CEELO POLICY BRIEF

Measuring Child Outcomes in the Early Years

By W. Steven Barnett, PhD, Shannon Riley-Ayers, PhD, & Jessica Francis, PhD

Introduction

As our nation increases public and private investments to support the care and education of young children, there is increased concern about how specific public policies affect children before they enter primary school. This desire to establish cause and effect and to estimate the magnitude of benefits to children's learning, development, and wellbeing (LDWB) puts increased technical demands on assessment (discussed below). In addition, causal attributions require more than simply describing children's development over time, it requires rigorous research methodologies that warrant strong causal inferences.

The National Institute for Early Education Research (NIEER), partner to CEELO, was commissioned by OECD to provide a scholarly discussion paper that presented the pros and cons of various methods of and instruments used for reporting on international data of children's cognitive and social outcomes. This report draws from the work for that paper to provide information to inform decision-making regarding the assessment of young children's LDWB for state and national assessments designed to inform early childhood education (ECE) policy and practice. We include "wellbeing" because ECE should not merely be a means to improve a young child's future success in school, or even life, but should enhance the child's current quality of life. The primary focus here is on the preschool years. As there are many, many assessments available, this report does not review all of the individual assessments. Several much broader reviews with exhaustive compendia are already available such a publication from the U.S. National Academy of Sciences (Snow & Van Hemmel, 2008). Instead, we describe and illustrate each of the general approaches from which policy makers can choose.

Use of and Concerns about Assessments for Young Children

Broadly speaking, the use of assessments can be described as formative or summative. Formative assessment is the use of assessment to inform teaching with some definitions going so far as to equate formative assessment with scaffolding. Formative evaluation is internal and takes place during the educational experience. It looks forward in a process that is responsive to the needs of the learner. Summative assessment is the use of assessment to judge progress

or attainment relative to a standard. Summative assessment of the performance of a child looks backwards and may be used to judge the contributions of a teacher or program to child progress. Summative assessment generally is external in its orientation. Summative assessments may be used to inform professional development and other supports for teachers and programs, but they also may be used to make "high stakes" decisions including to sanction or reward teachers, schools, and to inform decisions about public programs and policies. In addition, summative assessments are commonly used to make high stakes about individual children including the provision of additional supports (e.g., special education services and services for immigrant children who have limited proficiency in the local language) and opportunities (e.g., programs for gifted children), as well as to determine whether a child should enter primary school at the typical age or delay entry. The last use is quite controversial and may be viewed as indicating a lack of supports and individualization in the first year of primary school.

As it is the use of an assessment that is formative or summative rather than the assessment instrument itself, the same instrument can be used for summative or formative purposes. Confusion can arise because of instruments have been designed so as to be particularly useful for formative or summative purposes, and sometimes the instruments themselves are referred to as formative or summative measures. In addition, there is a tendency to think of qualitative assessments and teacher observations as formative tools.

Despite the widespread use of assessment, there is widespread concern regarding potential negative consequences. Among the greatest concerns are (1) narrowing of instruction in ECE to focus on what is most easily measured; (2) misuse of assessments for high stakes decisions about children, teachers and programs; and, (3) excessive burdens on children and teachers from time consuming assessments. These concerns have been greatest for direct tests used for summative purposes, but they may arise with any type of assessments regardless of the use for which it was developed. For example, screening tests are sometimes misused to make high stakes decisions about children rather than to refer them for additional assessment. Screening tests also are sometimes used to collect data on a large scale to inform policy because they are quick to administer and so impose minimal costs on everyone involved. However, this should be done in full recognition that screening tests often are designed to err on the side of over identifying problems and may measure better at the lower end than at the higher end of the range of abilities or skills.

Concerns about negative impacts of assessment on learning and teaching and misuse by policy makers are lessened when assessment is conducted with broad observational measures embedded in the educational process for formative purposes. As we discuss in more detail in later sections, teachers may document in detail children's interests, dispositions, learning, development, and wellbeing as a tool to assist them in providing the best care and education for each child. Yet, even the types of data teachers collect for these purposes can be turned to other, summative purposes. Moreover, because the broader and more detailed such assessments become, the greater the time burdens they may impose on teachers.

