OBSERVATIONAL MEASURES OF QUALITY IN CENTER-BASED EARLY CARE AND EDUCATION PROGRAMS

Research-to-Policy, Research-to-Practice Brief OPRE 2011-10c
December 2010
Observational Measures of Quality in Center-Based Early Care and Education Programs

Research-to-Policy, Research-to-Practice Brief OPRE 2011-10c

December 2010

Submitted to:
Ivelisse Martinez-Beck, Ph.D., Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Donna Bryant
FPG Child Development Institute, University of North Carolina at Chapel Hill

Contract Number: HHSP233200500198U

CONTRACTOR
Project Director: Martha Zaslow, Child Trends
4301 Connecticut Ave NW
Washington DC, 20008

Suggested Citation:

This Research-to-Policy, Research-to-Practice brief series focuses on issues related to the development and refinement of measures to assess the quality of early childhood settings. The views represented in this brief do not necessarily reflect the opinions of the Office of Planning, Research and Evaluation of the Administration for Children and Families.
OVERVIEW. Classroom observation measures that were originally developed and refined for early childhood research purposes are increasingly being used in state Quality Rating Systems (QRS), child care licensing, tiered reimbursement, and professional development. Understanding the characteristics and predictive power of these measures is critical to correctly interpreting and using the data that they produce. This brief reviews several widely used assessments and their relation to each other and to child outcomes. Particular attention is given to purposes for assessment, psychometric properties, inter-rater reliability, applicability of measures across ages, and content and cross-cultural validity. While several classroom observation methods have been shown to predict later child outcomes, classroom features and experiences still account for far less of child variability than family characteristics do. However, despite the modest sizes of the associations between child care quality and child outcomes, quality measures do consistently and significantly confirm these links; further development of quality measurement tools is warranted.

Classroom observation measures originally developed for early childhood research purposes are now being used in state Quality Rating Systems (QRS), child care licensing, tiered reimbursement, and professional development (PD). As financial consequences are attached to the scores obtained from these measures, policymakers want evidence about whether they are good measures. Researchers also want to use measures that are policy-relevant. Both policymakers and researchers want to know whether these measures reflect accurately the range of care that exists, whether improvement on these measures is possible, and whether improvement on the measures relates to improvement in children’s outcomes. A handful of well-established measures are typically used in research with center-based early care and education environments, most measuring broad aspects of classroom quality and some capturing quality in a specific domain. The purposes of this paper are to briefly review these assessments and then note key features of measures that should be considered when selecting a measure for use in quality improvement programs or early childhood policy initiatives.
What Should Our Quality Measures Assess?

Table 1 summarizes the domains covered by 11 widely used classroom research observation tools, specific data collection procedures, and applicable age range for each measure. Each measure typically includes multiple domains of classroom experience, but no measure covers all domains. These include frequent and warm interactions between teachers and children; rich language use; extending children’s knowledge through elaboration and contingent responsiveness; a variety of activities that encourage reasoning and problem solving and are culturally appropriate; opportunities for children to be with others in large and small groups and alone; consistent and positive use of behavior management strategies; safe and healthy daily routines; and good planning and time management.

