Inventory: A measure of adaptive behavior.* Bensenville, IL: Scholastic. http://www.ststesting.com/COPI.html.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales — Fifth edition examiner's manual.* San Antonio, TX: Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/Preschool-Language-Scales-%7C-Fifth-Edition/p/100000233.html.

ABOUT MDRC

MDRC IS A NONPROFIT, NONPARTISAN SOCIAL AND EDUCATION POLICY RESEARCH ORGANIZATION DEDICATED TO learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York; Oakland, California; Washington, DC; and Los Angeles, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff members bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-prisoners, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.