Current Policy and Practice

As education policy in the United States varies greatly amongst the 50 states, a brief review of such policies provides insights into the range of different policies that might be adopted. Of the 40 states that offer publicly funded preschool education programs (typically at age 4), the vast majority require the use of some assessments, though not necessarily specifying the assessment or even the type of assessment. Most often this assessment is to be used for formative purposes by teachers, but most states also seek to use this information to inform teacher professional development. A few states require assessments for high-stakes decisions about children (e.g., kindergarten entry) or for summative purposes including the evaluation of teacher and program performance for sanction or reward. Such states may specify a specific assessment to be used with every child enrolled. State polices regarding preschool assessment are summarized in Table 1 below (Schilder and Carolan, 2014).

Table 1: State Uses of Assessments in Pre-K

How Pre-K Assessment Data Are Used By the States	Number of State Programs
Guide teacher training, professional development, or technical assistance	35
Track child and program level outcomes over time	34
Make adjustments to curricula	32
Provide a measure of kindergarten readiness	17
Make changes to state policies regarding the preschool program	16
Make decisions regarding a child's enrollment in kindergarten	6
Identify programs for corrective action or sanctions	5
Make funding decisions about programs or grantees	5
Evaluate teacher performance	2

Most states have or will soon adopt kindergarten entry assessments (KEAs) that measure learning and development when children enter kindergarten (the first year of primary school) after turning age 5. The use of these assessments also varies considerably by states. Some states intend these assessments to provide a broad baseline measure that describes children as they enter school. This information would be used by teachers to inform their practice, but also could be aggregated to inform policy makers about the needs of young children and to assess growth between entry at age 5 to kindergarten and the next time the state mandates uniform assessment of every child, typically at the end of third grade. KEAs often have a "whole child" perspective and are not narrowly academic. However, some state KEAs focus primarily on early literacy and, sometimes, a few other academic domains such as mathematics. A few states plan to use these assessments to judge the educational effectiveness of individual ECE providers and for this purpose the KEA may be aligned with an earlier assessment in the preschool years.

Deciding What and How to Assess

From the perspective of obtaining national or international data that can be used to inform policy rather than practice, there are key criteria to be used in deciding what and how to assess. These criteria are as follows:

1. **Measure what matters**. What aspects of LDWB are important and of concern to policy makers and the public?
2. **Measure well.** To be useful measures of what matters must be valid, reliable, fair, and age and developmentally appropriate.
3. **Assessments must be practical and affordable.** The younger the child, the more difficult it is to accurately assess their LDWB. The broader and deeper the assessment the higher the cost. In addition some aspects of LDWB are more difficult and expensive to assess. Time demands on children, teachers, parents, and others can be substantial (opportunity costs such as lost time from teaching), and the costs of professionals specifically hired (and trained) to administer assessments or interviews may be high as well.
4. **Results of assessments should be comparable.** This should be within and across programs/sites and over time.

Measuring What Matters

Children's LDWB encompasses virtually every possible outcome of ECE including children's happiness and life-satisfaction, habits and dispositions, attitudes and beliefs, cognitive abilities, social abilities, emotional development, physical development, health, and nutritional status. Such a broad view is consistent with the early childhood field's emphasis on attending to the needs of the whole child. In addition, one might add measurement of the extent to which a child's rights are respected, for example, the right of children to have a voice or active role in determining the activities in which they are engaged in ECE. This could be viewed as a means to producing outcomes for the child (for example, life satisfaction and attitudes toward society and schools). However, it could be viewed as an additional category.

Both common values and research indicate the importance of comprehensive measures. In most or perhaps all states, the goals of ECE are to support the development and well-being of the whole child. This is evident in the U.S. National Academy of Sciences report on the science of early childhood development (Shonkoff & Phillips, 2000) which recognized the value of:

> (1) the development of curiosity, self-direction, and persistence in learning situations; (2) the ability to cooperate, demonstrate caring, and resolve conflict with peers; and (3) the capacity to experience the enhanced motivation associated with feeling competent and loved (p.5).

Note that we have not described any of these domains or their measures as "outcomes." The use of the term "outcomes" raises the question: Outcomes of what? Children's learning, development, and

wellbeing are affected by all of their experiences at home and with family more generally, in ECE arrangements, and in the community as well as of their personal attributes. Drawing valid inferences about the specific influence of ECE experiences and the policies that shape them is much more complex than simply looking at correlations between ECE and child LDWB measures in a cross-section or longitudinally. One might call for randomized trials, and it is sometimes possible to conduct these with special data collections or in such a way that they can use data that would have been collected anyway. However, randomized trials are not always possible or ideal. It is much more likely that comparisons of the impacts of ECE and ECE policies within and across states will be conducted using complex statistical models that are more successful in producing valid inferences when there are assessments at multiple time points (at least one "pretest") and when the assessments are accurate and precise. These statistical methods also benefit from linked information on each child's family, home experiences, and ECE experiences.

What should be assessed does depend on the purposes for which an assessment will be used. If policy makers wish to evaluate differences in ECE quality and services, these may be expected to influence some aspects of learning and development more than others. For example, if the vast majority of young children are healthy and have good motor development, and these are carefully monitored by health professionals, then ECE programs may not much affect these domains. In this case, there may be little reason for educators to assess them. If there is a strong concern that children's rights to engagement and active decision making are not adequately respected, then this aspect of wellbeing may be an important focus of assessment.

Measuring Well: Desirable Features of Assessments

To be useful assessments should be valid and reliable. Assessments also should be fair. In early childhood there is particular concern that assessments be age and developmentally appropriate. This applies equally to all types of assessments, performance assessments as well as tests, qualitative as well as quantitative.

Validity is a fundamental criterion for selecting instruments to measure LDBW. The *Standards for Educational and Psychological Testing* state, "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). A valid instrument--whether an observation, interview, questionnaire, or test--should measure what it purports to measure (Williams & Monge, 2001). Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from an instrument so that it is always judged in the context of the purpose for which those inferences are made (Borg, Gall, & Gall, 1989). In assessing validity, what we wish to know is the extent to which interpretations of a measure hold across persons and contexts.

In essence, validity is established by producing and evaluating evidence on how well an assessment represents the construct it purports to measure (Messick, 1995). Validity depends on the extent to which an assessment represents the entire construct (i.e., it is not enough that items be from the appropriate domain, they must be fully representative of it). Validity also requires that a measure not include irrelevant items (as, for example, when language demands obscure the demonstration of math or social skills). In other words, an assessment can be invalid because it is too narrow and shallow or because it is too broad. An assessment also can be invalid because accurate representation does not generalize across populations and contexts.

There are multiple types of evidence that help to establish construct validity. These include assessments of content by experts, structural evaluation, comparison against a criterion, and prediction. The extent to which experts concur that an assessment fully covers the key dimensions of the construct being measured and does not tap irrelevant areas is sometimes referred to as face validity. Validity is also judged based on the structure of the assessment. Do patterns of results across items conform to theoretical expectations regarding the underlying concepts? Criterion validity is can be assessed by examining patterns of performance across ages and concurrent correlations with other assessments of the same construct. A high degree of correlation with an instrument that has well-established validity provides evidence supporting the validity of the target assessment. At the same time a valid measure should not be highly correlated with a measure that is believed to measure a completely different construct. Other approaches include estimating the extent to which the assessment predicts current or subsequent performance in "real life" that is contingent on what is measured.

Assuring validity for many assessments is not simply a matter of design, but also of assuring that procedures are appropriate for individual children. An obvious issue occurs when a child's home language differs from that of the assessment. Another is when a child has a disability, and this is most easily understood with respect to vision and hearing impairments. With respect to both issues accommodations often must be made to the child in order to maintain the validity of an assessment.