Table 1

**Early Childhood Classroom Observation Measures for Global Quality or Dimensions of Quality**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Domains Observed</th>
<th>Observation Procedure</th>
<th>Age Range</th>
<th>Key references</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIS*: Caregiver Interaction Scale</td>
<td>Emotional tone, discipline style, and responsiveness of teachers</td>
<td>45 minutes; rating of 26 items; 4 point scale</td>
<td>Toddlers – Kindergarten</td>
<td>Arnett, 1989</td>
</tr>
<tr>
<td>CLASS: Classroom Assessment Scoring System</td>
<td>Teacher-child interactions in 3 domains: instructional support, emotional support &amp; classroom organization</td>
<td>2-3 hours; 30-minute cycles of observe-code; 10 items; 7 point scale</td>
<td>PreK &amp; K-3 versions; toddler soon</td>
<td>Pianta, La Paro, &amp; Hamre, 2007</td>
</tr>
<tr>
<td>ECCOM: Early Childhood Classroom Observation Measure</td>
<td>Quality of instruction, management, social climate, cultural sensitivity, and resources</td>
<td>3 hours; time sample of specific behaviors</td>
<td>Ages 4-7</td>
<td>Stipek &amp; Byler, 2004</td>
</tr>
<tr>
<td>ECERS-R: Early Childhood Environment Rating Scale –Revised</td>
<td>Global quality &amp; 7 subscales: space and furnishings, personal care, language and reasoning, activities, interactions, program structure, and parents/staff</td>
<td>3 hours + 20 minute interview; 43 items; 7 point scale</td>
<td>Ages 2.5-5</td>
<td>Harms, Clifford, &amp; Cryer, 1998</td>
</tr>
<tr>
<td>ECERS-E: Early Childhood Environment Rating Scale – Extended</td>
<td>Developed to supplement the ECERS-R with more focus on academic achievement: literacy, math, science &amp; diversity; reflects the British national pre-k curriculum</td>
<td>2 hours + 5 minute interview; 18 items; 7 point scale</td>
<td>Ages 4-6</td>
<td>Sylva, Siraj-Blatchford, &amp; Taggart, 2003</td>
</tr>
<tr>
<td>ELLCO: Early Language and Literacy Classroom Observation</td>
<td>3 tools: (1) Literacy environment checklist; (2) Classroom rating of 14 dimensions of literacy; &amp; (3) Literacy Activities Rating Scale with a summary rating</td>
<td>1.5 hours; 24 checklist items; 14 observed items on a 5 point scale</td>
<td>Ages pre-k to 3rd grade</td>
<td>Smith, et al., 2002</td>
</tr>
</tbody>
</table>
Table 1 (continued)

**Early Childhood Classroom Observation Measures for Global Quality or Dimensions of Quality**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Domains Observed</th>
<th>Observation Procedure&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Age Range</th>
<th>Key references</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITERS-R: Infant/Toddler Environment Rating Scale – Revised</td>
<td>Global quality &amp; 7 subscales: space and furnishings, personal care, listening and talking, activities, interactions, program structure, and parents/staff</td>
<td>3 hours + 20 minute interview; 39 items; 7 point scale</td>
<td>Ages birth – 3 years</td>
<td>Harms, Cryer, &amp; Clifford, 2003</td>
</tr>
<tr>
<td>ORCE: Observational Record of the Caregiving Environment</td>
<td>Focuses on an individual child’s interactions with adults; sensitive, warm, and responsive caregiving; several discrete behaviors and 5 qualitative ratings</td>
<td>2 observation cycles of 44 minutes; discrete behaviors and global ratings</td>
<td>Ages 6-54 months</td>
<td>NICHD ECCRN, 1996 &amp; 2001</td>
</tr>
<tr>
<td>PQA: Preschool Program Quality Assessment – 2&lt;sup&gt;nd&lt;/sup&gt; edition</td>
<td>3 observed domains: learning environment, daily routines, and adult-child interaction; 4 domains via interview: curriculum planning and assessment, parent involvement, staff qualifications and program management.</td>
<td>2-3 hours + teacher interview; 63 items; 5-point scale</td>
<td>Ages 3-5</td>
<td>High/Scope Educational Research Foundation, 1989 &amp; 2003</td>
</tr>
<tr>
<td>Profile: Assessment Profile for Early Childhood Programs</td>
<td>5 subscales: learning environment, scheduling, curriculum, individualizing, interacting</td>
<td>2-3 hrs; 60-item checklist; Yes/No response</td>
<td>Ages 3-7</td>
<td>Abbott-Shim &amp; Sibley, 1998 (the research version)</td>
</tr>
<tr>
<td>Snapshot&lt;sup&gt;b&lt;/sup&gt;: Emerging Academics Snapshot</td>
<td>Child’s exposure to instruction and engagement in 6 academic activity settings, 11 content areas, &amp; 6 levels of teacher responsivity</td>
<td>2-4 hours; time sample of specific settings and behaviors</td>
<td>Ages 1-8</td>
<td>Ritchie, Howes, Kraft-Sayre, &amp; Weiser, 2001</td>
</tr>
</tbody>
</table>

<sup>a</sup> Minimum observation time recommended; number of items on measure; and type of rating scale

<sup>b</sup> Measure can also be used with caregivers in family child care homes

Most parents would agree that these classroom dimensions are all important, but a quality enhancement consultant or a state child care administrator choosing a quality measure might wonder whether some dimensions are more important than others. Some researchers urge a stronger focus on measures that solely assess teacher-child interactions, setting aside physical features of the environment (Pianta, 2006); others emphasize language and literacy preparation (Dickinson, 2002). Although research is making some progress in linking specific components of quality to specific child outcomes (Burchinal, et al., 2009), currently, measures that reflect multiple and broad dimensions tend to predominate in quality rating systems (QRSSs) and program improvement efforts, often supplemented by measures with more specific focus.
Unable to specify that one or two explicit dimensions are the most important, we should heed Lambert’s advice (2003) that the choice of a measure should reflect the purpose of its use. For example, a measure that emphasizes environmental stimulation for language and literacy development in early childhood classrooms may be most appropriate if the purpose is to assess a policy initiative focusing on improving young children’s early literacy. The measures in Table 1 originated in research, but many have now been used for the purposes of self-assessment, program improvement, accreditation, or licensing.

What criteria should be considered when selecting a measure? Content- and age-appropriateness are primary. Validity, reliability, and ease of use are important, as well as ability to detect changes that might result from PD and other quality enhancement interventions. Most importantly, a good measure should relate positively to children’s outcomes. These considerations are discussed in the next sections with illustrative data from the measures described in Table 1.

Content: Succinctly Describing Quality and Various Dimensions

Observational measures are comprised of many individually scored items that can generally be averaged into a global quality score, the most frequently reported measure of quality. Individual item scores can also be grouped into subscale scores; for example, the Curriculum subscale of the Profile is the sum of 6 observed items. Authors create subscales conceptually, not empirically, so one should be cautious about over-interpreting subscale results, but for self-assessment or program enhancement, subscale use seems reasonable. Statistically rigorous research typically uses factors – the way individual items go together regardless of their subscale membership -- to answer research questions. For example, evaluations using the ECERS-R often report on the Teaching and Interactions Factor and the Provisions for Learning Factor, although no ECERS-R subscales exist by those names. These two factors have emerged from statistical analyses conducted in over 20 studies using the ECERS (see Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005, for the largest of these). Similarly, two main factors have also been found with the CLASS--Emotional Climate and Instructional Climate (Pianta et al., 2005). A large study of public pre-k found both ECERS-R and CLASS factors related to several hypothesized teacher and classroom characteristics (Pianta et al., 2005).

Although factor scores are efficient and statistically sound, they may reduce attention to potentially important domains of quality. For example, factor analyses of the ECERS and ITERS seldom contain items related to health, safety or facilities upkeep, yet these foundational elements of early childhood programs assure children’s health and safety and should be assessed, monitored, and improved when necessary. An unmeasured domain is not likely to receive attention. (Goodson and Layzer, 2010, this series, question this assumption.)

Researchers often warn against much emphasis on individual item scores, but specific items of the ECERS-R were used purposefully by the New Jersey Abbott pre-k program evaluators. Individual item-level data on indoor and outdoor space and equipment repair documented the extreme needs of typical programs (Lamy et al., 2004), resulting in a special legislative appropriation targeted to facilities, an area sometimes costly and hard to improve.

Whether factors, subscale scores, or even individual item scores are reported and used is usually related to the purpose of measurement. For research, factors are preferred; for program improvement purposes, subscales are often used; and for regulatory purposes, global scores predominate. Some domains of quality such as health and safety, may be better summarized as scales where assessment determines if standards are met or not.
Applicability across Ages

The age range for which an observational quality measure is needed quickly narrows one's choice of assessment. Most measures listed in Table 1 are intended for classrooms of preschool-aged children, while only three are indicated for use in infant-toddler classrooms (ITERS, ORCE, and Snapshot). No measure covers the age range from birth to 5, although the theoretically and procedurally similar ITERS-R and ECERS-R together will do so. The CIS, which captures interactional style and emotional tone, spans the widest age range, but even it is not applicable for infant and toddler classrooms. The ECERS was modified for use in kindergarten (Bryant, Clifford, & Peisner, 1991), but not for higher grades. The Profile has been extended to be applicable for early elementary grades and was used in the national Head Start Transition Demonstration Program (Ramey et al., 2000). As preschool is becoming more a part of school, the CLASS also fills the need to have a measure of classroom instructional processes spanning ages 3-8, and a toddler version of the CLASS is in development. The ELLCO and ECERS-E are relatively more difficult to use in classrooms of 3-year-olds or mixed-aged classes of 3s/4s because several items concern pre-academic teaching, group teaching, or particular activities that are generally not seen in, or even recommended for, younger children. Given the cost of valid instrument development, we are fortunate to have these well-known measures to choose from, however, if programs and policymakers want to include infants and toddlers in QRSs, more work is needed on observational measures in this age range.

Validity

One indicator of a measure’s validity is whether it captures the target construct well. Each of the measures considered here has shown adequate validity, typically by demonstrating high correlations to other measures of the same construct, indicating that the domains measured are, if not the same, quite similar. For example, in the Michigan School Readiness evaluation, the PQA and ECERS global scores were correlated at .86 (Xiang & Schweinhart, 2002). Two studies cited by Abbott-Shim, Lambert, and McCarty (2000) reported correlations between the Assessment Profile and the ECERS of .64 and .74. The ECERS-E and ECERS-R are correlated at .78 (Sylva, et. al., 1999).

The factors or subscales of these global measures of quality are also correlated. The ELLCO Classroom Observation score correlated .44 with the Learning Environment subscale of the Profile, as would be expected, but was not significantly correlated with Scheduling, also as expected (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). In a large study of public pre-k, the CLASS Emotional Support factor was highly correlated with the ECERS-R Teaching and Interactions factor (r=.58), but the CLASS Instructional Support factor was less correlated with Teaching and Interactions (r=.41), indicating that it measures a similar but somewhat different dimension than ECERS Teaching and Interactions (Early et al., 2006).

Policymakers frequently ask whether one classroom observational measure does a better job than others in measuring “good practice.” The relatively high correlations among these measures suggest, once again, that one’s choice should be based primarily on the specific domain(s) of information needed. Beyond that, concerns such as ease of training or effort needed to maintain reliability should be considered.
Cross-cultural Validity

As America becomes increasingly diverse, classroom quality observation data would be more useful programmatically and more accurate descriptively if our observational measures would reflect a program’s ability to provide culturally appropriate care and reinforce cultural values and heritage (Maher, 2007). Studies of cross-cultural validity exist for only one measure. Burchinal and Cryer (2003) showed that in the cultural variations found in the U.S., quality as measured by the ECERS was a good predictor of child outcomes. Studies in Western Europe (Clifford, 2005) and even in Bangladesh (Aboud, 2006) have demonstrated the relation between the ECERS and child outcomes. The CLASS, ECERS-E, ELLCO, PQA and Profile include items that address cultural sensitivity, but more thorough cross-cultural studies are needed. The lack of cross-cultural validity does not preclude use of measures other than the ECERS, but suggests doing so with awareness of a shortcoming. Meanwhile, new measures that focus solely on cultural sensitivity in early childhood settings are being developed (Castro, 2005).

Training and Reliability

The purchase cost of most observational quality measures is relatively inexpensive, if not free, but the costs of training observers and assuring their continued accuracy are realistic concerns for programs and policymakers. For training, funds may be needed for registration or trainer consultation fees, travel to training events, and 2-5 days typically needed for a trainee to obtain reliability with the trainer. To maintain reliability, observers should make ongoing joint observation visits to assure that they have not “drifted” from the standard item interpretation; otherwise, results could be contested. While most state QRSs include observational measures, cost of administration has been a deterrent in some instances. For example, Wisconsin policymakers considered observational measures for their QRS, but ruled them out because of these ongoing administration costs.

Training for the ECERS, PQA and CLASS are offered frequently by the authors and many well-trained individuals have become second-generation trainers in their region or state. For a person who is knowledgeable about early care and education, training on these measures takes about a week to achieve reliability required in research. Similar time is recommended for the Profile. The ELLCO can be self-taught in two days, according to the authors.

The availability of training manuals and other supports varies among measures. The ELLCO training manual is detailed and well-documented (Smith et al., 2002). The CLASS uses videotapes for training and recertification of trainers. The ECERS/ITERS include videotapes for training and comprehensive books with photos and examples; these have made reliability and PD using these measures much easier (Cryer, Harms, & Riley, 2003; Cryer, Harms, & Riley, 2004).

The ORCE is not widely used outside of the community of researchers who participated in the NICHD Study of Early Child Care, likely because it is complicated to learn and maintain reliability. It produces both quantitative scores and qualitative ratings and can thus contribute much to a research study. Lay people find data summaries from the Snapshot easy to understand, but it also requires extensive training and might be difficult to adopt in a state licensing system.
The rigor with which reliability has been demonstrated varies. Until recently, studies typically used a standard of two observers scoring 85% of individual items exactly the same or differing by only one point (e.g. one person scores 3, the other scores 4). On 5-point rating scales such as the PQA, the one-point-apart standard is very easy to meet and not considered a good reliability estimate. Even on 7-point scales, some studies more rigorously have used a standard of > 85% exact agreement (Epstein, 1999; Goelman, et al., 2006). The Cohen’s Kappa statistic is emerging as the preferred reliability method among researchers because it takes into account chance agreements. The standard of reliability should depend somewhat on the intended use of the data. For quality improvement programs or distinguishing between high and low quality, a within-one standard is probably sufficient; for research or licensing with consequences, our goal for reliability should be higher.

Although no rule mandates a certain percentage of visits to be conducted jointly, in research, inter-rater reliability is typically documented about every 10th visit. Even well-trained observers can drift in their interpretations of item scoring, especially if one sees mainly very poor quality programs and another collects data in very high quality programs. Budgeting time and travel for these joint visits is a data collection cost that must be considered.

Who should collect the observational rating data is one of the most important points of consideration for directors of PD programs and policymakers considering observations for QRSs. Ideally, observers have some background in early childhood education and the ability to code accurately according to the specific measure. As observations have become part of QRIS and licensing systems, some states have separated the observer role from the state rating or licensing agency to allow observers to focus solely on data collection and maintain their independence. An independent observer is also required for PD programs where consultants collect rating scale data and use it as the basis of program enhancement. Consultants’ observation accuracy depends on their level of training. Reliable consultants may be able to collect valid data at the beginning of a consultation, but after working closely with a provider, a consultant is surely too vested in the program and her work with staff to be considered an unbiased collector of post-consultation quality data. For valid data, the observer in any type of evaluation or ratings system must be independent of the program.

**Measurement of Classroom Change in Response to Intervention**

Witte and Queralt (2004) have shown that just making observational data available on a public website has small but significant effects on the overall quality of programs. What about specific interventions designed to enhance quality such as training or consultation? Are these observational measurements sensitive to change? Several studies of PD have shown changes in the ECERS or ITERS as a result of training, technical assistance, or consultation (Sakai, Whitebook, Wishard, & Howes, 2003; Palsha & Wesley, 1998; Wesley, 1994; Whitebook, Sakai, & Howes, 1997). Some quality enhancement interventions used the ECERS or ITERS as the basis for developing action plans to address areas of weakness, and indeed the endpoint observations (made by independent observers) showed improvement. A Heads Up Reading intervention where mentors focused on weak ELLCO items found classroom improvements on the ELLCO but also, unexpectedly, on the ECERS-R (Jackson, Larzelere, Clair, Corr, Fichter, & Egertson, 2006). All 5 subscales of the Profile showed treatment group differences in the K-3rd grade Head Start Transition demonstration classes (Ramey et al., 2000). Three domains of the CLASS showed treatment effects in a study of web-based consultation based on CLASS dimensions (Planta, Mashburn, Downer, Hamre, & Justice, 2008). These studies show that we have many observational measures that can reflect significant change in classroom practices as a result of technical assistance. Close alignment of the measure to the type of intervention can assure adequate assessment of improvement.
Predicting Child Outcomes from Classroom Observational Measures

Whether an observational tool relates to child outcomes is called criterion or predictive validity. Evidence from dozens of studies using the observational measures reviewed here shows that all of them have been related in a positive way to one or more aspects of children’s development, some to several outcomes in several studies (see Table 2 for exemplars).

Table 2

Child Outcomes Associated with Preschool Classroom Observation Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Child Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIS</td>
<td>Social initiations in 2-year-olds</td>
<td>Vernon-Feagans &amp; Manlove, 2005</td>
</tr>
<tr>
<td>CLASS Emotional Support</td>
<td>More social competence &amp; fewer problem behaviors</td>
<td>Mashburn et al., 2008</td>
</tr>
<tr>
<td>CLASS Instructional Support</td>
<td>Expressive &amp; receptive language &amp; math in pre-k</td>
<td>Mashburn et al., 2008</td>
</tr>
<tr>
<td>CLASS</td>
<td>Task-oriented behavior and aggression towards peers</td>
<td>Rimm-Kaufman et al., 2005</td>
</tr>
<tr>
<td>ECERS</td>
<td>Language &amp; academic skills in 2nd grade</td>
<td>Peisner-Feinberg et al., 2001</td>
</tr>
<tr>
<td>ECERS-R</td>
<td>Expressive language in pre-k</td>
<td>Mashburn et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Receptive language in pre-K &amp; K</td>
<td>Burchinal, Howes, et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Verbal &amp; non-verbal reasoning in preschool</td>
<td>Aboud, 2006</td>
</tr>
<tr>
<td></td>
<td>Pre-reading skills in preschoolers</td>
<td>Jackson et al., 2006</td>
</tr>
<tr>
<td></td>
<td>Cooperation, independence, concentration</td>
<td>Sylva, et al., 2006</td>
</tr>
<tr>
<td>ECERS-E</td>
<td>Pre-reading, math, reasoning in 5 yr. olds</td>
<td>Sylva, et al., 2006</td>
</tr>
<tr>
<td>ELLCO</td>
<td>Pre-reading skills &amp; vocabulary in preschools, Eng. &amp; Spanish-speaking</td>
<td>Jackson, et al., 2006</td>
</tr>
<tr>
<td>ORCE</td>
<td>Positive peer interactions at 36 mo</td>
<td>NICHD, 2001</td>
</tr>
<tr>
<td></td>
<td>Cognitive &amp; language scores at 54 mo.</td>
<td>NICHD, 2000 &amp; 2002</td>
</tr>
<tr>
<td>PQA</td>
<td>Cognitive scores in preschoolers</td>
<td>Epstein, 1999</td>
</tr>
<tr>
<td>Profile</td>
<td>Fewer problem behaviors</td>
<td>Lambert, Abbott-Shim &amp; McCarty, 2002</td>
</tr>
<tr>
<td></td>
<td>Print concepts &amp; story memory</td>
<td>Gallagher &amp; Lambert, 2006</td>
</tr>
<tr>
<td>Snapshot</td>
<td>Teacher ratings children’s language and literacy skills</td>
<td>Howes, et al. 2008</td>
</tr>
</tbody>
</table>

While the associations between quality and outcomes are significant, they are generally very modest, with family background characteristics typically accounting for much more of the variance in child outcomes than the classroom measure(s) (Burchinal et al., 2009). Nevertheless, given the amount of time children spend with families and the genetic influence of parenting, the fact that particular classroom practices can have a significant added effect on child outcomes is a notable finding and one on which to build pre-service and in-service training.

It would be unexpected for any single measure to be the best predictor of school readiness because we have so many different desired outcomes for children, and, even though these relationships are modest, it is reassuring that most studies show some relationships. Our most widely used measures of children’s classroom environments are describing well at least some of the conditions that are important for children’s development. Further efforts are currently underway to strengthen the measurement of quality so that those facets most important to children’s outcome are a focus of measurement.

**Summary**

This paper has identified key issues to take into account when selecting classroom quality measures as components of quality rating and PD systems. The increased use of assessment tools is commendable, provided that the process of selecting appropriate measures is thoughtful and closely tied to the purpose for their use. A detailed plan should address training, administration, reliability and objectivity of assessors. When financial stakes are placed on the results of quality assessments, communities must use measures as carefully as do researchers.

The authors thank Nancy Eisenberg and anonymous reviewers for their extremely helpful comments on earlier drafts of this research brief when under review by *Child Development Perspectives*. These comments were valuable in strengthening the brief.
References


Overview for OPRE Research Brief Series on Measuring Quality in Early Care and Education Settings

Measures to assess the quality of early care and education environments, originally developed as research tools and, in some cases, as guides for improving practice, now play a prominent role in the early childhood policy arena. Many states use information from on-site observations and environmental rating scales to make decisions about inclusion of programs in publicly funded initiatives and interventions, to target quality improvement dollars and to target incentives when programs meet higher quality standards. To date, the majority of states that have developed statewide Quality Rating Systems combine scores on observational measures of quality with other quality indicators to provide a rating that is available to the public. The intent is to provide better information to parents, and to provide a framework within which quality benchmarks, financial support, technical assistance, and monitoring create leverage for quality improvements in early care and education.

Yet the use of quality measures in “high-stakes” policy and programmatic decisions raises important new questions about their content, reliability, validity, and applicability with diverse populations across a broad range of settings. To address these questions, the Office of Planning, Research and Evaluation in the Administration for Children and Families of the U.S. Department of Health and Human Services and other federal partners convened a meeting of researchers, state policymakers, practitioners and other key stakeholders. The meeting provided a forum for analyzing current quality measures, engaging in critical discussion about the use of quality measures in the policy arena, and outlining the steps needed to improve measurement strategies.

The four coordinated research briefs in this series were developed based on presentations made at the meeting, with the intent of informing policymakers, researchers and practitioners about new developments in quality measurement being generated at the intersection of child development research and early childhood policy.

- The first paper (by Martha Zaslow, Kathryn Tout and Ivelisse Martinez-Beck) describes why and how quality measures are currently used in policy and practice contexts and the issues and concerns that arise as a result of this widespread use.

- The second paper (by Margaret Burchinal) reviews the literature on the dimensions of quality that have been measured in early care and education settings and identifies the quality dimensions that have received a more thorough treatment in the literature compared to those that have not been studied as extensively.

- The third and fourth papers review the quality measures that have been developed for use in center-based early care and education programs (paper by Donna Bryant) and home-based settings (paper by Barbara Goodson and Jean Layzer). In addition to highlighting the types of measures used, their psychometric properties, and their value in predicting child outcomes, the authors discuss the importance of the findings for policymakers and practitioners.

Overall, we hope that the four papers provide a useful review of the current state of the field of quality measurement and suggest important next steps that policymakers, researchers, and practitioners can take to assure the integrity of measurement strategies and the appropriate use of data on the quality of early care and education settings especially when measures are widely implemented in policy and practice initiatives.

Those interested in the issue of the measurement of quality in early childhood settings may also want to read these OPRE briefs:

