MATHEMATICA
Policy Research, Inc.

Measuring Children's Progress from Preschool Through Third Grade

Prepared for The National Early Childhood Accountability Task Force

This paper was prepared with support from The Pew Charitable Trusts, the Foundation for Child Development, and the Joyce Foundation

Sally Atkins-Burnett

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Ave. S.W., Suite 550
Washington, DC 20024-2512
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

MEASURING CHILDREN'S PROGRESS FROM PRESCHOOL THROUGH THIRD GRADE

More and more states are answering the call to provide preschool programs for children, particularly those at a higher risk of academic failure. By the 2004-2005 school year, 38 of the 50 states had funded programs for four-year-old children (Barnett, Hustedt, Robin, & Schulman, 2005). By the end of end of 2006, state spending on such programs for these children was more than three billion dollars with almost 950,000 children enrolled (Barnett et al., 2007). With such large investments and so many children involved,, states want to be assured that this investment is making a positive difference in children's readiness for school.

On average, children who have preschool experiences enter kindergarten with more academic skills than those who do not, but preschool does not ensure that all children will have the skills needed for success in kindergarten (Denton, Germino-Hausken, & West, 2000). Variation in quality of care and support for learning (i.e., different instructional opportunities) is associated with different outcomes for children. (Belsky et al., 2007; Burchinal & Cryer, 2003; Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002; National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network [ECCRN], 2000; Peisner-Feinberg et al., 2001). Evaluating the effectiveness of preschool programs in supporting children's readiness is challenging for states. Evaluation efforts are hindered by concerns about what should be measured and how it should be measured..

Policymakers are concerned not only about the immediate effects of this investment, but also the long-term effects on the performance of children in elementary schools and beyond. However, congruence or alignment between beginning elementary school expectations and what happens in preschools is limited (Scott-Little & Martella, 2006). Monitoring the effectiveness of programs in preparing children for success in kindergarten and beyond requires that the assessed criteria be aligned in ways that allow examination of change over time. Assessments need to be aligned with state standards, that in turn should align vertically, horizontally, and temporally (Kauerz, 2006).

This paper will discuss the measurement of child outcomes in the context of evaluating the effectiveness of preschool programs for children. Little is known about how individual districts and states are evaluating early childhood programs, so this discussion will highlight some of the ways in which this challenge is being addressed. After a brief discussion of the importance of focusing on the whole child rather than just their language and cognitive domains, most of the paper will explore what is known about current assessment methods used with young children. Problems related to relying solely on traditional, on-demand standardized tests to assess achievement of young children will be explained. Although young children who are English Language Learners (ELL) represent an increasing proportion of preschool children, it is beyond the scope of this paper to discuss in-depth the issues involved in assessing these children (see Lazarin, 2006 for some discussion of K-12 efforts). Observational measures that span the preschool to elementary age range offer an alternative to direct testing. The use of these measures in formative evaluation efforts will be discussed with the caution that high stakes should never be attached to these measures. Using a multimethod approach would provide a richer portrayal of children's performance. Innovative and alternative approaches to assessment used by some

states will be highlighted, and concerns about reliability of teacher judgments discussed. The paper concludes with a brief discussion of measuring classroom quality and recommendations for next steps.

CONTENT OF ASSESSMENTS

Early childhood traditionally has assessed children by developmental domains examining key expectations or milestones in cognitive, social-emotional, language, approaches to learning, and fine and gross motor development. In elementary and secondary schools, the focus often shifts to an examination of specific academic areas, with an emphasis on literacy, mathematics, and science. With the stakes for academic achievement increasingly high at the elementary level, this emphasis on cognitive development has led to a similar narrowing of focus in preschool assessments, and little attention has been paid to the interdependence of other types of development in early childhood. However, a child's readiness for success in school is dependent upon more than their cognitive abilities, so social-emotional, motor, and other developmental areas also should be assessed for this age group (Hair, Halle, Terry-Humen, Lavelle, & Calkins, 2006).

Few measures are available for direct measurement of physical development in the preschool years. Direct measures of motor development are often a part of screening instruments, and usually require space (for gross motor assessments) and equipment. This area seldom receives attention in evaluating preschool environments.

Evidence for the importance of approaches to learning and the social-emotional domain in early development have continued to build in the past decade (Agostin & Bain, 1997; Hauser-Cram, Warfield, Shonkoff, & Krauss, 2001; Henricsson & Rydell, 2006; Ladd, Birch, & Buhs, 1999; Ladd, Kochenderfer, & Coleman, 1997; Meltzer et al., 2004; Pianta, Nimetz, & Bennett, 1997; Raver, 2002; Raver & Knitzer, 2002; Rubin, Coplan, Nelson, Cheah, & Lagace-Seguin, 1999; Shonkoff & Phillips, 2000; Sroufe, 2005; Sroufe, Egeland, Carlson, & Collins, 2005; Tur-Kaspa, 2004). A child's ability to regulate his or her emotions and attention, to persist in the face of challenges, to approach learning with interest and enjoyment, to form friendships, and to interact positively with others are among the skills that have been found to be related to academic as well as social-emotional outcomes. For example, two groups of preschool children with average cognitive ability but different levels of social skills were followed through first grade and had different academic outcomes that year: the children with higher social skills scored significantly higher on tests of academic achievement (Konald & Pianta, 2005). Alternatively, the absence of social-emotional skills and/or presence of problem behaviors such as aggression, hyperactivity, and bullying are related to negative academic as well as social outcomes (Le, Kirby, Barney, Setodji, & Gershwin, 2006). Too much emphasis in preschool programs on cognitive development with too little attention to social and emotional development could lead to negative outcomes for children.

We need to acknowledge the effects that testing can have on curriculum and instruction. An unintended consequence of gathering information solely on academic outcomes is that parents, teachers, and program administrators may not pay enough attention to other critical areas of development (National Research Council, 2001). This may be particularly true for social-emotional development and approaches to learning. Recent longitudinal research suggests that early childhood environments can have long-term negative effects on children's social and emotional development even when the quality of those early environments is rated positively (Belsky et al., 2007). Because more children are spending time in group environments, it is important that we evaluate their social, motivational, and emotional development.

Although the value of examining social-emotional development is clear, the methods for examining these areas are more complex and less developed than the methods for examining early cognitive and

language development. A full discussion of measuring social-emotional development is beyond the scope of this paper, although some instructional measures that include teacher reports of children's social-emotional development will be described. For further discussion of this important topic, see Denham, 2006; Keith & Campbell, 2000; Ladd, Herald, & Kochel, 2006; Meisels, Atkins-Burnett, & Nicholson, 1996; Printz, Borg, & Demaree, 2003; and Raver & Zigler, 2004. Information about reliability and validity evidence for some measures used in research as well as published measures is available on the internet (Berry, Bridges, & Zaslow, 2004). Additional direct measures, particularly of self-regulation, are currently under development and will be important additions to our understanding of children's ability to benefit from the learning environment (Blair & Razza, 2007; Carlson, 2005; Denham, 2006; Riggs, Blair, & Greenberg, 2003; Smith-Donald, Raver, Hayes, & Richardson, in-press).

DIRECT ASSESSMENTS

Norm-referenced on-demand standardized tests are the most commonly used assessments in program evaluation and accountability efforts. They provide a common framework for making comparisons among programs and children, and can be administered by an outside evaluator providing more objectivity to the measurement. However, direct assessments can be problematic for measuring outcomes with young children. They are not valid for all children, often lack congruence with curriculum, and have added measurement error in young children.

Group Administration. Direct assessments usually are administered individually to young children, although some group administered assessments are available for early elementary school. Districts often prefer the standardized tests because they consider them more objective and consider it more cost-effective to administer a group assessment than to test children individually in first through third grade. However, there are problems with this approach. Although by first grade, differences in the reliability of a group- versus individually-administered standardized test are not detectable in the standard errors, questionable validity is evident in observations of children taking the tests (Atkins-Burnett, Rowan, & Correnti, 2001). Even though children in first grade receive much of their instruction in group settings, the group administration of tests leads to behaviors that increase both the number of omissions (skipped items) and the frequency of multiple answers on items, even when tests are given in smaller group settings (Atkins-Burnett, Rowan, & Correnti, 2001). These problems with attention to task and staying on the correct item lead to underestimates of the ability of the middle- and lower-performing children.

Given these problems, tests for children in kindergarten and first grade include more items to assess a specific area than would be necessary with older children, In order to attain adequate reliability estimates. For example, on the TerraNova (CTB/McGraw Hill, 1997), one of the most widely used standardized assessments in elementary schools, the mathematics form for third graders has 30 items, while the form for first graders has 47 items. Shorter survey forms are available only for third grade and beyond, while first and second graders need to complete the lengthy basic test battery. These longer tests tax young children who experience fatigue and lose focus when responding to the unfamiliar format of standardized assessments. In addition, these group-administered assessments are grade specific and often have problems with the ceiling and floor of the tests.

The best measurement on a test occurs when the items are targeted specifically to a child's ability. Assessments that are group-administered work best with children who are average, that is, in the middle of the scale. Information about children who are most at risk for academic failure—typically those in poverty with more limited experiences and less opportunity to learn outside of school—is sparse and less reliable because the measurement error is greater at the ends of the distribution.

At both ends of the distribution, the item gradients often are very steep, making it difficult to assess progress reliably. On some standardized measures, a difference in performance on one or two items can

cause large changes in standard scores at the ends of the continuum (Meisels & Atkins-Burnett, 2000). This is particularly problematic on grade-specific (usually group-administered) assessments. For example, on the Terra Nova (CTB/McGraw Hill, 1997), a maximum score on the mathematics test in first grade would assign a child a standard score of 680. If, in second grade, the same child got one item incorrect, he or she would have a standard score of 646–661 (depending on which item was missed). The difference between the two scores makes it appear that the child has lost skills. In the middle of the first- or second-grade mathematics scale, a correct or an incorrect response to one more item means an average difference of about 5 standard score points. At the low end of the scale, missing one item (or getting an additional item correct) can change a score by more than 40 standard score points.

In addition to questions about reliability, the overlap and vertical alignment (grade-to-grade alignment of content) between forms on standardized tests at successive grade levels often is poor to nonexistent. This is particularly problematic when assessing children at either end of the continuum, because there may not be items available to measure where they are on the scale. At the lower end of the continuum, there are those who may be working on skills or behavior learned by others the previous year, while at the upper end, there are those who may be improving their knowledge from the year beyond their current placement.

Group tests often are administered by the classroom teacher, and in these cases limited availability of alternate forms and the security of test forms become additional issues. It is difficult for many teachers to understand that the tests are designed to sample information and behaviors that are representative of the behaviors and knowledge in a domain. If the specific information in a particular items is taught, those items are no longer representative of the domain and so are no longer a good measure of the child's ability in that domain. When the stakes are high, the temptation to teach to the test items also is high (Domenech, 2000; Pedulla, et al. 2003; Porter & Olson, 2003; Stecher & Barron, 2001). Teaching the specific content of the test has been reported more frequently among elementary school teachers than among middle or secondary school teachers (Pedulla et al., 2003), suggesting that it could be very prevalent among early childhood teachers.

Individual Administration. In preschool and kindergarten, individual test administration is recommended and is most typical. Several different types of assessments are available including content specific assessments (such as the Test of Early Reading Ability–3rd Edition [TERA–3; Reid, Hresko, & Hammill, 2001]; and the Test of Early Mathematics – 3 [TEMA-3; Ginsburg & Baroody, 2003), some curriculum-based measures (such as the Individual Growth and Developmental Indicators [Missall & McConnell, 2004; Missall, McConnell, & Cadigan, 2006]; Dynamic Indicators of Basic Early Literacy Skills [DIBELS; Good & Kaminski, 2002]), standardized performance-based probes (such as the Early Literacy Advisor [ELA; Bodrova & Leong, 2001], and adaptive assessments (such as the Woodcock-Johnson Tests of Achievement III [WJ III; Woodcock, McGrew, & Mather, 2001]). Longer content-specific tests such as the TERA-3 and the TEMA-3 can take 45 minutes to administer to a single child and provide information on only one content area, so these are seldom used in large-scale program evaluations.