Reliability is the extent to which an assessment produces stable or consistent results because it produces little random error in its results (Creswell, 2008). A reliable assessment produces the same or highly similar results for a child on different occasions (assuming only a brief interval between assessments) and with different assessors (e.g., one teacher would not rate the same child differently from another teacher). A reliable assessment is also robust with respect to the circumstances of the assessment.

Reliability can be improved through several means. Optimizing the length or detail of an assessment is one way to increase reliability. The more items, or samples, obtained the less random error affects the results, unless, for example, a longer "test" results in fatigue or distraction for the child or assessor. Another is to construct items and their scoring so as to maximize clarity and minimize uncertainty or misunderstandings. Minimizing the influence of incidental factors in the environment or assessment

circumstances and subjective (idiosyncratic) interpretation also increase reliability as does guidance and training for the assessors.

Multiple approaches are available to evaluate reliability. One of the most common is examining internal consistency, or how the items (or samples) in the assessment relate to one another. Historically, reliability as judged by internal consistency has been assessed using Chronbach's Alpha, though recently this approach has been challenged and others recommended as more appropriate (Yang & Green, 2011). All of these approaches produce reliability coefficients (a measure of correlation among items). In general tests that have a reliability of .80 or higher can be considered as sufficiently reliable for most research purposes (Borg, Gall, & Gall, 1989). However, reliability coefficients should be judged carefully since the value adequacy depends on the phenomenon studied (Hancock & Mueller, 2010). Values of .90 have been recommended for assessments used for high stakes decisions about individuals (Yang & Green, 2011).

Other common measures of reliability are the correlations of repeated assessments of the same child by the same assessor and inter-rater agreement of different assessors. Inter-rater agreement also may be assessed as criterion-related observer reliability, which is the extent to which a trained observer's scores agree with those of an expert observer (Borg, Gall, & Gall, 1989). It is important because it declares that the trained observer understands the variables measured in the instrument with the same efficacy as an expert observer. Again, there are norms with respect to the extent of agreement required and this depends on the use with the highest levels of agreement required when use relates to an individual child. A high level of reliability is important not just when use is summative but also when used to inform individualized education of a child.

Fairness refers to the ways in which assessments are used rather than a property of assessments per se. In addition, it is socially defined rather than scientifically defined. In our view, fairness does depend on validity and reliability because for an assessment's use to considered fair most would agree that the assessment should be free of bias (e.g., with respect to gender, family background, or national origin) and that random error should not be higher for some types of children than others (at least at the same age). However, even a valid and reliable assessment can be applied in ways that are not fair.

One concern in the early childhood field is that assessments developed for older children not be pushed down to younger children when they are neither age nor developmentally appropriate. This concern arises, in part, because of the much greater availability of assessments for older children than for younger children. As demands grow to assess young children on a broader set of domains for which fewer assessments are available, for example, creativity and subjective wellbeing this temptation to use inappropriate assessments only increases. The problem can be avoided by limiting assessments to those with substantial evidence of validity and reliability, which depend on instruments being age and developmentally appropriate.

Practical Issues: Feasibility and Cost

In addition to meeting the criteria for validity, reliability, and fairness, a desirable assessment or set of assessments is feasible and affordable. Otherwise, it will not be used or, if used, will create unintended negative consequences. Depending on how information is collected assessments impose costs for children, parents and teachers as respondents or assessors. More detailed and comprehensive assessments impose higher costs on respondents and assessors. In addition there are costs for purchasing assessment tools and training those who administer and use them. If assessment results are made available to people other than those collecting the data there are costs of the systems for storing and sharing data, as well. Finally, the younger the child, the more practical difficulty there is in obtaining information without placing unrealistic demands on the capacities of the child or assessor. As discussed earlier, one of the costs of excessive demands is deterioration in the quality (reliability) of the information obtained. In addition, imposing (unreimbursed) costs on teachers, parents, and others will increase nonresponse rates.

The costs to purchase assessments or the tools for their use are minor compared to the actual costs of training assessors and administering assessments. Yet, policy makers sometimes ignore the latter and act as if there is no cost for administration if teachers (who are already paid) conduct the assessments. This assumption seems especially likely when assessment is formative and integrated into teaching. However, there is always an opportunity cost, and for teachers this can be quite high. The cost of time spent in classrooms collecting, recording, and reviewing assessment information is best measured by the value of the activities that teachers forego as a result--this can be other forms of planning, but may is likely to include direct caring and education of children. Similarly, it should be recognized that parent time is not “free,” and while it is desirable to obtain multiple perspectives, requesting that they provide information imposes opportunity costs on them, as well. Ultimately, the time costs imposed on parents and teachers may result in costs to children by decreasing the time they have to interact with children.

Different types of measures not only have different costs, but differ in who bear those costs. For example, brief direct tests or parent interviews (to obtain ratings) impose some costs on children and parents, but could have substantial costs for specially trained assessors who administer the instruments. Direct tests administered by teachers might be brief individually, but require substantial time if obtained for every child in a setting. Depending on the nature of the test it might be perceived by children as enjoyable (e.g., a game) or stressful. On the other hand, portfolios or rating scales completed by teachers may be collected unobtrusively without interfering with the children’s activities and requiring no parent time or outside staff. However, teacher assessments may require many hours observing children and recording the results rather than interacting children in ways that directly enhance their wellbeing, learning, and development.

Conclusions

Current assessments available offer many choices for measuring children's physical, social, emotional, linguistic, and cognitive development with respect to age, mode of assessment, the source or respondent, and burdens on respondents. There are fewer choices for assessments of executive functions and for some cognitive measures in the areas of math and science. Very few options are available for assessing development in the arts and culture and for approaches to learning; this is primarily done through performance assessments including clinical interviews (conversations and story telling would be included here). Measures that address aspects of approaches to learning including the specific topics of curiosity, creativity, critical thinking, and problem solving are very few. Even rarer are measures of self-esteem, self-efficacy, values and respect, or subjective states of wellbeing such as happiness. To our knowledge, there are not any comprehensive assessments for young children that address these domains. For those domains that are measured rarely or not all by comprehensive assessments, specific assessments are sometimes available.

Clearly, some assessments have stronger evidence of technical adequacy than others. Concerns with technical adequacy are greatest for performance assessments and ratings, particularly in the domains that are not well-covered by tests. The technical adequacy of performance assessments can be improved by standardization of assessment procedures and training of assessors. This has costs, of course.

Policy makers are cautioned to use the results from assessment of young children cautiously. Although progress has been made in measurement of children's LDWB, the progress is not equal in all domains. It is clear that in early childhood there is a need to measure development that is not necessarily assessed in upper elementary and beyond. However, the data collected on young children's LDWB tend to be less reliable as young children develop at vastly different rates and their developmental and learning patterns can be episodic, uneven, and rapid (Bowman, Donovan, & Burns, 2001; Ackerman & Coley, 2012).

Approaches to Assessment

Information on children's LDWB can be collected through a variety of methods, both quantitative and qualitative. Assessments vary in the extent to which they are standardized and in the source (or sources) of their information. Information on children can be obtained directly from children or from those who observe them, most often parents and teachers or other adult caregivers. It may even be obtained from other children (nominations of friends or evaluations of peers to assess relationship status and social skills).

Tests

- Widely used to assessment cognitive abilities, particularly to assess academic achievement in specific content areas.
- Increases the reliability, validity, and the fairness of assessments by reducing assessor (particularly teacher) bias.
- Standardization aims to reduce random fluctuations in the circumstances and procedures, and to eliminate systematic biases by the assessor through variations in procedures as well as subjective judgment.
- Can be group or individually administered. (As our focus is only assessment prior to primary school we consider only individually administered tests; group administered tests are not recommended for children in this age range due to inadequate reliability and validity.)

Performance assessments and qualitative interviews

- Observation of children in their everyday activities is the primary basis for data collection (Dunphy, 2008).
- Typically are embedded in teaching and data are collected continuously during the year and as part of ordinary activities.
- Documentation can include notes, and observation records, artifacts, art, dictation and children's writing, photographs, and video and audio recordings.
- Procedures for conducting and reporting or scoring performance assessments vary from highly standardized to completely unstandardized.
- Some performance assessment systems are linked to specific curricula and provide tools and detailed procedures for data collection and scoring based on rubrics.
- Performance assessments can be scored using a checklist or rating scale (and accompanying rubric) either at one point in time or recorded periodically over a year.

Checklists and rating scales.

- Parents, teachers, or other adults rely on their general knowledge of the child or a brief current observation to answer questions about the child's capabilities, personality, dispositions, behavior, or other characteristics.
- May be standardized in the sense that the precise form and order of the questions has been devised based on research and are not be varied.

Time diaries

- Collect data about children's activities including information about the types of activity, duration of each activity, the place of each activity, and who else was engaged with child as well as what else may have been going on at the same time.
- Provide a unique and very detailed, approach to assessing children's engaged capacities and wellbeing.

The different types of informants

Informants can include parents and other (informal) caregivers, preprimary and primary teachers (we include here all those responsible for the care and education of children in formal settings including some family home care), and health professionals. In addition, children themselves are key informants and can be active participants in their assessment. There are advantages and disadvantages for each informant when assessing young children's LDWB. Informants may provide information directly or professionals specifically trained to administer an assessment may be employed to obtain information, typically from children and parents and other caregivers.

Parents and Caregivers

- Parents and other caregivers are valuable informants because of the intimate knowledge they acquire of a child due to their relationship and the time they spend with the child.
- However, when caregivers are asked to provide ratings relative to an implicit standard or expectation (for example regarding learning, development, relationship quality, life satisfaction or happiness) they may differ greatly from one socio-economic environment and culture to another regarding what is typical or normative (Ertem et al., 2008).
- Caregivers also tend to provide socially desirable answers.
- Despite these disadvantages, caregivers' information about children can be valuable and nationally representative information can be readily obtained through household surveys.

Teachers

- Teachers in preschool settings often provide valuable insights into children's LDWB, though they can only report on those children who attend ECE programs.
- Teachers make good informants because they tend to spend a great deal of time with the children and have working knowledge of and/or training in learning and development.
- However, teachers vary considerably in their preparation and training.
- For many instruments, specialized training of the teacher (or other assessor) may be required.

Health Professionals

- Health professionals have advantages as informants of children's development because of their understanding of how children progress through development and in some instances the health services may be the only professional services available to young children.
- However, for some health professionals, monitoring child development can be a new concept (Ertem et al., 2008).
- The familiarity with child and the level expertise possessed by health professionals can vary by socio-economic context.

The child

- Children are always, in a sense, the basic source of the information on their LDWB. Often, this is indirect and mediated by others. However, young children can provide direct responses in tests, other direct assessments, and interviews. They can be asked to provide ratings. The younger the child, the greater the difficulty of obtaining direct information that is valid and reliable.

Additional Resources

Below are selected assessment resources produced by CEELO. For more information, see CEELO's assessment resource page.

Early Childhood Assessment highlights resources on CEELO's website on topics related to early childhood assessment, including policy reports, FastFacts, webinars, and videos. (Annotated Bibliography)

Formative Assessment: Guidance for Early Childhood Policymakers serves as a guide and framework to early childhood policymakers considering formative assessment, outlining issues for consideration in implementing formative assessment. This guide provides a practical roadmap for decision-makers by offering several key questions to consider in the process of selecting, supporting, and using data to inform and improve instruction. (Policy Report)

Resources to Inform Technical Assistance on Formative Assessment gives recommendations about research and practical resources to inform technical assistance conducted with state education staff on Formative Assessment. (Fast Fact)

State Early Childhood Comprehensive Assessment System: Mapping and Priority Setting Tools (Word) (PDF). Part 1 of this tool is designed to assist State Teams in mapping the current status of efforts to implement a comprehensive assessment system for children from birth through 3rd grade and teacher/classroom and program assessment and evaluation. Part 2 is designed to assist State Teams in setting priorities for planning, implementing and sustaining initiatives in their comprehensive assessment system in the coming year. (Tool)

State of the States Policy Snapshot: State Early Childhood Assessment Policies addresses the questions: What child assessments are required of pre-K and Kindergarten providers? How are child assessment data used? The brief is based primarily on secondary analysis of data collected in the NIEER State of Preschool Yearbook and presents a snapshot of responses to questions about child assessment. (Policy Brief)

References

Ackerman, D., & Coley, R. (2012). *State pre-k assessment policies: Issues and status.* Princeton, NJ: Educational Testing Service.

Atkins-Burnett, S. (2007). *Measuring children's progress from preschool through third grade* (No. 5687). Plainsboro, NJ: Mathematica Policy Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barbot, B., Besançon, M. & Lubart, T. (2011). Assessing creativity in the classroom. *The Open Education Journal, 4*, 58-66.

Barnett, W. S., & Boyce, G. C. (1995). Effects of children with Down syndrome on parents' activities. *American Journal of Mental Retardation: AJMR*, *100*(2), 115-127.

Benítez, I., & Padilla, J. L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52-68.

Berry, D.J., Bridges, L.J., & Zaslow, M.J. (n.d.). *Early childhood measures profiles.* Washington, DC: Child Trends.

Borg, W. R., Gall, M. D., & Gall, J. P. (1989). *Educational research: An introduction* (5th ed.). New York: Longman.

Bowman, B., Donovan, M. and Burns, M. (Eds). (2001). *Eager to learn: Educating our preschoolers.* Washington, DC: Committee on Early Childhood Pedagogy, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy Press.

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* (3rd ed.). Saddle River, NJ: Pearson.

Dunn, L. & Dunn, L. (2006). *Peabody Picture Vocabulary Test, Fourth Edition (PPVT -1V)*. Bloomington, MN: NCS Pearson.

Dunphy, E. (2008). *Supporting early learning and development through formative assessment: a research paper.* Dublin: National Council for Curriculum and Assessment.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3-4), 199-215.

Ettling, D., J. T. Phiri, et al. (2006). *Child Development Assessment in Zambia: A study of developmental norms of Zambian children aged 0-72 months.* Lusaka, Zambia, Ministry of Education, Republic of Zambia.

Fernald, L. C., Kariger, P., Engle, P., & Raikes, A. (2009). *Examining early child development in low-income countries.* Washington DC: The World Bank.

Fink. G., Matafwali, B., Moucheraud, C., & Zuilkowski, S. S. (2012). The Zambian Early Childhood Development Project 2010 Assessment Final Report. Cambridge: Harvard University.

Frongillo, E. A., Tofail, F., Hamadani, J. D., Warren, A. M., & Mehrin, S. F. (2014). Measures and indicators for assessing impact of interventions integrating nutrition, health, and early childhood development. *Annals of the New York Academy of Sciences*, *1308*(1), 68-88.

Glascoe, F.P. (2002). The Brigance Infant and Toddler Screen: Standardization and validation. *Journal of Developmental & Behavioral Pediatrics, 23*, 145-150.

Godfrey, J. R., & Galloway, A. (2004). Assessing early literacy and numeracy skills among Indigenous children with the Performance Indicators in Primary Schools test. *Issues in Educational Research*, *14*(2), 144-155.

Hamilton, S. (2006). Screening for developmental delay: Reliable, easy-to-use tools. *Journal of Family Practice* 55*,* 415.

Hofferth, S. L., & Sandberg, J. F. (2001). How American children spend their time. *Journal of Marriage and Family*, *63*(2), 295-308.

Kim, D. H., & Smith, J. D. (2010). Evaluation of two observational assessment systems for children's development and learning. *NHSA Dialog, 13*, 253-267.

Korkman, M., U. Kirk, et al. (1998). *NEPSY: A developmental neuropsychological assessment.* San Antonio, TX, The Psychological Corporation.

Lau, Sing, et al. (2013). Bicultural effects on the creative potential of Chinese and French children. *Creativity Research Journal, 25*, 109-118.

Matafwali, B. (2010). *The relationship between oral language and early literacy development: Case of Zambian languages and English* (Ph.D. Dissertation in progress). Lusaka, University of Zambia.

Melton GB. Young children's rights. In: Tremblay RE, Boivin M, Peters RDeV, eds. *Encyclopedia on Early Childhood Development* [online]. Montreal, Quebec: Centre of Excellence for Early Childhood Development and Strategic Knowledge Cluster on Early Child Development; 2011:1-8. Retrieved from http://www.child-encyclopedia.com/documents/MeltonANGxp1.pdf.

Merrell, C., & Tymms, P. B. (2001). Inattention, hyperactivity and impulsiveness: their impact on academic achievement and progress. *British Journal of Educational Psychology*, *71*(1), 43-56.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Rossbach, H. G. (1988). *Daily routines of young children.* Paper presented at the Annual Meeting of the American Educational Research Association.

Sheridan, S., & Pramling Samuelsson, I. (2001). Children's conceptions of participation and influence in pre-school: A perspective on pedagogical quality. *Contemporary Issues in Early Childhood*, *2*(2), 169-194.

Slentz, K.L., Early, D., & McKenna, M. (2008). *A guide to assessment in early childhood: Infancy to age 8.* Washington State Office of Superintendent of Public Instruction.

Snow, C. E., & Van Hemel, S. B. (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Snowling, M. J., Hulme, C., Bailey, A. M., Stothard, S. E. & Lindsay, G. (2011). *Better communication research project: Language and literacy attainment of pupils during early years and through KS2: Does teacher assessment at five provide a valid measure of children's current and future educational attainments?* (DFE-RR172A). London: Department for Education. Retrieved from http://dera.ioe.ac.uk/13689/1/DFE-RR172a.pdf

Standards and Testing Agency (2013). *2014 Early Years Foundation Stage Profile Handbook.* London: Department of Education, Standards and Testing Agency.

Stevens, P. A., & Dworkin, A. G. (Eds.). (2014). *The Palgrave Handbook of Race and Ethnic Inequalities in Education*. Palgrave Macmillan.

Teaching Strategies (2013). *Teaching Strategies GOLD Assessment System: Technical summary.* Washington, DC: Author.

Tinajero, A.R., & Loizillon, A. (2012). *Early childhood development and wellbeing. The review of care, education, and child development indicators in early childhood*. Paris: OECD.

William, F., & Monge, P. (2001). *Reasoning with statistics: How to read quantitative research.* (5th ed.). Belmont, CA: Thomson Higher Education.

Woodhead, M., & Brooker, L. (2008). *A sense of belonging. Early Childhood Matters* (No. 111, 3-17). The Hague, The Netherlands: Bernard van Leer Foundation

ABOUT CEELO:

One of 22 Comprehensive Centers funded by the U.S. Department of Education's Office of Elementary and Secondary Education, the Center on Enhancing Early Learning Outcomes (CEELO) will strengthen the capacity of State Education Agencies (SEAs) to lead sustained improvements in early learning opportunities and outcomes. CEELO will work in partnership with SEAs, state and local early childhood leaders, and other federal and national technical assistance (TA) providers to promote innovation and accountability.

For other *CEELO Policy Reports*, *Policy Briefs*, and *FastFacts*, go to http://ceelo.org/ceelo-products.

Suggested citation: Barnett, W.S., Riley-Ayers, S, & Francis, J. (2015). *Measuring child outcomes in the early years* (CEELO Policy Brief). New Brunswick, NJ: Center on Enhancing Early Learning Outcomes.

This policy brief was produced by the Center on Enhancing Early Learning Outcomes, with funds from the U.S. Department of Education under cooperative agreement number S283B120054. The content does not necessarily reflect the position or policy of the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by the federal government.

The Center on Enhancing Early Learning Outcomes (CEELO) is a partnership of the following organizations: