CEELO Short Take: State Approaches to Evaluating Preschool Programs

Shannon Riley-Ayers, PhD and W. Steven Barnett, PhD

June 2015

Consideration of measurement and evaluation systems is an important component for states developing or expanding their preschool programs. When well conceived, these systems not only provide data that are important for continuous program improvement, they also serve as report cards on programs to policymakers and the public. A comprehensive assessment system includes data on the implementation and quality of delivery of preschool services and on the outcomes for children in multiple areas of child learning, development and well-being.

This paper presents guidance for state policy makers for evaluating the quality and effects of a preschool program. The information here will be valuable as states consider monitoring for program quality and continuous improvement as well as conducting a program evaluation for effectiveness. Presenting in a concise manner, this paper is organized into the following key considerations for evaluation planning:

- Study Design
- Sample Size
- Child Outcome Measures
- Classroom Quality Measures

Short Takes are papers and other materials that are designed to be quick resources on key issues of importance to Preschool Development Grant (PDG) states. While geared towards a PDG audience, the information in Short Takes is often of interest to other states and early childhood policymakers .

Study Design

Four options for studying program effectiveness are briefly outlined below. These represent the most often used designs used in preschool evaluation studies.

1. Randomized Trial. A randomized trial is the best approach to evaluation, because it prevents selection bias, a common problem in preschool evaluations. Here children are randomly assigned to either attend preschool or not. This can be done using a lottery to select children for enrollment when there are many more children than can be served by expansion.

2. Nonequivalent Groups, Post-test Only. This study design answers specific questions about children's academic achievement and social skills at kindergarten entry and over time. This study approach examines the children beginning at kindergarten entry and creates two groups—those children who attended the preschool program and those who did not. This can create issues of selection bias where the groups may differ inherently, so that it is a weak approach for measuring a pre-K program's impacts. Nevertheless, this approach provides a look at the same students over time and it is stronger if it uses measures of children's family backgrounds and standardized child assessments that provide a report of academic achievement and social skills over time.

3. Nonequivalent Groups, Pre- and Post-test. The third design, nonequivalent groups with a pre- and post-test, is similar to the above design with the exception that the children are selected and assessed before the beginning of the preschool program. This approach requires either a waiting list for entry into the program, or some other type of screening measure to determine eligibility, so that the two groups (one receiving the treatment of preschool and one not receiving it) can be determined and both groups assessed at this time. This approach can reduce selection bias issues and provides more statistical power for reliable identification of program effects. It requires one additional year of assessments compared to the post-test-only option and a means of identifying children who are not served at preschool entry age rather than later (e.g., a wait list group).

4. Regression-Discontinuity Design. The strongest research design next to a randomized trial is the regression-discontinuity design (RDD). It offers protection against selection bias beyond simply controlling for family background and "pre-treatment" scores in a statistical analysis. One RDD approach employs a statistical model that uses stringent age eligibility cut-offs to define groups. Testing groups using the age cut-offs and then statistically adjusting for age variation reduces the likelihood that selection bias has an appreciable impact on study results. (Another RDD approach is to enroll children below a firm cutoff for eligibility such as 200 percent of the poverty line; this requires accurate information on family incomes for each child.)

For states considering this approach, it is important to have enough children enrolled to provide a large sample to provide confidence in the estimates from an RDD study. For this approach to be most successful, it is best conducted when there is no significant expansion in the provision of the preschool program from the prior year in the communities participating in the study. Confidence in RDD studies is also increased if children can be assessed very early in the school year. RDD studies can be combined with an ordinary nonequivalent comparison group longitudinal study of children who did and did not attend preschool. In this case, the children are followed from kindergarten to third grade. The combination of these two approaches enables investigators to assess the extent to which selection bias affects the estimates at kindergarten entry. If sufficient family background and school attendance data can be obtained, it may be possible to model some of the differences that contribute to the bias and thereby reduce that bias. The non-equivalent group longitudinal design adds to the study the ability to estimate effects on social and emotional development at the end of kindergarten, and to estimate effects on achievement, grade retention, special education, and socio-emotional development into the elementary years. These outcomes are important for increasing confidence in estimates of the economic value of the program's benefits.

Sample Size

There is no hard and fast rule for minimum sample size for all circumstances. What is important is that before beginning a study, you calculate the minimum sample size needed to provide adequate power to detect the effect size (i.e., program impact) you expect to produce or consider acceptable. State administrators may wish to obtain the assistance of an expert to calculate the necessary sample size, as it depends on the research design and assessments chosen as well as your specific research and evaluation questions. However, there are tools that can be used to calculate the necessary sample and guides for their use should you wish to do this yourself, for example *PowerUp*.

As concrete examples of studies with adequate sample size, we reference several recent reports. A regression discontinuity study of the Boston's preschool program included 2045 children in 69 schools. A randomized trial of Tennessee's Voluntary Pre-K program used a sample of 3000 children in 80 schools. These studies are designed to provide estimates of program impacts for subgroups as well as the entire population served and can detect even modest effects. In contrast, most studies in the IES Preschool Curriculum Evaluation Research Consortium (PCER) had sample sizes in the range of about 200-300 children in 14 to 40 classrooms. The PCER studies used pre-tests to increase their statistical power, but nevertheless lacked statistical power to detect moderate sized differences among classrooms and some impacts on children that would be considered meaningful. Although most studies employ equal sample sizes for treatment and controls, it should be noted that this is not required and there is little loss of statistical power for a control or comparison group up to twice as large as the treatment group.

Child Outcome Measures

A research study design dictates the strength of the results, but the information collected and assessments chosen as part of that study will determine the content of the results and affect the reliability and validity of the study results.

Children's Background

It is desirable to have as much background data on children when they enter preschool as possible for those in the new program, and a comparison group of children in other programs or not attending pre-K. Background measures of particular importance are: a pre-test measure of children's abilities, parent education, income (free lunch status), ethnicity, and home language.

Children's Learning, Development, and Well-being (LDWB)

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. Often, it is desirable to collect data from the child in his or her home language wherever possible.

Assessment Methods

In education, the first type of assessment that comes to mind for many people is standardized tests. However, these are just one method of assessing children's learning and well-being. This section outlines the characteristics of standardized tests, checklists and rating scales, and performance-based assessments.

Standardized tests. Standardized tests are widely used for assessment of cognitive abilities, particularly to assess academic achievement in specific content areas. Standardization refers not just to the instrument itself, but also to the process of its administration. It 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. Tests for young children are generally administered one-on-one by a trained assessor. Tests provide standardized and often norm-referenced information and often cut across larger age spans. There are also some disadvantages to using a one-time test. Young children's learning is often not accurately captured in one snapshot. Also, tests can be expensive and sometimes require training to administer. Tests for young children require time to be administered because they are generally administered one-on-one. A few examples of tests are listed next.

- The Woodcock Johnson Achievement Test is a nationally normed standardized test that has been widely used in preschool evaluation studies. It has several sub-tests that identify specific skills such as letter-word identification and applied problems for math that capture growth in young children's cognitive development.
- The Peabody Picture Vocabulary Test (PPVT IV; Dunn & Dunn, 2007) is a norm-referenced standardized measure of receptive vocabulary from age 2.6.
- Short tests with specific tasks for children like Head-Toes-Knees-Shoulders Task (HTKS, Ponitz, McClelland, Jewkes, Connor, Farris, et al., 2008) and Peg Tapping Task (PT; Diamond & Taylor, 1996) are used to assess self-regulation. Similarly, the test Dimensional Change Card Sort (DCCS; Zelazo, 2006) is a measure of cognitive flexibility.

Checklists and rating scales. A second type of assessment frequently used is the checklist or rating scale. 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. However, in this section we refer to measures that do not necessarily require continuous data collection over time (and are summative in use). Instead, 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. Such assessments 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. These types of assessments are valuable because they do not require any training and are often not too time-consuming to complete. However, they are reported by adults and may not necessarily accurately represent the child.

These report measures from caregivers or teachers often report social skills and adaptive behaviors: The Social Skills Rating System (Elliott, 1990); Behavior Assessment System for Children, Second Edition (BASC-2; Reynolds & Kamphaus, 2004); and Vineland Adaptive Behavior Scales (Sparrow, Cicchetti, & Balla, 2005) are examples of report measures.