Curriculum-based measures (CBM) are designed to be administered frequently as ongoing monitoring tools. CBM, such as Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Kaminski & Good, 1996) and Get It, Got It, Go (McConnell, 1998), typically are fluency measures that can be administered quickly (less than five minutes) and have multiple items or forms. They are designed as quick indicators of children's status and growth, rather than comprehensive measures (McConnell, Priest, Davis, & McEvoy, 2000), and they are created for areas that have been predictive of later outcomes, such as vocabulary and letter naming. CBM typically are administered by the classroom teacher, but sometimes

by outside examiners. They are more prevalent for the assessment of language and literacy skills than for math skills in early childhood, although new efforts are being made to create CBM for this domain as well (Fuchs, Fuchs, Compton, Bryant, Hamlett, & Seethaler, 2007; VanDerheyden, Broussard, Fabre, Stanley, Legendre, & Creppell, 2004). As these are developed, it will be important for them to examine concepts that have predictive validity, such as number constancy, magnitude judgments, and other number concepts and applications (Fuchs et al., 2007; Mazzocco & Thompson, 2005). Some research suggests that single time-points of CBM, even at third grade, are not reliable enough to be used in evaluation, and that CBM with third graders lack evidence of validity unless more than one administration is aggregated (Jiban & Deno, 2007).

Get It, Got It, Go (GGG; McConnell, Priest, Davis, & McEvoy, 2002) is a CBM designed specifically for 3 to 5 year old children. It assesses expressive vocabulary (picture naming), rhyming, and initial sounds (alliteration). CBM such as GGG are evaluated according to "the extent to which they (a) measure important outcomes for children; (b) can be used efficiently and economically; (c) are standardized and replicable; (d) rely on generalized or 'authentic' child behaviors; (e) are technically adequate; and (f) are sensitive to growth and change over time and to the effects of intervention" (Missall & McConnell, 2004, pp. 3-4). GGG tasks are free, are designed to be administered in five minutes. The developers report evidence of test-retest reliability (r=.44 to .89), moderate to strong concurrent validity with other measures (r=.56 to .79), and ability to show developmental changes (correlations of .41 to .60 between children's scores and chronological age; growth curve analysis found 76% of variance in a child's score was related to chronological age) (Missall, McConnell, & Cadigan, 2006).

One concern about CBM is that the timed aspect of the administration may increase measurement error for some children. For example, the picture naming task involves using 4 of approximately 120 picture cards to teach the task of naming a picture and the remaining cards are shuffled and shown to the child one at a time (random sample of cards presented each time). The score is the number of cards that the child correctly names in one minute. Children who have more difficulty with processing, and those with more limited vocabulary would be expected to have lower scores. However, sometimes children who are highly verbal score poorly because they want to talk about each picture as it is presented, rather than just name it. As a result, they are not able to name enough pictures in the allotted time to achieve a score that reflects their extensive vocabulary. Inexperienced assessors may have difficulty in keeping some children on task. In addition, more research is needed (in particular, about the meaning of CBM scores and growth, number of time points needed for reliable measurement in preschool; and timeframes for data collection) before CBM can be helpful for program evaluation.

Standardized Performance-based Assessments that involve standard probes for tasks administered to children are more common in the elementary school years; there are no direct performance-based measures that span preschool to third grade. The Early Literacy Advisor (ELA; Bodrova & Leong, 2001) is a preschool to kindergarten measure has been used as an accountability measure for kindergartens by at least 30 districts (Bodrova & Leong, 2001). The ELA is unique in its use of technology to provide feedback and recommendations to teachers based on the child's performance. However, because it is designed for children who are 4 to 6 years old, its use in elementary schools is limited.

Adaptive assessments are designed to measure children's knowledge and skills longitudinally. These tests present items in order of difficulty, and most of them establish starting and stopping rules for children based on the child's performance on the tests, thus targeting items to the child's level. This allows the test to be brief enough that the child's attention and fatigue do not interfere with the reliable assessment, while still providing enough items for strong measurement of an area. The measurement error on an instrument is lowest when the items are targeted to the child's ability. Many large-scale studies use instruments that are adaptive; usually they use ceiling and floor rules, such as those found in

the scales from the Woodcock-Johnson Tests of Achievement III (WJ III) and the Peabody Picture Vocabulary Test Fourth Edition (PPVT-4). The WJ III and the PPVT-4 are two of the few instruments that have established norms for children age two through elementary school (and beyond), so large-scale studies of early childhood often use the PPVT and the WJ III Letter-Word Identification and Applied Problems tests. The most recent revision of the WJ III was standardized with a larger sample of preschool children than the previous version (N=963 children 2 to 4 years old), increasing the reliability of the norms. The most recent PPVT-4 has revisions that improve its appropriateness for younger children. The newest version is in color and improved the representation of word types (such as naming, categories, and attributes) across levels of difficulty, thus providing a stronger indication of children's general knowledge and cognition.

Because they target children's skill levels, each of these adaptive tests can be administered to young children in less than 10 minutes, but they do not address the full range of language and literacy or mathematics goals included in state and national standards, nor do they capture skills and processes in other academic content areas. These scales often have been found to be predictive of school achievement (although the predictive validity is not as strong from preschool estimates as from those in elementary school), making them useful for policymakers who want to know the likelihood that children will be successful in school. However, the tests also show strong associations with socioeconomic status and may not reveal whether a preschool program is supporting children across the developmental domains. The tests also do not provide information to teachers and programs about areas of strengths and weaknesses in the curriculum or in the children's development. The dilemma is that more comprehensive, longer assessments tax the energy and attention span of young children; currently available standardized tests are either not comprehensive enough, or they are too lengthy (for reviews, see U.S. Department of Health and Human Services, 2002).

The challenge in creating a comprehensive measure that does not take an inordinate amount of time to administer is to find items in the different strands or areas of learning that also tap a range of levels of difficulty. By including different types of words across the range of difficulty, the PPVT-4 has made a step forward in increasing the comprehensiveness of the preschool items. To create adaptive measures of development that are brief, yet comprehensive, one needs to use items that are different from those found in a diagnostic test. On a diagnostic test, the goal is to assess areas in which a child is strong or weak, so the items must be independent of performance in other areas. For short comprehensive measures of development, the most desirable items assess more than one area in that domain (for example, number concept and spatial reasoning).

Given the growth in early childhood assessment efforts in the past decade, we are in a strong position to create shorter, more comprehensive, adaptive cognitive assessments. Item response theory (IRT) (Embretson & Hershberger, 1999; National Research Council, 2001; Van der Linden & Hambleton, 1997) has made it possible to improve the way we measure young children by targeting items to their ability level. IRT uses information on all of the items and all of the children's responses to estimate, through an iterative process, item difficulty and the ability of the child with respect to the domain being measured. The PPVT-4 and the WJ III used IRT to design the tests. Information from IRT analysis of items allows the creation of comparable alternate measures, and allows us to compare children who take different versions of tests.

The PPVT-4 and WJ III use starting and stopping rules to target items. An alternative way to develop an adaptive assessment is to use a two-stage design similar to that used in the measures

developed for the Early Childhood Longitudinal Studies.[1] Adaptive testing was used both in the Early Childhood Longitudinal Study—Birth Cohort (ECLS-B) and the Early Childhood Longitudinal Study—Kindergarten Class of 1997-98 (ECLS-K) to assess a range of skills based on national standards. The ECLS-B recently conducted a study to place the items from ECLS-K and ECLS-B on the same scale. This will allow continued longitudinal assessment of the children of ECLS-B as they enter elementary school and also will allow a comparison of the ECLS-B children as kindergartners with the ECLS-K kindergartners. Using a similar procedure, two-stage adaptive assessments could be developed that have adequate overlap of items and could be administered to all children. This would allow us to measure performance across the entire early child period. A bank of comprehensive items could be developed and field-tested with different age/grade levels to create a strong longitudinal scale. However, this is an expensive development process and would most likely require a pooling of resources as states collaborated on this endeavor.

In developing these assessments, one must consider additional factors. For instance, it is faster and easier to administer assessments to young children if the items allow them to choose among several answers. On the other hand, this approach usually requires a more complex three-parameter measurement model[2] and larger samples of children to develop the scale. Analysis of differential item functioning also is needed to ensure that the instrument is valid for the diverse population of children in the United States. To be used for evaluation or accountability, the measure needs to be kept secure so that the actual test items are not taught. Using outside evaluators to administer the assessments to the children would help both to maintain security and guard against effects of assessor bias; however, unfamiliar assessors can present challenges to young children.

Using IRT allows the creation of an interval scale. With the collection of data at multiple time points, growth curves can be examined. It is these changes in the growth curve that is of particular interest in evaluating the effectiveness of preschool programs, so a sample of children should be followed for at least three points of measurement with the same adaptive measure. Matrix sampling of children and/or domains assessed (Childs & Jaciw, 2003; National Research Council, 2001; Popham, 1993) can help to contain the cost of this approach so that financial resources can be invested in making better use of curriculum-based and observational assessments in a multimethod approach to evaluation. Although measurement of individual students is poor if matrix sampling of items is used and thus individuals cannot be compared, such sampling provides a broader level of information about the quality of the curriculum than standardized assessments of preschool children currently used in evaluation and accountability efforts.

CHALLENGES AND CONCERNS ABOUT DIRECT ASSESSMENTS

For young children, the validity of direct, on-demand tests must be considered against what is known about the child from other sources (such as the observational measures discussed later). A score on these direct tests may tell only the extent to which the child is familiar with a given type of question or task, or has the ability to stay focused on the task (Meisels & Atkins-Burnett, 2006; Meisels, 1994). Young children have a more limited response repertoire – preschool and kindergarten children are more apt to show than tell what they know (Scott-Little & Niemeyer, 2001). They also may have difficulty responding to situation cues and verbal directions, particularly the more complex, multi-step directions

[1] Two-stage adaptive tests involve a small initial set of items administered to all children that are used to target the specific level of those who will be assessed more carefully in the second stage. This procedure was used in the ECLS-K direct assessments of children's cognitive and academic skills.

[2] Three-parameter models adjust for different discrimination of items and for the added measurement error involved in multiple choice tests (probability that a child guessed correctly).

(Meisels & Atkins-Burnett, 2006). *What* is being measured may be confounded by *how* it is being measured. Young children are not familiar with the structure of test questions, and test formats pose cognitive demands that may be unrelated to the criteria being assessed. Children may not understand what it means to weigh alternatives; for example, when questions ask them to 'choose the best answer,' young children may choose the one that is most attractive to them even if they know it is not the correct answer to the question. Also, language demands may obscure what is being assessed. Young children may not be able process negatives or subordinate clauses, or they may focus only on the last part of a question. While these cognitive demands relate to a child's ability to process language, they do not say anything about a child's knowledge in the content being assessed when other areas are the focus.

Temperament and experience also can influence a child's performance on standardized tests. For instance, many parents who have asked their preschool child to do something as simple as saying "hello" or waving to someone will tell you that sometimes the child will do it, and sometimes they won't. In addition, many young children will respond or not to a question or a demand, depending on their relationship with the adult who is asking the question. A child who responds readily to a parent may not respond as readily to an adult administering an assessment if the adult is unfamiliar.

Culture also can shape a child's perception of, and thus her response to, questions. In some cultures, direct questions, or questions to which the answer is obvious are considered rude, thus making a child uncomfortable if asked such questions in an assessment. In such cases, a failure to respond should be interpreted as a temperamental or cultural norm, not as an indicator of inability or limited knowledge.

In addition to being inconsistent with some home cultures, the questioning found on standardized tests may be inconsistent with the approach to classroom curriculum. More traditional early childhood curricula – as well as curricula such as the Reggio Emilia and the Project Approach – build on children's interests and creativity. Children who experience this type of pedagogy may not be comfortable answering disconnected questions, and may not respond to questions in areas that are not of interest to them. Children in these types of programs typically perform well on a standardized assessment only if they come from home environments that utilize direct questioning, but if not, they are less apt to respond to standardized assessments. Although they may have the skills being assessed, they may not be willing or able to respond to the out-of-context questioning style of norm-referenced standardized tests (Fagundes, Haynes, Haak, & Moran, 1998; Laing & Kamhi, 2003; National Research Council, 2001). Under such circumstances, on-demand standardized assessments may be less a measure of what children know and can do and more a measure of how well children have acculturated to this type of questioning and on-demand performance.

These various problems may not be indicated by the traditional ways of documenting validity (concurrent, construct) because children usually are consistent in approaching this type of task and therefore the responses on different direct measures will be correlated. Some of these temperament and culture-based problems can lead to lower predictive validity (as is evident in preschool measures) because as they become acculturated to test-taking; the scores increase commensurate with the children's true ability. Performance on direct assessments typically show lower correlations with teacher judgments of children who are not good test-takers, since teachers are able to rate children based on a wider repertoire of tasks and observations.

Standardized assessments used in early childhood evaluation efforts, many of which are adaptive but draw from a limited set of constructs, show weak predictive validity (LaParo & Pianta, 2000; Konold & Pianta, 2005), and the predictive validity coefficients obtained for early childhood assessments are different from one study to another (Kim & Suen, 2003). Based on a meta-analysis from 70 longitudinal studies, LaParo and Pianta found that, on average, only about 25% of the variance in academic

achievement in the primary grades is predicted by the assessments administered in preschool or kindergarten. Konold and Pianta tried a different approach, using cluster analysis to create profiles from multiple different measures of children at 54 months and analyze the ability of the profiles to predict first-grade achievement. The measures they used in creating the profiles included both social-emotional and cognitive assessments with prior evidence of predictive validity. Once again, variability in development was the rule rather than the exception. The R-square statistics at the aggregate level ranged from .08 to .18 (i.e., 8 to 18% of the variance accounted for) on first-grade measures of the Woodcock-Johnson Tests of Achievement - Revised (Woodcock & Johnson, 1989). Kim and Suen (2003) conducted a generalizability study using hierarchical linear modeling (HLM) with 716 coefficients reported in 44 studies. They concluded that "the predictive power of any early assessment from any single study is not generalizable, regardless of design and quality of research. The predictive power of early assessments is different from situation to situation" (p. 561). Together, these studies warn of the dangers inherent in relying solely on standardized assessments of child outcomes in early childhood. Evaluation and accountability efforts must address more than just how children perform on standardized assessments.

Despite their drawbacks, individually administered standardized assessments have been helpful in large-scale research and program evaluations by raising and providing answers to important policy questions (see, for example, Burchinal et al, . 2002; Walston & West, 2004; Yu & Qiuyun, 2005). Specifically, the assessments have been used to demonstrate the positive effects of Head Start and Prekindergarten at the state level (Garces, Thomas, & Currie, 2002; Gormley & Gayer; 2005; Gormley, Gayer, Phillips, & Dawson, 2005; Henry, Gordon, Mashburn, & Ponder, 2001; Henry, Gordon, & Rickman, 2006). These studies used individually administered assessments, and the sample sizes were large enough that the measurement error associated with children's unfamiliarity with the test format was less of an issue, but care must be taken when interpreting the data.

For instance, when analyzing assessment data, it is assumed that the error is randomly distributed. This may not be the case, however, if one is examining program outcomes for young children, because the wording and tasks on a specific assessment will be more familiar to children in some programs than to children in others. For example, teachers with more education who are familiar with standardized tests, or those using a direct instruction curriculum, may use question formats in day-to-day instruction that are similar to standardized test formats. In tests given to children taught by these teachers, there would be less measurement error than in tests taught by teachers who are not familiar with standardized tests or by those not using a direct instruction curriculum. It is therefore important for researchers to consider whether the measurement error is randomly distributed, or if it is related to the findings of interest.

In addition to measurement error, sample size can affect the interpretation of mean test results. Smaller samples introduce the problem of missing data. Who is present on the day a test is administered can strongly affect findings when the size of a program is small. One or two children who are at one end of the normal distribution of scores and leave the program during the school year can more strongly change the mean score for a classroom or program. When the stakes are high in programs, parents of poorly performing children are sometimes asked to keep their child home on the days the tests are administered so that their scores do not pull down the mean. Therefore, care should be taken in interpreting test results from small samples.

The use of a multimethod approach to program evaluation would provide a more complete indication of child outcomes. The next section will discuss some of the ongoing observational assessment options.

OBSERVATIONAL AND ONGOING PERFORMANCE ASSESSMENTS

Although there are many commercially available assessments, few cover preschool through grade 3. Preschool classroom assessments usually address multiple domains of development and learning, and the most commonly used ones include the Preschool Child Observation Record (COR; High/Scope, 2004), the Creative Curriculum® Developmental Continuum for Ages 3–5 (Dodge, Colker, & Heroman, 2005), Galileo System for the Electronic Management of Learning (Galileo; Assessment Technology, Inc., 2004), and Work Sampling System™ (WSS; Meisels & Jablon et al., 2001). In addition, California recently developed the Desired Results Developmental Profile—Revised (DRDP-R, California Department of Education, 2006). Among these assessments, only WSS and the DRDP-R assess children from preschool through grade 3 on a continuum. The WSS and the DRDP-R are based on standards rather than being tied to a specific curriculum. These assessments will be discussed in more detail in this section. A number of websites and other sources provide additional information about the other preschool assessments (Berry, Bridges, & Zaslow, et al., 2004; NIEER, n.d.; Pai-Samant, et al., 2005; Shillady, 2004).

DESIRED RESULTS DEVELOPMENTAL PROFILE—REVISED (DRDP-R)

California recently developed a curriculum-embedded longitudinal measure to assess development from birth through age 12. The Desired Results Developmental Profile—Revised (DRDP; California Department of Education, 2006) was developed in alignment with the California learning standards and the research base on developmental levels. IRT was used to assign performance level indicators to different forms and to create scale scores for tracking progress across as well as within forms. All teachers in programs funded by the Child Development Division are required to complete the DRDP-R for the children in their program. Scale scores are created from the submitted ratings, and the data are aggregated for reporting to the state.

Content. The preschool form includes 36 items (or measures, as they are termed in the DRDP-R) that assess development in the following domains: social-emotional, language, cognitive (including literacy and mathematics), and physical (gross and fine motor, health and safety). Each item includes a rubric with a description and exemplars for each of four ratings: exploring, developing, building, and interpreting. On the basis of documented classroom observations, the teacher determines the level at which a child easily, confidently, and consistently demonstrates these four behaviors over time and in different settings. Teachers can check "not yet at this level" if a child is not yet 'exploring'; they also can indicate whether a child is 'emerging' to a next level. Teachers rate children only on developmentally appropriate indicators. Teachers and parents can see the developmental progression on charts that trace results over time from infancy through age 12. There is a separate form with guidance on assessment adaptations for children with disabilities.

Psychometrics. Because the DRDP-R is a very recent development, the psychometric information on it is limited. Preliminary data presented at recent conferences (Wilson et al., 2006) indicate high reliability of the scales, and that there is inter-rater agreement about the difficulty of items on adjacent forms.

Spanish versions of the DRDP-R are under development, as are additional revisions to align more closely with the newly drafted state Preschool Learning Foundations (California Department of Education, April 9, 2007).

Training. Training is available for teachers on how to document observations and make the ratings on the DRDP-R.

WORK SAMPLING SYSTEM™ (WSS)

The Work Sampling System™, or WSS, (Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 2001) is a standards-based, curriculum-embedded performance-assessment system. It is designed to be an ongoing collection of evidence of children's knowledge, skills, and behavior in a variety of classroom contexts. The WSS includes developmental guidelines and checklists, a focused collection of children's work, and summary reports. The developmental guidelines describe development on indicators from preschool (3 years old) through grade 3.

Content and Features. The WSS addresses language and literacy, mathematics, science, social studies, art, physical development, and social-emotional development (including approaches to learning). The most recent version reflects the changes in standards in the last decade. The language and literacy area includes indicators for listening, speaking, reading, writing, and spelling. The mathematical thinking area includes indicators for the areas addressed in the National Council for Teachers of Mathematics (NCTM) standards. Both concepts and procedures are addressed in the indicators. The scientific thinking area covers how children show evidence of understanding and of using the scientific processes of observing, describing, recording, posing questions, making predictions, forming explanations, and drawing conclusions. The social studies section includes indicators of a child's knowledge, skills, and understanding of the similarities and differences between people, roles, rules that govern behavior, and the environment around them. The arts section includes indicators of children's expression and representations of dance, drama, music, and the visual arts. The section on physical development addresses both fine and gross motor development, as well as health and safety indicators. The personal and social development section addresses a child's self-concept, self-control, approach to learning, social problem-solving, and interaction with adults and other children.

In addition to the regular preschool WSS guidelines, there is a Work Sampling for Head Start (Dichtelmiller, Jablon, Meisels, & Marsden, 2001), and several states (e.g., Arizona, Florida, Illinois, New Jersey, and New York) have created their own version of WSS, or cross-walks of their state's early learning standards to the WSS. The WSS also is available online. Known as Work Sampling Online (WSO), this feature allows teachers and administrators to generate reports easily based on the ratings that teachers enter. It also allows them to create customized child reports easily and aggregate summaries of outcomes (*T.H.E. Journal*, 2002). Strengths and weaknesses of a program can be examined by using the aggregated data.

One of the strengths of the WSS is that it allows users to examine the same areas (for example, number concepts) from preschool through the third grade. The developmental guidelines provide exemplars for each year or grade; this helps teachers to identify the level of skill expected of children at a given grade level and also to see how the skills in one year build upon previous skills, knowledge, or behavior (the developmental progression). Each indicator has several examples, thus showing the variety of ways in which a child may demonstrate the skill, knowledge, or behavior. Using multiple observations of a child and information gleaned from work samples, the teacher rates the child on the developmental checklist for different skills, knowledge, or behavior as "not yet," "in progress," or "proficient." Information from both the portfolio (focused collection of work samples) and the developmental checklist are summarized on the Summary Report at least twice a year. In addition to rating current performance, the teacher rates the child's progress in each area in the Summary Report.

Psychometric Information. The WSS does not provide a scale score or norms. Several states have devised ways to create scores from the developmental checklists. The reliability and validity of the WSS have been examined on the basis of ratings given by teachers in both the developmental checklists and the summary reports. Most of the WSS psychometric work has been done with the kindergarten through third-grade versions.

The reliability and validity of the WSS were examined in a study with experienced teachers (K-3) in low-income urban schools who had both received training in WSS and implemented the assessment system for at least two years (Meisels, et al., 2003; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, 2001). The teacher ratings on the WSS language and literacy and mathematical thinking sections were moderately to strongly correlated with the children's performance on the corresponding scales of the Woodcock-Johnson Revised Tests of Achievement (WJ-R; Woodcock & Johnson, 1989). In kindergarten and first grade, the scores from the teacher ratings on the checklist demonstrated a unique contribution to the prediction of the spring standard scores on the WJ-R, even after controlling for the fall WJ-R score and the child's age, race, ethnicity, and poverty status.

Parents and teachers expressed satisfaction with the assessment system and agreed that children benefited from the use of the WSS (Meisels, Bickel, et al., 2001; Meisels, Xue, et al., 2001). Analysis of children's academic achievement in subsequent years suggested that the children's involvement in the WSS facilitated continued progress over time. In a matched sample study (no random assignment), children in schools that used the WSS in kindergarten through third grade had greater gains in achievement in the fourth grade than children in classrooms most closely matched on demographic characteristics, as well as children in the remainder of the same grade classrooms in the school district (Meisels et al., 2003). For two reasons, these results should be interpreted cautiously, however. First, other curriculum initiatives were taking place in this district at the same time. Second, the study design raises concerns about selection bias and poor comparability on baseline achievement. Nonetheless, the findings do suggest that the WSS could support innovations in curriculum by focusing the teacher's attention on a child's progress in that particular curriculum.

Evidence of inter-rater reliability is not available for the current edition of the WSS or for all age/grade levels. High inter-rater reliability was found between kindergarten teachers on the Summary Reports using an earlier version of the WSS (Meisels, Liaw, Dorfman, & Nelson, 1995). The raters used both a child's portfolio collection and the teacher ratings on the developmental checklists to complete the Summary Report ratings for a child.

Training. The WSS provides a teacher's manual and other materials to help teachers understand how to do focused observations, how to document their observations and the work samples they collect, and how to report on what they know about the child in the Summary Report. The publishers offer a variety of training, from a half-day initial awareness session to individualized, extended train-the-trainer development. As with any assessment, training is fundamental to the reliable implementation of the WSS.

Several states are implementing adaptations of the WSS at the preschool and primary grade levels. The WSS has been tailored to the individual state-level standards and some states select a subset of indicators to monitor and report on a statewide and county-level basis. Adaptations of the WSS are being used in preschools (e.g., Illinois), preschool and primary grades (South Carolina and Maryland), or in the primary grades only (Delaware). See Appendix A for a more detailed discussion of how Maryland and South Carolina have adapted the WSS for these purposes.

OBSERVATIONAL ASSESSMENT: SUMMARY OF USE

As with all assessments, training in administering the assessment (data collection) and subsequent analysis of the evidence (data) needs to be provided in order to attain reliable results. In both Maryland and South Carolina, certified teachers in the primary grades completed WSS ratings used to assess child-level school readiness. It is not clear what level of training is needed to obtain reliable, valid ratings from preschool teachers, some of whom may not have a teaching certificate or much background in

early childhood development or assessment (Barnett et al., 2005). However, it is clear that providing examples and training increases the reliability of observational assessment, as administered by teachers. Several states already train individuals, such as retired teachers and college students, to reliably score writing and other performance-based assessments in their K-12 testing programs. Ohio offers training and exemplars in how to score its writing diagnostic assessment for K-2. For each level of the rating scale, teachers have several examples against which they can compare a child's work to rate it.

Well-defined rubrics or scoring guides also are helpful in establishing the reliability of teacher judgments (National Research Council, 2001). Sample rubrics are readily available online (see, for example, http://www.nwrel.org/assessment/; http://www.sdcoe.k12.ca.us/score/actbank/srubrics.htm; http://rubistar.4teachers.org/index.php). Guidance for developing valid and reliable rubrics is also available (Moskal, 2003; Moskal, 2000; Moskal & Leydens, 2000; Tierney & Simon, 2004).

In terms of states' individual uses of the WSS, South Carolina limited the number of domains assessed, and Maryland reduced the number of indicators from the WSS, ostensibly to reduce teacher burden. It appears from the documentation that both states made these changes, not on the basis of empirical data, but on the basis of expert opinion or consensus. IRT could be used to inform decision-making about item selection and to help set the criteria for the different ratings by providing relative item difficulties. IRT also would allow states both to examine the item difficulties in the process of selecting the indicators and criteria for monitoring progress, and to empirically validate the criteria set for that indicator. Having indicators that represent a range of difficulty levels would be helpful. In addition, matrix sampling of items could be used with assessments to ensure that all areas of the domain are represented across programs. Unfortunately, selecting only some domains or areas of domains makes it more likely that programs will address only those areas, (National Research Council, 2001). IRT also could help to facilitate data monitoring of questionable and atypical ratings. Unusual ratings could be investigated, standards could be established empirically, and districts would know what skills a given child is most likely to possess on the basis of his or her score.

Only California took advantage of current psychometric methods to inform the selection of indicators and to create scores using IRT. IRT creates an equal interval scale, making it easier to track growth when data are collected across multiple years or at multiple points in time. Both Maryland and South Carolina collected the data on child outcomes only at the beginning of kindergarten. The advantage of doing so is that teachers have no motivation for inflating ratings, since outcomes at that time of the year are baseline data for them. However, in terms of assessing the relative benefits of different programs, the data collected in kindergarten is problematic. How the children from different types of programs perform in kindergarten may be related more to their initial status before starting a specific prekindergarten program than to the preschool program itself. Resolving this problem means developing measures that are collected at the start as well as the end of preschool, and then at the beginning of kindergarten, thus making it possible to examine how children are learning, rather than how many children from disadvantaged areas attend a particular kindergarten classroom. Policymakers do not want to know which programs are recruiting the most able children, but rather, which programs are most beneficial in terms of raising the achievement level of children. Raudenbush (2005) asserts that it is "scientifically indefensible to use average achievement test scores of a school [to judge how good a job a school is doing]. We need to know how much kids are learning, not just how much they know" (p. 11).

Although WSS has provided evidence of concurrent validity, this evidence comes from studies that were conducted with children in kindergarten through the third grade. We need to amass more evidence of validity of observational tools used in preschools by using the most valid, direct instruments available as well as having observers and teachers discuss the available evidence of children's skills, knowledge,

and abilities. Teacher reports of children's activities are informed by previous experiences with the children as well as by what happens that day (Camburn & Barnes, 2004). It is even more probable that their reports on a child's current skills, knowledge, and abilities are informed by their previous knowledge of that child. For example, an outside observer might classify a child's response as 'inference,' which indicates a certain level of developmental sophistication, but the teacher, who knows that the topic surrounding the response was discussed in-depth the day before, would classify the response as 'recall,' which represents a less sophisticated, level of development. Understanding what influences teachers' ratings, and how those influences may affect validity, should be examined.

The additional information that a well-trained teacher brings to an assessment allows for examination of more complex learning. The ongoing nature of classroom observational assessment makes available information about how recently a skill was acquired. Using this method, teachers have a greater sense of the whole child and can consider how development in one area affects performance in another.

CHALLENGES AND CONCERNS WITH ONGOING OBSERVATIONAL ASSESSMENTS

One of the greatest challenges for ongoing assessments is establishing trust in teacher judgments. Reviews of research have established the conditions under which teacher ratings are reliable, including the need for items that are behaviorally anchored (Hoge & Coladarci, 1989; Perry & Meisels, 1996). Establishing and maintaining inter-rater reliability has been a concern in on-demand performance-based assessments as well. Most states have experience in training raters to agreement criteria. This task becomes more challenging for ongoing instructional assessments when the ratings are made in different geographic areas and on the basis of different types of evidence. One solution is to select specific examples of a work sample to be collected, or provide an observational chart with specific descriptions of behaviors to observe. Another is to provide a range of examples to establish the level of competence/difficulty involved in different types of samples of work or observational evidence.

The use of some standard means of documentation, as well as examples of the types of information that should be collected as evidence for an indicator, will be helpful in training teachers. Teachers also may need training in what information they need to add to work samples in order to best evaluate the learning. Using a set of materials to train teachers in reliably evaluating work samples will be important for ensuring the reliability of the data. During teacher training, Maryland established inter-rater reliability by using a common set of items. California incorporates video into their training on how to reliably document observed behavior. Inter-rater agreement would need to be verified at a minimum of once a year to prevent rater "drift." When teachers are able to compare the work of their children to a specific work sample, it helps them to apply the set criteria. Preschool teachers who have not had teacher training (those who are not certified teachers) are likely to need more training in observational skills in order to make reliable ratings (Mashburn & Henry, 2004).

When a specific example is not available for review, a normative framework can influence teachers' ratings. One of the problems with teachers rating only their own students is that the normative framework for them becomes their own students. To prevent the problems associated with a classroom normative rather than a criterion reference, and to extend a teacher's understanding of how to evaluate their students' work, teachers could be asked to rate samples of work and documentation for a few children from a classroom in a different program and to provide documentation from randomly sampled children in their classrooms, which would be rated by another teacher. This approach also would help teachers to understand what information needs to be documented to provide strong evidence for ratings. It also could generate additional ideas for the types of evidence that they could collect about children's skills, knowledge, or behavior. It may be that teachers will discover more efficient methods of documenting children's learning as they share their ideas with one another.

Different data collection tools, such as checklists, and ideas about appropriate work samples to be collected, could be made available to teachers. For example, in WSS, one of the indicators used by the Maryland Model for School Readiness was "shows understanding of number and quantity." The rating in this area at the kindergarten level is based on children's ability to "count objects to at least 20 . . . count using one-to-one correspondence reliably, use objects to represent numbers, and use numerals to represent quantities" (MSDE, 2002, p. A3). States could provide teachers with checklists that direct teachers to verify a child's ability to count different numbers of objects and to note how many objects the child reliably counts. Alternatively, the criterion might be for the child to count 20 items in at least three contexts before the teacher enters the rating. For preschool teachers who may have more limited educational background, the checklist could prompt them to designate the number of items that a child counts correctly with one-to-one correspondence when the items are arranged in different ways (spilled from a cup versus lined up) or when they are different sizes or shapes. The criteria for ratings of different items could be presented in a computer program that supports the data entry system. Teachers would indicate the descriptor that most resembles what the child did and the program would decide whether that descriptor meets the criteria for a specific rating.

Teachers have the ability to collect data in a variety of contexts and over time to gain a more valid and reliable measure of a child's ability. In addition, when teachers are good observers, they are more apt to provide specific feedback to children. Feedback is one of the strongest instructional predictors of achievement (Hattie & Jaeger, 1998). It is therefore wiser to invest in training teachers to be better observers and more reliable assessors than to spend those resources training and paying for outside assessors to administer on-demand tasks to young children in unfamiliar contexts that will provide data with the added measurement error inherent in assessing young children from diverse backgrounds (Meisels Atkins-Burnett, 2006; National Research Council, 2001). Unfortunately, not all teachers will be good assessors, so there still may need to be periodic assessment of samples of children to ensure the validity of the data being collected, and that positive outcomes are being achieved. Thus, continued work on direct measures should be undertaken to improve their comprehensiveness and validity, as well.

MEASURES OF INSTRUCTION AND CLASSROOM QUALITY

In addition to examining child outcomes, measures of the program itself—particularly measures of instruction and teacher-child interaction—need to be collected. If we want to know that the programs in which we are investing are high in quality, we should be assessing what we know about what ingredients create high-quality early childhood programs. If we want to know about school readiness, we should be asking whether our schools are ready to support the development of children who come with a diverse set of skills, rather than whether those children already have certain skills. The research on the measures of child outcomes indicates that children's performance is "situation specific" (i.e., children may demonstrate a skill in one situation and not in another) and that rapid changes in skills can occur (Pianta, 2003; La Paro & Pianta, 2001). It is through the interaction between what children and families bring to the school environment and what schools bring to children that success is engendered—or not. The quality of either environment (school and home) can strongly influence child outcomes. Accountability efforts in early childhood need to focus on the quality of environments provided to children from preschool through the third grade (Pianta, 2003).

Fortunately, as attention to early childhood education in the past two decades has grown, so has what we know about instructional and program factors that make a difference for young children (Burchinal et al,. 2000; La Paro, Pianta, & Stuhlman, 2004; Landry, n.d.; Peisner-Feinberg & Burchinal, 1997; Peisner-Feinberg et al., 2001; Pianta, 2003; Pressley, 2006; Dickinson, 2006). Highly effective teachers use positive classroom management, establish routines, provide feedback to students, engage them in extended conversations, and promote positive relationships in the classroom, all of which create a sense of trust and community; the instruction is also very deliberate (Matsumura, Patthey-Chavez,

Valdes, & Garnier, 2002; NICHD ECCRN, 2003; Mashburn & Pianta, 2006; Pressley, 2006; Dickinson, 2006).

In research on the preschool years, measures of the environment, such as the Early Childhood Environment Rating Scale—Revised (ECERS-R; Harms, Clifford, & Cryer, 1998) and measures of teacher-child interaction have been used widely in large-scale studies, as well as in program evaluation and accountability efforts. For example, South Carolina evaluated the quality of its preschool programs with the ECERS-R and found that children in classrooms that received quality ratings of good to excellent (5 or higher) also showed greater readiness on the SCRA than children in classrooms with lower quality ECERS-R ratings (Brown et al., 2006). However, global quality as measured on the ECERS-R may be a necessary, but not sufficient, evaluation of quality. While the ECERS-R includes scales that examine the interaction, activities and language, and reasoning opportunities, these areas are rated globally and the rating categories mix availability of materials with the activities/interactions around the materials.

Research supports the pivotal role of the teacher in supporting children's early development (Matsumura, Patthey-Chavez, Valdes, & Garnier, 2002; NICHD ECCRN, 2003; Mashburn & Pianta, 2006; Pressley, 2006; Dickinson, 2006). Measures of quality in early childhood programs have been used for many years to evaluate the adult-child interaction, with a focus on the relationship between the two and the supportiveness of the interaction. Examples include the Caregiver Interaction Scale (Arnett, 1989), the Adult Involvement Scale (Howes & Stewart, 1987), the Teacher Interaction Scale (Phillipsen, Burchinal, Howes, & Cryer, 1997), and the Observational Record of the Caregiving Environment (ORCE: Early Child Care Research Network (ECCRN), 2001), and the Child Caregiver Observation System (Boller & Sprachman, 1998). Recently, new measures have been used to assess the teacher-child relationship (preschool through third grade) in combination with a closer look at the instructional aspects of the classroom. These measures include Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2004), the Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2002), Classroom Observation System (COS; NICHD Study of Early Child Care and Youth Development, n.d.), the Early Language and Literacy Classroom Observation Toolkit (ELLCO; Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). The last measure focuses on a single academic area. Dickinson (2006) argues that both fine-grained (using time sampling and examining discrete categories such as those found in the CLASS and COS) and more global approaches (ratings of the classroom in different areas, such as the ECERS-R) are needed in evaluating programs.

The Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2004) examines "the emotional climate, instructional climate, and classroom management" (Mashburn & Pianta, 2006, p. 166). In all there are nine scales. Early childhood classrooms typically do well on measures of emotional climate, but the instructional climate in many preschools and elementary classrooms is characterized by passive engagement of children, lower-level basic skills, and instructional approaches that are not intentional or deliberate (little evidence that teachers design instruction specifically to challenge or extend children's skills) (Mashburn & Pianta, 2006; Pianta, 2003; Pianta & La Paro, 2003). This limited instructional climate is not found in all classrooms. After examining data from more than two thousand preschool and early elementary classrooms, Pianta (2003) noted that large variability in early childhood classrooms is found in every grade, and the entire range of codes or ratings is used across classrooms. Within classrooms, however, there appears to be stability. Within the NICHD study, some classrooms were observed more than one time (when more than one study child was in a classroom). The average correlations for both the global ratings and the more discrete time-sampled codes across days ranged from .71 to .91, indicating that the ratings were stable across time and could act as a reliable indicator of classroom instruction (Pianta, 2003). Pianta (2003) noted that teacher education and class size, the long-standing indicators of quality in elementary schools, were not related to child outcomes or to measures of instructional quality. The CLASS was developed from classroom

practice variables that were found to be associated consistently with child outcomes (Pianta, 2003). If policymakers want to monitor programs for young children, they must pay attention to what actually happens within classrooms, as well as to how teachers can be supported in implementing practices known to be associated with positive child outcomes.

Preschools and elementary schools should offer children a positive, caring emotional climate and stimulating, engaging instructional opportunities. Unfortunately, often this is not the case (Bryant et al., 2002; NICHD ECCRN, 2003). The documented long-term benefits of early childhood programs have been found to be associated with high-quality programs. Accountability efforts should include an examination of program quality, while at the same time working to ensure that all programs are high in quality by providing support to programs displaying weaknesses.

RECOMMENDATIONS AND CONCLUDING THOUGHTS

The assessments most important to accountability efforts are measures that assess both the instructional environment and what we know to be important aspects of quality, including measures of emotional climate, teacher-child interaction, and the quality and frequency of intentional instruction. Measures of child outcomes should include authentic tasks and use multiple sources of information, while recognizing the difficulties inherent in obtaining reliable assessments of young children (NEGP, 1998). The areas assessed should be important and meaningful for the child's development. If standardized direct assessments are included as one of the measures, they should be adaptive in nature so that the items are targeted measures of the child's skills, knowledge, and behavior.

Teachers have knowledge about response to intervention, background interests and prior experiences of a child that can be invaluable in interpreting the evidence of child performance. There can be additional advantages to increasing teacher use of ongoing assessment. When teachers develop strong assessment skills, they are more apt to target instruction in ways that scaffold learning and provide more specific feedback to children (Alexandrin, 2003; Hattie & Jaeger, 1998). Investing in increasing both the reliability of teacher judgments and the level of evidence that teachers use in making those judgments would not only inform accountability efforts, but also potentially increase the quality of the instruction. However, simply implementing performance assessments without providing teacher training and ongoing professional development can have unintended negative consequences, including narrowing curricula to include only areas of learning targeted for inclusion in reporting to the state, increased teacher stress, and decreased time devoted to instruction (Mehrens, 1998). We must be careful not to subvert the purpose of instructional assessment by attaching high stakes to them and failing to help teachers understand how to collect data within the context of instruction and to use that data to inform instruction.

With the advent of new technology, innovative ways of supporting teacher professional development are emerging (National Research Council, 2001). The ELA (Bodrova & Leong, 2001) analyzes child responses and estimates the range of skills that will be emerging next. It is designed to "emulate the decision-making process of master teachers by making connections between an individual student's raw assessment data and effective instructional strategies that are most likely to benefit a particular student at a specific time" (Bodrova & Leong, 2001, p. 23). Similar links could be programmed into other classroom-based assessment systems to support teachers in making inferences about student learning and in learning how to use what they know about children to alter instruction and scaffold learning. Pianta and colleagues (Bromley, 2006) are using an Internet-based conference system to provide ongoing professional development via expert consultation/coaching for new teachers. With the low cost and increasing accessibility of webcams and digital videography, teachers can document student learning in new ways, discuss their judgments about student learning, and reflect on the process with other professionals.

Sharing the evidence of student learning with others, and observing what children do in other settings, will help teachers to form a wider normative frame of reference. If teachers could share evidence of the progress of a random selection of a few children in their classrooms with another teacher or teachers outside of their program or school, they may better understand what documentation is helpful for understanding children's skills, knowledge, and behavior, and they would have at their disposal new ideas about alternative types of work that can be collected as evidence. If the ratings of one teacher in a program are verified by another teacher outside of the program, it would address concerns about the reliability and validity of teacher ratings.

The lessons learned from assessment in elementary and secondary schools should be heeded when deciding about early childhood assessments. The National Research Council (NRC, 2001) argued the following points regarding the use of assessments in our schools. First, there is "ample evidence of accountability measures negatively impacting classroom instruction and assessment" (p. 252). Second, effective assessments should be better designed and used as part of a system that "is aligned . . . vertically, across levels of the education system; horizontally, across assessment, curriculum, and instruction; and temporally, across the course of a student's studies" (p. 253). Third, measurement approaches should be comprehensive, and the different assessments should be coherent and complementary, with large-scale assessments examining more broadly while classroom assessments focus more closely on the same areas. Fourth, the assessments should be longitudinally designed to allow measurement of progress over time, moving away from a cross-sectional approach toward an approach geared to the "processes of learning" (p. 257). Fifth, as urged by professional organizations, assessments should yield information that ultimately improves learning. And finally, the NRC makes specific recommendations regarding program evaluation and large-scale assessments:

> Alternatives to on-demand, census testing are available. If individual student scores are needed, broader sampling of the domain can be achieved by extracting evidence of student performance from classroom work produced during the course of instruction. If the primary purpose of the assessment is program evaluation, the constraint of having to produce reliable individual student scores can be relaxed, and population sampling can be useful. . . *More of the research, development, and training investment must be shifted toward the classroom, where teaching and learning occur.*
>
> *A vision for the future is that assessments at all levels – from classroom to state – will work together in a system that is comprehensive, coherent, and continuous* In such a system, assessments would provide a variety of evidence to support educational decision making. (NRC, 2001, p. 258-259).

The early childhood years are an important time. We should ensure that we are providing programs commensurate with the overwhelming promise that these years hold for children's brighter futures. To do this, we must first examine program quality—the environment, the opportunities for learning, and the responsiveness, deliberateness, and supportiveness of adult-child interactions. We need to use multiple sources of evidence to assess the different dimensions of child outcomes and address the development of children more comprehensively. We must ensure that all of the measures used to assess children provide valid, reliable, and important information about their development. We will not know whether programs are benefiting children unless we are able to measure how the programs affect a child's development. More work is needed on measures of child outcomes, (particularly measures that assess children's development longitudinally from preschool through the third grade) and on assessment of teacher training, but the theory and technological advances to support this work are well within reach. Above all, we must heed the maxim to "do no harm" by seeing to it that assessments are used to inform how we can better support programs, teachers, families, and children.

REFERENCES

Agostin, T. M., & Bain, S. K. (1997). Predicting early school success with developmental and social skills screeners. *Psychology in the Schools, 3*(3), 219-228.

Alexandrin, J. R. (2003). Using continuous, constructive classroom evaluations. *Teaching Exceptional Children, 36*(1), 52-57.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10(4),* 541-552.

Assessment Technology, Inc. (2004). *Alignment of Ohio scales in Galileo.* Retrieved online September 19, 2006 from www.ati-online.com.

Atkins-Burnett, S., Rowan, B., & Correnti, R. (2001). *The use of group versus individual settings for assessing student achievement in kindergarten and first grade.* Consortium for Policy Research in Education, Study of Instructional Improvement, Research Note S-1. Ann Arbor: University of Michigan.

Bagnato, S. J., Neisworth, J. T., & Munson, S. M. (1989). *Linking developmental assessment and early intervention: curriculum-based prescriptions: Second edition.* Rockville, Md: Aspen Publications.

Barnett, W. S., Hustedt, J. T., Hawkinson, L. E., & Robin, K. B. (2007). *The state of preschool: 2006 state preschool yearbook.* New Brunswick, NJ: National Institute for Early Education Research. Accessed April 14, 2007 from www.nieer.org.

Barnett, W. S., Hustedt, J. T., Robin, K. B., & Schulman, K. L. (2005). *The state of preschool: 2005 state preschool yearbook.* New Brunswick, NJ: National Institute for Early Education Research. Accessed May 7, 2006 from www.nieer.org.

Belsky, J., Burchinal, M., McCartney, K., Vandell, D. L., Clarke-Stewart, K. A., & Owen, M. T., & The NICHD Early Child Care Research Network (2007). Are there long-term effects of early child care? *Child Development, 78*(2), 681-701.

Bergan, J. R. (1991). *Path-referenced assessment in the service of teaching and learning: The role of assessment in home, preschool, and school initiatives to promote an effective transition to elementary school.* ERIC # ED354073.

Bergan, J. R., Sladeczek, I. E., Schwartz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal, 28*(3), 683-714.

Bergan, J. Richard, Bergan, J. Robert, Rattee, M., Feld, J. K., Smith, K., Cunningham, K. & Linne, K. (2003). *The Galileo system for the electronic management of learning.* Assessment Technology, Inc. Retrieved online September 19, 2006 from http://www.ati-online.com/galileoPreschool/overview/index.htm

Berry, D. J., Bridges, L. J., Zaslow, M. J., Johnson, R., Calkins, J., Geyelin Margie, N., Cochran, S. W., Ling, T. J., Fuligni, A. S., and Brady-smith, C. (2004). *Early childhood measures profile.* Washington, DC: ChildTrends. Retrieved September 15, 2006 from http://aspe.hhs.gov/hsp/ECMeasures04/report.pdf.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647-663.

Bodrova, E., & Leong, D. J. (2001). *Tools of the mind: A case study of implementing the Vygotskian approach in American early childhood and primary classrooms.* Geneva, Switzerland: International Bureau of Education. Retrieved online September 21, 2006 from http://www.ibe.unesco.org/publications/Monograph/inno07.pdf#search=%22tools%20of%20the%20mind%22.

Boller, K., Sprachman, S. (1998). *Child caregiver observation system instructor's manual.* Princeton, NJ: Mathematica Policy Research, Inc.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (Eds.) (2001). *Eager to learn: Educating our preschoolers.* Washington, DC: National Academy Press.

Brigance, A. H. (2006). Brigance inventory of early development –II (IED-II). North Billerica, MA: Curriculum Associates, Inc.

Bromley, A. (2006, August 25). Curry's Pianta gets $10 million for national preschool study. *Inside UVA online, 36*(14). Retrieved October 8, 2006 from http://www.virginia.edu/insideuva/robert_pianta.html.

Brown, G., Scott-Little, C., McIntee, C., Hooks, L., Marshall, B. J., Weisner, A., & Amwake, L. (2006, March). *The South Carolina classroom quality research project: analyses of ECERS-R data and teacher attitudes toward the ECERS-R process.* Accessed June 25, 2006 from http://ed.sc.gov/news/more.cfm?articleID=628

Browning, K., Daniel-Echols, M., & Xiang, Z. (2006). *From implementation to impact: an evaluation of South Carolina First Steps to School Readiness Program.* Ypsilanti, MI: High/Scope.

Bryant, D., Clifford, R., Early, D., Pianta, R., Howes, C., Barbarin, O., & Burchinal, M. (2002, November). *Findings from the NCEDL Multi-State Pre-Kindergarten Study.* Paper presented at the annual meeting of the National Association for the Education of Young Children, New York, NY.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear modeling.* Thousand Oaks: Sage Publications.

Burchinal, M. R., & Cryer, D. (2003). Diversity, child care quality, and developmental outcomes. *Early Childhood Research Quarterly, 18*(4), 401-426.

Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology, 40,* 415-436.

Burchinal, M. R. , Roberts, J. E., Riggins, J. R., Zeisel, S. A., Neebe, E., & Bryant, D. M. (2000). Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development, 71*, 339-357.

California Department of Education (2006). *Desired Results for Children and Families- Revised.* Sacramento CA: Author.

California Department of Education (2007, April 9). *DRAFT preschool learning foundations.* Retrieved April 16, 2007 from http://www.cde.ca.gov/sp/cd/re/psfoundations.asp.

Camburn, E., & Barnes, C. (2004). Assessing validity of language arts instruction log through triangulation. *Elementary School Journal, 105*(1), 49-73.

Carlson, Stephanie M. (2005).Developmentally sensitive measures of executive function in preschool children.*Developmental Neuropsychology*, *28(2),* 595-616.

Childs, R. A., & Jaciw, A. P. (2003). *Matrix sampling of test items*. ERIC # ED482268. Retrieved October 13, 2006 from http://www.ericdigests.org/2005-1/matrix.htm.

CTB/McGraw Hill. (1997). Terra Nova Technical Bulletin 1. Monterey CA: Author.

Delaware State Department of Education (2001). *Delaware student testing program: Work Sampling assessment kindergarten and first grade development guidelines.* ERIC Document Reproduction ED 469 769) Dover, DE: Author.

Denham, S. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education and Development,* 17(1) 57-89.

Denton, K., Germino-Hausken, E., & West, J. (2000). *America's kindergartners*. [NCES 2000-070]. Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Dichtelmiller, M., Jablon, J., Meisels, S., & Marsden, D. (2001). *Work sampling for Head Start* New York: Pearson Early Learning.

Dickinson, D. K. (2006). Toward a toolkit approach to describing classroom quality. *Early Education and Development, 17*(1), 177-202.

Dodge, D. T., Colker, L., & Heroman, C. (2005). *Creative curriculum® developmental continuum for ages 3-5*. Washington, DC: Teaching Strategies, Inc.

Domenech, D. A. (December, 2000). My stakes well done: The issue isn't the academic benchmarks, it's the misguided use of a single test. *The School Administrator* (web version). Retrieved August 25, 2006 from http://auto.search.msn.com/response.asp?MT=cl.utoledo.edu&srch=5&prov=gogl&utf8.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test – Third Edition.* Circle Pines, MN: American Guidance Service.

Early Child Care Research Network (ECCRN) (2001). *Observational record of caregiving environment.* Retrieved August 8, 2006 from http://secc.rti.org.

Embretson, S. E., & Hershberger, S. L. (Eds.) (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fagundes, D., Haynes, W., Haak, N., & Moran, M. (1998). Task variability effects on the language test performance of southern lower socioeconomic class African American and Caucasian five year olds. *Language, Speech, and Hearing Services in Schools, 29,* 148-157.

Feld, J. (May 2005). *Using Assessment data to child outcomes: A look at the Head Start National Reporting System and Galileo Online.* Retrieved September 19, 2006 from http://www.nhsa.org/research/research_nrs_nhsa.htm.

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*(3), 311-330.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review, 92*(4), 999-1012.

Ginsburg, H. P. (2007). *Computer guided comprehensive mathematics assessment for preschool age children.* Presentation at the Biennial Meeting of the Society for Research in Child Development. March 29-April 1. Boston, MA.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability- 3*. Austin, TX: ProEd.

Good, R. H. & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades* (Technical Report No. 10). Eugene, OR: University of Oregon. Retrieved April 14, 2007 from http://dibels.uoregon.edu/techreports/index.php.

Gormley, Jr., W. T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-K program. *Journal of Human Resources, 4*(3), 533-558.

Gormley, Jr., W., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872-84.

Greenfield, D. B. (May 2005). *NRS Roundtable: Florida's coordinated child assessment initiative: Galileo & NRS,* Retrieved September 19, 2006 from http://www.nhsa.org/research/research_nrs_nhsa.htm.

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly, 21,* 431-454.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72,* 625-638.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale- Revised Edition.* New York: Teachers College Press.

Hattie, J., & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education: Principles, Policy, & Practice, 5*(1), 111-122.

Hauser-Cram, P., Warfield, M. E., Shonkoff, J. P., & Krauss, M. W. (2001). Children with disabilities: A longitudinal study of child development and parent well-being. *Monographs of the Society for Research in Child Development, 66*.

Henderson, K. (2002). The next step in assessment. (ERIC Document Reproduction # ED 481 224). *Little Prints, 2(2)* 1-2.

Henricsson, L., & Rydell, A. M. (2006). Children with behaviour problems: The influence of social competence and social relations on problem stability, school achievement and peer acceptance across the first six years of school. *Infant and Child Development, 15*(4), 347-366.

Henry, G., Gordon, C., Mashburn, A., & Ponder, B. (2001). *PreK longitudinal study: Findings from the 1999-2000 school year.* Atlanta: Georgia State University, Applied Research Center.

Henry, G. T., Gordon, C. S., & Rickman, D. K. (2006). Early education policy alternatives: Comparing quality and outcomes of Head Start and state prekindergarten. *Educational Evaluation and Policy Analysis, 28*(1), 77-99.

High/Scope Education Research Foundation. (2004). *Preschool child observation record, 2nd ed. (COR).* Ypsilanti, MI: High/Scope Press.

Hirsh-Pasek, K., Kochanoff, A., Newcombe, N. S., & de Villiers, J. (2005). Using scientific knowledge to inform preschool assessment: Making the case for empirical validity. *Social Policy Research, 19*(1).

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A Review of the literature. *Review of Educational Research, 59,* 297-313.

Howes, C., & Stewart, P. (1987). Child's play with adults, toys, and peers: An examination of family and child-care influences. *Developmental Psychology, 23*(3), 423-430.

Illinois State Board of Education (2002). The Work Sampling System Illinois: piloting soon in pre-k classrooms near you – why not yours. (ERIC Document Reproduction # ED 481 224). *Little Prints, 2(2),* 2-7.

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics performance: are simple one-minute measures technically adequate. *Assessment for Effective Intervention, 32*(2), 78-89.

Kaminski, R. A., & and Good, R. H. (1996). *Dynamic indicators of basic early literacy skills (DIBELS).* Retrieved April 26, 2007 from http://dibels.uoregon.edu/

Kauerz, K. (2006). *Ladders of learning: fighting fade-out by advancing PK-3 alignment.* Washington, DC: New America Foundation. Retrieved October 6, 2006 from http://www.newamerica.net/files/archive/Doc_File_2826_1.pdf

Keith, L. K., & Campbell, J. M. (2000). Assessment of social and emotional development in preschool children. In B. Bracken (Ed.) *Psychoeducational assessment of preschool children,* (3d ed., pp. 364-382). Boston: Allyn & Bacon.

Kim, J. & Suen, H. K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly, 18*, 547-566.

Klein, L., & Knitzer, J. (2006). Effective preschool curricula and teaching strategies. *Pathways to early school success: Issue brief n. 2.* New York City: National Center for Children in Poverty.

Kochanoff, A. T. (Ed.). (2003). *Report of the Temple University forum on preschool assessment: Recommendations for Head Start.* Philadelphia, PA: Temple University.

Konold, T. R., & Pianta, R. C. (2005). Empirically-derived, person-oriented patterns of school readiness in typically-developing children: Description and prediction to first-grade achievement. *Applied Developmental Science, 9*(4), 174-187.

Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development, 70*, 1373-1400.

Ladd, G. W., Herald, S. L., & Kochel, K. P. (2006). School readiness: are there social prerequisites? *Early Education and Development, 17*(1), 115-150.

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1997). Classroom peer acceptance, friendship, and victimization: Distinct relational systems that contribute uniquely to children's school adjustment? *Child Development, 68,* 1181-1197.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools, 34,* 44-55.

Lamy, C., Barnett, S., & Kwangee, J. (2005). *The effects of South Carolina's early childhood program on young children's readiness.* New Brunswick, NJ: The National Institute for Early Education Research, Rutgers University.

Landry, S. (n.d.). *Teacher Behavior Rating Scale TBRS*. Houston, TX: Center for Improving the Readiness of Children for Learning and Education.

LaParo, K. M. & Pianta, R. C. (2001). Predicting children's competence in the early school years. A meta-analytic review. *Review of Educational Research*, *70* (4), 443 – 484.

La Paro, K. M., Pianta, R., & Stuhlman, M. (2004). The classroom assessment scoring system: findings from the prekindergarten year. *Elementary School Journal, 104,* 409-426.

Lazarin, M. (2006).Improving assessment and accountability for English language learners in the No Child Left Behind act. *National Council of LaRaza Issue Brief, 16*. Retrieved April 14, 2007 from http://www.nclr.org/content/publications/download/37365.

Le, V-N., Kirby, S. N., Barney, H., Setodji, C. M., & Gershwin, D. (2006). *School readiness, full-day kindergarten, and student achievement: an empirical investigation*, (MG-558-EDU), Washington DC: RAND Corporation. Retrieved January 15, 2007 from http://www.rand.org/pubs/research_briefs/RB9232/

Maryland State Department of Education (June 2001). *Children entering school ready to learn. School readiness baseline information: final report.* (ERIC Document Reproduction: ED 458 939). Baltimore, MD: Author.

Maryland State Department of Education (2002). *Children entering school ready to learn: school readiness baseline information.* Baltimore, MD: Author. Retrieved September 18, 2006 from http://www.marylandpublicschools.org/.

Maryland State Department of Education (2006). *Children entering school ready to learn. Maryland school readiness information 2005-2006 state data.* Baltimore, MD: Author. Retrieved September 18, 2006 from http://www.marylandpublicschools.org/.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practices, 23*(4).

Mashburn, A. J., & Pianta, R. C. (2006). Social relationships and school readiness. *Early Education and Development, 17*(1), 151-176.

Matsumura, L. C., Patthey-Chavez, G.G., Valdes, R. & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in higher and lower achieving schools. *The Elementary School Journal, 103*, 3-25.

Maxwell, K. L., & Clifford, R. M. (2004). School readiness assessment. *Young Children: Beyond the Journal.* Retrieved October 12, 2006 from http://www.journal.naeyc.org/btj/200401/

Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*(3), 142-155.

McConnell, S. R., Priest, J. S., Davis, S. D., & McEvoy, M. A. (2000). *Best practices in measuring growth and development for preschool children (Draft).* Retrieved January 15, 2007 from http://education.umn.edu/ceed/projects/ecri/BestPracP2.pdf.

McConnell, S. R. (1998). *Get It, Got It, Go.* Retrieved January 15, 2007 from http://ggg.umn.edu/

McConnell, S. R., Priest, J. S., Davis, S. D., & McEvoy, M. A. (2002). Best practices in measuring growth and development for preschool children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology.* (4th ed., Vol. 2, pp. 1231-1246). Washington, DC: National Association of School Psychologists.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13).

Meisels, S. J. (1994). Designing meaningful measurements for early childhood. In B. L. Mallory & R. S. New (Eds.) *Diversity and developmentally appropriate practices: Challenges for early childhood education* (pp. 202-222). New York: Teachers College Press.

Meisels, S. J., & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J. P. Shonkoff & S. J. Meisels (Eds.) *The Handbook of Early Childhood Intervention* (second edition) (pp. 231-257). New York: Cambridge University Press

Meisels, S. J., & Atkins-Burnett, S. (2006). Evaluating early childhood assessments: A differential analysis. In K. McCartney & D. Phillips (Eds.) *Handbook on Early Childhood Development.* Oxford: Blackwell Publishing.

Meisels, S. J., Atkins-Burnett, S., & Nicholson, J. (1996). *Assessment of social competence, adaptive behaviors, and approaches to learning.* Working Paper Series, National Center for Education Statistics. Washington, D. C.: U. S. Department of Education, Office of Educational Research and Improvement.

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives*, 11(9). http://epaa.asu.edu/epaa/v11n9/.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in Kindergarten - Grade 3. *American Educational Research Journal*, *38* (1), 73–95.

Meisels, S. J., Jablon, J. R., Marsden, D. B., Dichtelmiller, M. L., & Dorfman, A. B. (2001). *The Work Sampling System.* New York: Pearson Early Learning.

Meisels, S. J., Liaw, F-R., Dorfman, A., and Nelson, R. (1995). The Work Sampling System: reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly,* 10(3), 277-296.

Meisels, S. J., Xue, Y., Bickel, D. D., Nicholson, J., & Atkins-Burnett, S. (2001). Parental reactions to authentic performance assessment. *Educational Assessment, 7*(1), 61-85.

Meltzer et al, (2004). *Learning disabilities: Research & Practice, 19*(1)

Missall, K. N., & McConnell, S. R. (2004). *Psychometric characteristics of individual growth and development indicators: picture naming, rhyming, and alliteration.* Minneapolis, MN: University of Minnesota Center for Early Education and Development. Retrieved January 15, 2007 from http://ggg.umn.edu/pdf/ecript8.pdf.

Missall, K. N., McConnell, S. R., & Cadigan, K. (2006). Early literacy development: skill growth and relations between classroom variables for preschool children. *Journal of Early Intervention, 29(*1), 1-21.

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*, 8(14). Retrieved September 25, 2006 from http://PAREonline.net/getvn.asp?v=8&n=14

Moskal, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation*, 7(3). Retrieved September 25, 2006 from http://PAREonline.net/getvn.asp?v=7&n=3 .

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). Retrieved September 25, 2006 from http://PAREonline.net/getvn.asp?v=7&n=10.

NAEYC & NAECS/SDE (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8. A joint position statement of the National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE).* Accessed online September 15, 2006 from http://www.naeyc.org/about/positions/pdf/pscape.pdf

National Education Goals Panel (1997). *Getting a good start in school.* Washington, DC: National Education Goals Panel.

National Education Goals Panel (1998). *Principles and recommendations for early childhood assessments.* Washington, DC: National Education Goals Panel.

National Institute of Child Health and Human Development NICHD Early Child Care Research Network, 2000. The relation of child care to cognitive and language development. *Child Development, 71,* 958-978.

National Institute of Child Health and Human Development Study of Early Child Care and Youth Development (n.d.) http://secc.rti.org/.

National Institute of Child Health and Human Development, Early Child Care Research Network (2003). A day in third grade: Observational descriptions of third grade classrooms and associations with teacher characteristics.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pelligrinno, J., Chudowsky, N., & Glaser, R., (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Science and Education. Washington, DC: National Academy Press.

NIEER (n.d.). *Assessment database.* http://nieer.org/assessment/.

Ohio Department of Education (2004). *Ohio diagnostic measures. Ohio observation measures.* Columbus, OH: Author.

Office of Assessment South Carolina Department of Education (2005). Alternate scoring supplement: South Carolina readiness assessment kindergarten and first grade developmental guidelines. Accessed online June 25, 2006 from http://ed.sc.gov/agency/offices/assessment/programs/scra-alt/documents/SCRAGuidelines_Final.pdf.

Pai-Samant, S., DeWolfe, J., Caverly, S., Boller, K., McGroder, S., Zettler, J., Mills, J., Ross, C., Clark, C., Quinones, M., & Gulin, J. (2005). *Measurement options for the assessment of Head Start quality enhancements: final report volume II.* Washington, DC: Mathematica Policy Research. Retrieved October 6, 2006 from http://www.mathematica-mpr.com/publications/PDFs/measurementoptions.pdf.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, J. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Boston College.

Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relation between preschool children's child care experiences and concurrent development: The Cost, Quality, and Outcomes Study. *Merrill-Palmer Quarterly, 43,*451-477.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72,* 1534-1553.

Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?* (NCES 9608). Washington, DC: National Center for Education Statistics.

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). *The prediction of process quality from structural features of child care.*

Pianta, R. C. (2003). *Experiences in p-3 classrooms: The implications of observational research for redesigning early education.* New York: Foundation for Child Development.

Pianta, R. C., & La Paro, K. M. (2003). Improving school success. *Educational leadership, 60*(7), 24-29.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2004). *Classroom assessment scoring system (CLASS).* Unpublished Measure. Charlottesville, VA: University of Virginia.

Pianta, R. C., Nimetz, S. L., & Bennett, E. (1997). Mother-child relationships, teacher-child relationships, and school outcomes in preschool and kindergarten. *Early Childhood Research Quarterly, 12,* 263-280.

Popham, W. J. (1993). Circumventing the high costs of authentic assessment. *Phi Delta Kappan 7,* 470-473.

Porter, A. C., & Olson, L. (Spring 2003). Standards and tests: keeping them aligned. *Research Points, 1*(1), 1-4.

Pressley, M. (2006, April 29). *What the future of reading research could be.* Paper presented at the International Reading Association's Reading Research 2006, Chicago, IL.

Printz, P. H., Borg, A., & Demaree, M. A. (2003). A look at social, emotional, and behavioral screening tools for Head Start and Early Head Start. Educational Development Center. Retrieved March 11, 2004 from http://notes.edc.org/CCF/ccflibrary.nsf/0/4ef7c268ca492cef85256e2f006d178c/$FILE/screentools.pdf.

Raudenbush, S. (2005). Newsmaker interview: How NCLB testing can leave some schools behind. *Preschool Matters, 3*(2), 11-12. Retrieved September 26, 2006 from http://nieer.org/resources/printnewsletter/MarApr2005.pdf.

Raver, C. C. (2002). Emotions matter: Making the case for the role of young children's emotional development for early school readiness. *Social Policy Report, 16,* (3), 3-18.

Raver, C. C., & Knitzer, J. (2002). *Ready to enter: What research tells policymakers about strategies to promote social and emotional school readiness among three- and four-year-old children.* New York: Columbia University National Center for Children in Poverty. www.nccp.org

Raver, C. C. & Zigler, E. F. (2004). Another step back? Assessing readiness in Head Start. *Young Children, 59,* 58-63.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability–3rd Edition.* Austin, TX: ProEd.

Riggs, N. R., Blair, C. B., & Greenberg, M. T. (2003). Concurrent and 2-year longitudinal relations between executive function and the behavior of 1st and 2nd grade children. *Child Neuropsychology, 9*(4), 267-276.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weise, B. (2002). *Snapshot.* Los Angeles: University of California, Los Angeles.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Rubin, K. H., Coplan, R. J., Nelson, L. J., Cheah, C. S. L., & Lagace-Seguin, D. G. (1999). Peer relationships in childhood. In M. H. Bornstein & M. E. Lamb (Eds.) *Developmental psychology: An advanced textbook (*4th Ed., pp. 451-501). Mahwah, NJ: Erlbaum.

Sarama, J., Clements, D. H., Liu, X. (2007). *Development of a measure of early mathematics developmental progressions using a Rasch model.* Presentation at the Biennial Conference of the Society for Research in Child Development, March 29-April 1. Boston, MA.

Scott-Little, C., & Martella, J. (2006, March). *Early learning standards: now that we have them, what next?* Presentation at the March 2006 National Smart Start Conference. Retrieved online June 25, 2006 from http://www.ccsso.ort/content/PDFs/EarlyLearningGuidelines.pdf.

Scott-Little, C., & Niemeyer, J. (2001). *Assessing kindergarten children: What school systems need to know.* Greensboro, NC: SERVE. Retrieved August 24, 2006 from http://www.serve.org/_downloads/publications/rdakcg.pdf

Scott-Little, C., Kagan, S. L., & Clifford, R. M. (Eds.) (2003). *Assessing the state of state assessments: Perspectives on assessing young children.* Greensboro, NC: SERVE. Retrieved August 24, 2006 from http://www.serve.org/_downloads/publications/ASSA.pdf

Shaul, M. S., Ward-Zukerman, B., Edmondson, S., Moy, L., Moriarty, C., & Picyk, E. (2003). *Head Start: curriculum use and individual child assessment in cognitive and language development: Report to congressional requestors.* ERIC Document Reproduction # ED 480 621. Washington, DC: General Accounting Office.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness.* New York: Pergamon.

Shillady, A. L. (2004). Choosing an appropriate assessment system. *Beyond the Journal.* Retrieved September 15, 2006 from http://www.journal.naeyc.org/btj/200401/shillady.

Shonkoff, J. P., & Phillips, D. A. (Eds.) (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, DC: National Academy of Sciences.

Shore, R. (1998). *Ready Schools.* Washington, DC: National Education Goals Panel. Retrieved October 13, 2006 from http://govinfo.library.unt.edu/negp/Reports/readysch.pdf.

Smith, M. W., & Dickinson, D. K. (2002). *User's guide to early language and literacy classroom observation toolkit.* Baltimore, MD: Brookes Publishing.

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (in press). Preliminary construct and concurrent validity of the preschool self-regulation assessment (PSRA) for field-based research. *Early Childhood Research Quarterly.*

South Carolina Department of Education (2005, November). The South Carolina Readiness Assessment. Retrieved June 25, 2006 from http://ed.sc.gov/agency/offices/assessment/programs/scra/.

South Carolina Department of Education (2005). *Alternate scoring supplement South Carolina Readiness Assessment* kindergarten and first grade developmental guidelines. Retrieved September 18, 2006 from http://ed.sc.gov/agency/offices/assessment/programs/scra/

Sroufe, L. A. (2005). Researchers examine the impact of early experience on development. *Brown University Child and Adolescent Behavior Letter, 21(*11), 1-3.

Sroufe, L. A., Egeland, B., Carlson, E., Collins, W. A. (2005). The development of the person: The Minnesota study of risk and adaptation from birth to adulthood. New York: Guilford Press.

Stecher, B. M., & Barron, S. (2001). Unintended consequences of test-based accountability: When testing in 'milepost' grades. *Educational Assessment,* 7(4), 259-282.

Stipek, D., & Byler, P. (2004). The early childhood classroom observation measure. *Early Childhood Research Quarterly, 19 (3),* 375-397.

Teaching Strategies, Inc. (2003?). *Creative curriculum for preschool.* Washington, DC: Author.

____ (2002). SchoolSuccess, Pearson launch assessment reporting system. *T. H. E. Journal, 29*(10), 30.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Retrieved September 25, 2006 from http://PAREonline.net/getvn.asp?v=9&n=2.

Tur-Kaspa, (2004). *Learning Disabilities: Research & Practice, 19*(1),

U. S. Department of Health and Human Services (June 17-18, 2002). *Early childhood education and school readiness: conceptual models, constructs, and measures. Workshop summary.* Washington, DC: U.S. Government Printing Office Retrieved June 22, 2006 from http://www.nichd.nih.gov/publications/pubs_details.cfm?from=&pubs_id=5665

VanDerheyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27*(1), 27-41.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer

VORT Corporation (1995). *HELP for preschoolers.* Palo Alto, CA: Author.

Walston, J., & West, J. (2004). Full-day and half-day kindergarten in the United States: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999. (NCES 2004-078). Washington, DC: National Center for Education Statistics.

Wilson, et al., (2006, April 6). *Desired results developmental profile: Observational assessment of child development.* Presentation to National Head Start Conference 33rd Annual Training Conference.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-educational Battery – Revised.* Chicago: Riverside.

Woodcock, R. W., & McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Chicago: Riverside.

Xue, Y. & Meisels, S.J. (2004). Early literacy instruction and learning in kindergarten: Evidence from the Early Childhood Longitudinal Study—kindergarten class of 1998–99. *American Educational Research Journal, 41* (1), 191–229.

Yu, W., & Qiuyun, L. (2005). Effects of class size and length of day on kindergarteners' academic achievement: Findings from Early Childhood Longitudinal Study. *Early Education and Development, 16*(1), 49-68.

Appendix A

MARYLAND MODEL FOR SCHOOL READINESS

The Maryland Model for School Readiness (MMSR) is Maryland's early childhood assessment initiative. The MMSR uses 28 indicators from the WSS developmental guidelines and checklist to assess children's readiness for kindergarten and to identify areas in which children need additional support (Maryland State Department of Education, 2001). The Maryland State Department of Education (MSDE, 2001) reported that the WSS was being used in some child care and Head Start programs, in addition to many of the prekindergarten classrooms and all of the local elementary schools. The MMSR, aligned with the Head Start Outcomes, has been used since 2000 in most Head Start programs in Maryland. It became available to child care centers in 2002.

Beginning in the fall of the 2000-2001 school year, MSDE collected baseline kindergarten data on a random sample of children in the state using the WSS (census data were collected in subsequent years). Teachers rated each child's performance on the developmental checklist, based on evidence collected in the first week of school through the end of October. MSDE created summative scores (range 4-12) in each domain and assigned cutpoints to different levels of readiness. Each domain initially included four indicators. Children with domain summary scores equal to or greater than 10 were considered "fully ready." Children with domain summary scores from 7 through 9 were considered "approaching readiness," and children with domain summary scores below 7 were considered to be "developing readiness" in that domain. MSDE reported aggregate scores by county by race/ethnicity, sex, disability status, and English proficiency status (yes/no). After the 2000-2001 school year, MSDE collected data on the type of preschool experience the child had and aggregated findings based on those categories.

MSDE now collects data on all of the children in the beginning of kindergarten. After the 2000-2001 school year, MSDE revised some of the indicators (increasing the difficulty or adding to the specificity of the indicators) and expanded the number of indicators to 30, adding ones for phonemic awareness and comprehension of fiction and nonfiction. Because the language and literacy domain now has six items, the state also adjusted the cut scores for that domain. Full readiness was indicated by a score of 15 or greater, approaching readiness, by scores of 10 to 14, and developing readiness, by scores of 6 to 9. The indicators have remained the same for the past four years.

The most recent report, based on kindergarten teacher reports of children on the MMSR in the fall, shows that the trend in children's readiness for kindergarten has been positive over the past four years, as indicated by the teachers' fall WSS developmental checklist scores on 30 items (MSDE, 2006). Both the composite score and the language and literacy score showed an increase in the percentage of Maryland children demonstrating full readiness. For the state overall, the share of children fully ready for kindergarten rose by 11 percent. The greatest increase was in language and literacy. In 2005-2006, 14 percent more children than in 2001 were fully ready in language and literacy. However, the differences were not consistent across subgroups. Among children who attended state-funded prekindergarten programs, there was an 18 percent increase. This finding suggests that the MSDE is sensitive to the different interventions being implemented in Maryland (i.e., state-funded prekindergarten, Head Start, child care, and so on).

MSDE uses several safeguards to ensure that the data collected are reliable and valid. Teachers participate in a professional development program staffed by expert consultants in the observation, documentation, and evaluation of student learning. These consultants use standard training materials, and the "teachers' accuracy in rating students' skills and abilities" (MSDE, 2006, p. C-4) is assessed during the professional development program using standard training materials and evaluation forms. The student assessment data are scanned and checked for reliability by an outside testing vendor

(MSDE, 2006, p. C-4). This reliability analysis data includes examining both the internal consistency of the assessment and the relative influence of each item on the scale (item-scale correlations); also included is a correlation analysis of the relationship between student scores and school scores. The data are disaggregated by race/ethnicity; sex; prior preschool experience, special education status, English proficiency, and free and reduced-price meal status (yes/no).

The demographic variables displayed expected relationships, for example, more children without disabilities showed full readiness than did children with disabilities, and more children with English proficiency showed full readiness than did children with limited English proficiency. Among the various preschool experiences, children in a nonpublic nursery school were most often rated as fully ready, and children who stayed at home or attended Head Start were least likely to be fully ready (MSDE, 2006).

In addition to using the data for its own purposes, MSDE shares the information with county districts and teachers build their understanding of the relative strengths and weaknesses of children as they enter kindergarten. Teachers can use the information immediately to plan instruction that is better targeted to the children they teach. In 2002-2003 school year, kindergarten teachers reported that the MMSR helped them in planning for individual children (92%), in determining how to group children (78%), in reporting to parents (86%), and as a source of evidence in making referrals for student evaluations (68%) (MSDE, 2006, p. C-3). The MSDE and counties can use the disaggregated data to examine the differences in areas of strength based on the different experiences of the children in their county. They can both examine how well the needs of different groups of young children are being served and use that information to target additional programming to those who need it. For example, children who stayed at home or with relatives before coming to kindergarten in 2001 were least likely to be rated as fully ready (39% of those in home/informal care compared to 67% in private nursery and 47% in prekindergarten; MSDE, 2002). Beginning in 2003, MSDE began distributing monthly "Parent Tips" on a variety topics related to supporting the development of preschool children at home.

SOUTH CAROLINA READINESS ASSESSMENT

The South Carolina Readiness Assessment (SCRA) is also based on the WSS. Unlike the MMSR, the SCRA focuses on only three domains: language and literacy, mathematics, and personal/social development. The selected indicators are aligned with the South Carolina standards in English language arts and mathematics. SCRA requires a minimum of two work samples per domain semiannually, and the South Carolina Department of Education (SC DOE) recommends that teachers consider what evidence would be necessary for another teacher to rate a given child in a given area (SC DOE, November, 2005). Teachers enter checklist ratings for the three domains online at least twice a year. SC DOE disseminated an alternative version of the SCRA to provide guidance to kindergarten and first-grade teachers and districts for students with significant disabilities (Office of Assessment, SC DOE, 2005).

In a recent evaluation of the First Steps Program (South Carolina's early childhood initiative), HighScope derived factor scores from the checklist ratings (Browning, Daniel-Echols, & Xiang 2006). Two factors were derived from the personal/social items, one addressing social skills (including self-control and interaction with others), and the other addressing approaches to learning (including self-concept and different approaches to learning). A language and literacy factor with 12 items and a mathematics factor with 14 items were the other factors in the analysis. Factor loadings for the language and literacy and mathematics items were greater than .75, the majority being greater than .80. Factor loadings for the social skills and approaches to learning scale were somewhat lower, although the majority of the loadings were greater than .70. These factors explained more than 67% of the variance for each grade.

Using the SCRA factor scores as outcomes, and controlling for child characteristics and demographic factors (age, ethnicity, special education status, mother's education, low birth weight, foster care, and several economic factors), the researchers found differences in academic achievement between children who did not receive classroom programming at the age of four and those that did. Children who were enrolled in a full-day program for four-year-olds had higher kindergarten scores than children who were in a half-day program or had no preschool program. This effect was stronger for children in minority groups. These findings again suggest that the assessment is sensitive to the intervention.

Counties in South Carolina are given the flexibility to decide how to provide services to at-risk children. Some areas use programs to strengthen parenting and families. Others devote funding to improving the quality of early childhood programs. Still others try to increase the number of children served. It does not appear that SC DOE disaggregates the SRCA kindergarten data by county or by child or program characteristics. The choice not to do so limits the usefulness of the data in understanding which strategies are effective for which groups of children.

TEXT FOR BOXES AND SIDEBARS:

ASSESSMENT STANDARDS

All assessments administered in early childhood should adhere to standards that have been agreed upon and supported by national professional groups such as the American Educational Research Association (AERA), American Psychological Association (APA), Chief Council of State School Officers (CCSSO), National Council on Measurement in Education (NCME), the National Association for the Education of Young Children (NAEYC), and the National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE). According to these standards, there must be evidence that the measures are reliable and valid, not only for the purpose for which they are used, but also for the sample of children who are being assessed. Evidence of children's abilities and skills should be collected in multiple ways (Meisels & Atkins-Burnett, 2006; NAEYC & NAECS/SDE, 2003), and decisions about children or programs should never rest on a single assessment. Assessors should be trained in interacting with young children and in administering the assessment. Children should be assessed in contexts that are meaningful, and the assessment should reflect the child's skills and abilities in realistic situations. What is assessed should be developmentally or educationally important. When assessments are used for program evaluation, multiple sources of data should be used and children's gains over time should be examined (rather than examining a single time point). When used for accountability purposes, the results of assessments should be employed for continuous improvement rather than to impose penalties (NAEYC & NAECS/SDE, 2003).

TYPES OF ASSESSMENT

Norm referenced – users should examine whether the sample used for norming the assessment included children who would be representative of the children they serve

Criterion referenced

- Standards-based: the standards are the criteria
- Performance-based: performance on tasks similar to daily activities; strong social validity; may examine process as well as product
- Developmental: developmental milestones and steps toward the milestones are the criteria

Both criterion-referenced and norm-referenced

Types of Administration

Direct On-Demand Administration – may include multiple choice questions, open-ended responses, performance-based responses to standard probes

- Group administration – not recommended for children younger than 8 years old; usually grade-specific and suffer from ceiling and floor problems
- Individual administration – most appropriate for young children
  - Adaptive administration – may use start/stop rules or two-stage design to obtain better measurement in a shorter administration
  - Curriculum-based measures – fluency measures designed to be administered in less than five minutes but frequently throughout the year

Observational Ongoing Assessment – allows a wider sampling of skills and behaviors to be assessed; High social validity

- Checklists – lists of skills or behaviors, may be lists of developmental milestones or standards and performance indicators
- Rating Scales – may be ratings of frequency or of how characteristic behaviors or skills are for the child
- Rubrics – scoring guides that describe several levels of performance. They can be used to describe multiple aspects of performance. They are particularly well suited when looking at qualitative differences in behavior or process differences.

Examples of Commercial Adaptive Assessments with Preschool Items

Expressive One-Word Picture Vocabulary Test – Third Edition (EOWPVT; Brownell, 2000)
*Peabody Picture Vocabulary Test Fourth Edition (*PPVT-4; Dunn, Dunn, & Dunn, 2007)
Woodcock-Johnson Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2001) Letter-Word Identification; Applied Problems; Phonological

Examples of Observational Assessments Preschool through Grade 3

Desired Results Developmental Profile – Revised (DRDP-R; California Department of Education, Child Development Division, 2006)

Work Sampling System (Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 2001).