Performance-based assessments. Another broad type of assessment is performance, or authentic, assessment for which observation of children in their everyday activities is the primary basis for data collection (Dunphy, 2008). These assessments typically are embedded in teaching and data are collected continuously during the year and as part of ordinary activities. Documentation can include notes, observation records, artifacts, art, dictation and children's writing, photographs, and video and audio recordings. The documentation obtained can be collected and organized in portfolios for each child. These approaches seek to maintain the whole-child perspective and recognize the inter-relatedness of children's dispositions, habits, skills, and knowledge, as well as the importance of context for understanding children's LDWB. Often, these measures are already in place in programs, which makes them an easy and valuable source of information, but they require training and support to implement. However, the reliability of this approach must be monitored carefully as it is completed by teachers, so quality checks of the data are important.

Teaching Strategies GOLD and the Early Learning Scale (ELS) from NIEER are two examples of performance-based observation assessment tools that examine multiple domains of learning and development.

Classroom Quality Measures

The quality of the program is generally examined by observation by outside observers trained to reliability on a battery of classroom observation instruments that examine several aspects of the classroom. By including this classroom-level aspect of an evaluation, information is summarized to examine such questions as:

- How does quality of preschool differ across auspice, provider or teacher?
- What is the impact of quality of the preschool experience on student outcomes?

Classroom observations are generally conducted once per year during the mid-year to assess quality. Observers should be trained to reliability and should be monitored for observer drift. Below are examples of classroom observation tools. The list is not exhaustive, rather provides examples of tools used in published evaluations.

- The Third Edition of the *Early Childhood Environment Rating Scale* (ECERS-3; Harms, Clifford, & Cryer, 2014) measures environmental factors as well as teacher-child interactions that affect the broad developmental needs of young children. It also emphasizes the role of the teacher in creating an environment conducive to developmental gains.
- The *Classroom Assessment Scoring System* (CLASS; Pianta, LaParo & Hamre, 2008) provides information that focuses specifically on teacher interactions and other features of instruction.
- The *EduSnap Classroom Observation* (Ritchie, Weiser, Mason, & Holland, 2015) quantifies students' experiences of their school day by providing an in-depth look at how students experience their day by recording the actual time they spend in activity settings, content areas, student learning approaches, and teaching approaches.
- The *Early Language and Literacy Classroom Observation* (ELLCO; Smith, Brady, & Anastasopoulos, 2008) identifies the practices and environmental supports that nourish children's literacy and language development. This observation focuses on important pre-literacy activities like storybook reading, circle time conversations, and child-originated story writing.

- The *Preschool Classroom Mathematics Inventory* (PCMI; Frede, Weber, Hornbeck, Stevenson-Boyd & Colon, 2005) This tool measures the materials and strategies used in the classroom to support children's early mathematical concept development, including counting, comparing, estimating, recognizing number symbols, classifying, seriating, geometric shapes, and spatial relations.

Important Considerations

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.

All data that are collected should be interpreted carefully. They should be collected and evaluated to make program decisions, to inform policy, and to guide instruction. Using multiple measures to generate an understanding of a program is recommended.

Guidance for State Policy Makers

This paper briefly presents options and considerations for state policy makers executing the evaluation of the effectiveness of preschool programs.

Recommendations:

- Conduct regular evaluations of programs and policies implemented in early childhood education. With a new program or policy, build up evaluation gradually by starting with the collection of data to establish a baseline (how are children and programs doing prior to the new policy or program). The next step is follow-up with process evaluations to assess quality of implementation. Child outcomes might be tracked to a get a general sense of whether they are moving in the right direction. However, rigorous child outcome evaluation is best reserved until after a program or policy has been found to be reasonably well implemented which may take a few (or even more) years. Patience is a virtue in evaluation.

- Select the most rigorous study design and the largest sample that is possible given the context (e.g., program design and eligibility criteria, funding). As there is always some uncertainty about the required sample size, it can be useful to plan for potential additional waves of data collection over additional years. Moreover, because unexpected events can cause any single year to be unusual, it is useful to spread the sample over multiple cohorts of children.

- Work collaboratively with a qualified contractor or consultant when planning and carrying out evaluations. Good evaluators engage with those administering the program in designing and implementing the evaluation so that it is fully informed by those who will use the information from the evaluation and are most knowledgeable about the program.

- Select measures of child outcomes and classroom quality that link directly to the program standards and goals and to policy makers' most critical questions. A broad set of measures of children's learning and development are likely to be more predictive of later life outcomes than narrow measures that focus only on literacy and mathematics. Assessing non-English speakers in their home language as well as in English is best practice.

- Engage policy makers and practitioners in interpreting program evaluation data to inform practice and policies in the context of both local knowledge and the broader body of scientific knowledge regarding learning and development and early education.

For Further Information

Riley-Ayers, S., Frede, E., Barnett, W. S., & Brenneman, K. (n.d.). *Improving early education programs through data-based decision-making.* New Brunswick, NJ: National Institution on Early Education Research. Available: http://nieer.org/pdf/Preschool_Research_Design.pdf

The National Early Childhood Accountability Task Force. (2007). *Taking stock: Assessing and improving early childhood learning and program quality.* Available: http://www.pewtrusts.org/en/research-and-analysis/reports/2007/10/31/taking-stock-assessing-and-improving-early-childhood-learning-and-program-quality

Dynarski, M. & Kisker, E. (2014). *Going public: Writing about research in every day language.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Available: http://ies.ed.gov/ncee/pubs/REL2014051/pdf/REL_2014051.pdf

Additional Explanation for terms used in this paper

Selection Bias

Selection bias becomes an issue when using groups that self-select to participate in the preschool program and/or who are selected based on program eligibility criteria. Selection bias in this case relates to the concern that the two groups could be inherently different *in ways that are not measured and are related to children's learning and development* at the start of the study before the treatment (the preschool program) is provided. If such bias exists, the estimates of the effects of the program are likely biased. Selection bias could, for example, be due to systematic differences between the educational aspirations of parents of the two groups. This would undoubtedly lead to the groups performing differently in school even if there were no difference in their preschool program participation. In studying preschool programs that serve disadvantaged children, selection bias most often appears to underestimate program effects.

Child and Family Characteristics

Data both on child and family characteristics of the sample children are necessary for matching samples. This information is generally collected through short family interviews done at the time of enrollment or at another time by phone. The interview is conducted in the family's home language.

The data that are essential include the following:

- Maternal education level;
- Primary language spoken in the home;
- Family income level; and
- Confirmation of preschool attendance status already collected from school records (if applicable)

Additional data that can be collected, but often have limited utility in analyses include information about:

- The child's health and hospital stays;
- The child's dental care;
- Number of siblings;
- Extended family members such as step members;
- Parent involvement in school such as attending parent-teacher conferences; and
- How many times the family has moved

Validity

Validity is a fundamental criterion for selecting instruments to measure Children's Learning, Development, and Well-being (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 in 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

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 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

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 well-being, 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.

References

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.

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

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

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental psychobiology*, *29,* 315-334.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test.* Bloomington, MN: Pearson Assessments.

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

Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System: Manual.* Circle Pines, MN: American Guidance Service.

Frede, E., Weber, M., Hornbeck, A., Stevenson-Boyd, J., & Colón, A. (2005). *Preschool Classroom Mathematic Inventory (PCMI).* New Brunswick, NJ: National Institute for Early Education Research.

Hancock, G. R., & Mueller, R. O. (2010). *The reviewer's guide to quantitative methods in the social sciences.* New York, NY: Routledge.

Harms, T., Clifford, R. M., & Cryer, D. (2014). *Early Childhood Environment Rating Scale.* New York, NY: Teachers College Press.

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.

Pianta, R. C., La Paro, K.M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) manual: K-3*. Baltimore, MD: Paul H. Brookes Publishing Company.

Ponitz, C. E. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly, 23*(2), 141-158.

Reynolds, C. R., & Kamphaus, R. W. (2004). *BASC-2: Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Ritchie, S., Weiser, B., Mason, E., & Holland, A. (2015). *The EduSnap Professional Learning System.* Durham, NC: Snapshot.

Smith, M.W., Brady, J.P., & Anastasopoulos, L. (2008). *Early Language & Literacy Classroom Observation (ELLCO) Pre-K tool.* Newton, MA: Paul H. Brookes Publishing Co

Sparrow, S.S., Cicchetti, D.V., & Balla, D.A. (2005). *Vineland Adaptive Behavior Skills – Second Edition (Vineland-II).* Minneapolis, MN: Pearson Assessment.

Williams, F., & Monge, P. R. (2001). *Reasoning with statistics: How to read quantitative research.* Fort Worth, TX: Harcourt College Publishers.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377–392.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1*(1), 297-301.

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: Riley-Ayers, S. & Barnett, W.S. (2015). *State approaches to evaluating preschool programs.* New Brunswick, NJ: Center on Enhancing Early Learning Outcomes.

This resource 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: