Dialog, *17(2)*, 105-120

ISSN: 1930-9325

DIALOG FROM THE FIELD

A Review of Empirical Evidence and Practical Considerations For Early Childhood Classroom Observation Scales

Lia E. Sandilos
Temple University

James C. DiPerna
The Pennsylvania State University

The current article presents a critical review of empirical evidence for six observation scales commonly used in practice to evaluate the quality of the early childhood classroom environment. Specifically, the theoretical foundation, content, and psychometric properties are reviewed for each scale. Based on the strengths and limitations of the evidence for each measure, recommendations are made regarding use of these specific systems in early education settings.

As research has indicated that improving teaching quality is crucial to students' academic success, early childhood education (ECE) has received increased attention and scrutiny (Bill & Melinda Gates Foundation, 2012; Zaslow, Martinez-Beck, Tout, & Halle, 2011). The emphasis on accountability in education (e.g., No Child Left Behind Act, 2001) also has led to growth in research examining specific teaching practices in early education that may influence later academic achievement, particularly for children considered to be at risk for poor outcomes (Zaslow et al., 2011). The identification of practical instruments that can provide reliable and valid data regarding "educational quality" is critical for bringing about positive changes in practice. One method commonly used to evaluate teaching practices is systematic classroom observation.

A number of systematic observation scales have been created to examine various aspects of instruction and classroom quality. These observation systems directly measure a wide range of teaching strategies, classroom and curricular resources, and administrative practices hypothesized to promote positive academic, social, and emotional growth in children. Data gathered from these scales have been used to identify best practices in the classroom, inform teacher professional development programs, and guide educational policy (Halle & Vick, 2007). However, selecting an appropriate observation scale for use in practice can be difficult given the range of available scales, and in some instances, the limited information available regarding each measure's theoretical and technical properties. As such, the purpose of this article is to provide a critical review of several observation systems available to assess early childhood classroom

environments and instructional practices.

Specifically, we systematically review six observation scales that assess various qualities and instructional practices within early childhood classrooms. In addition to providing a brief description of the content and constructs assessed, connections are made to developmental theories underlying each scale. Psychometric properties of scores also are reviewed in an attempt to identify strengths and limitations of observation scales measuring global classroom quality.

Several methods were used to identify the observation scales included in this review. First, a systematic search was conducted for relevant published literature via three electronic databases: PsycInfo, Proquest, and ERIC. Examples of descriptive terms used to conduct the literature searches included: *classroom, classroom quality, childhood, early childhood, ecological, pre-kindergarten, observation scale, observation schedule, observation system, teacher behavior, teaching methods, teaching quality.* These searches yielded a number of relevant studies (e.g., scale validation, empirical studies using classroom observation systems). In addition, three compendium reports on early childhood measures (Halle & Vick, 2007; Malone et al., 2010; Snow & Van Hemel, 2008) were consulted to identify prospective scales. Finally, the Buros Mental Measurements Yearbook was utilized to identify relevant scales that were not located through the initial search strategies.

Based on these search strategies, a total of 43 early childhood classroom observation scales were identified for initial consideration. To be retained for the review, each scale had to meet five specific criteria. Of the 43 instruments identified initially, 37 were excluded from this literature review because they failed to meet at least one of the required criteria. (A complete list of the excluded measures can be obtained by contacting the first author).
The specific inclusionary criteria and number of scales that failed to meet them (*n*) were as follows:

1. Developed for use in pre-kindergarten classrooms (*n* = 3).
2. Appropriate for use in early childhood classrooms across the U. S. (*n* = 1).
3. Assess global aspects of the classroom environment and instructional quality (*n* = 12).
4. Require direct observation in the classroom (*n* = 4).
5. Focused on classroom-level variables (i.e., whole class, teachers/caregivers; *n* = 17).

Thus, six scales were retained for inclusion based on the aforementioned criteria. The selected scales consisted of the following measures: Assessment Profile for Early Childhood Programs: Research Edition II (Assessment Profile; Abbott-Shim & Sibley, 1998), Child-Caregiver Interaction Scale (CCIS; Carl, 2007, 2010), Classroom Assessment Scoring System, Pre-K (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008), Early Childhood Classroom Observation Measure (ECCOM; Stipek & Byler, 2004), Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998), and the Preschool Classroom Implementation Rating Scale (PCI; Frede & Miller, 1990).

After identifying the observation scales for inclusion in the review, theoretical and psychometric (reliability and validity) evidence were examined for each scale. Reliability and validity of scores must be carefully considered when selecting assessments to use in research and practice (Gall, Gall, & Borg, 2007). Reliability is "the degree to which measurement error is absent from scores" yielded by a measure, and it is evidenced by the consistency of scores (Gall et al., 2007). Forms of score reliability reported in this study were interrater, internal consistency,

and test-retest. Interrater reliability is the agreement between the scores of two or more observers. Internal consistency reliability refers to the consistency of item scores within a single measure. Test-retest reliability (also referred to as stability) is the correlation between scores on the same measure at two different time points.

Validity is defined as "the degree to which evidence and theory support interpretation of test scores entailed by proposed uses of tests" (Standards for Educational and Psychological Testing; AERA, APA, NCME, 1999). The types of validity evidence considered in this review included concurrent, predictive, and structural. Concurrent validity is the extent to which the scores on a measure correlate positively with a criterion variable administered simultaneously. Predictive validity is an estimate of how accurately scores on one measure can predict a criterion variable obtained at some later time. Structural validity is the analysis of the way in which items/subscales on a measure reflect the constructs that they are purported to represent (Gall et al., 2007). To examine the psychometric evidence of the aforementioned observation scales, reliability and validity evidence are reported within the description of each measure. Specific descriptive labels for reliability and validity evidence have been applied consistently throughout this review to maintain uniformity among the descriptions of observation scales. Criteria for reliability were as follows: < .60 = unacceptable, .60 to .69 = marginally acceptable, .70 to .79 = partially acceptable, and ≥ .80 = acceptable (Nunally & Bernstein, 1994; Salvia & Ysseldyke, 2007; Sattler, 2001). Criteria for validity were as follows: < .30 = weak, .30 to .70 moderate, and > .70 = strong.

TABLE 1
Characteristics of Observation Scales of Early Childhood Classroom Quality

Observation Scale Key Characteristics					
Observation Scale	*Framework*	*Purpose*	*Domains Assessed*	*Age Range*	*Rating Format*
Assessment Profile: Research Edition II (Abbott-Shim & Sibley, 1998)	Developmental Systems, NAEYC DAP	Evaluate learning environment and teaching practices	Classroom: Learning Environment, Scheduling, Curriculum Methods, Interacting, Individualizing	Ages: 3 - 7 years (center-based child care, preschool programs, kindergarten classrooms)	12 items per subscale totaling to 60 items dichotomous items (yes/no)
Child-Caregiver Interaction Scale (CCIS; Carl, 2007)	Attachment, Constructivist, Ecological, NAEYC DAP	Assess caregiver interaction across age groupings and settings	Three domains: Emotional, Cognitive/Physical, Social	Infancy through school age (center- and home-based early childhood programs)	17-item Likert-type scale ranging from 1 (inadequate) to 7 (excellent)
Classroom Assessment Scoring System, Pre-K (CLASS; Pianta La Paro, &Hamre, 2008)	Developmental Systems, Attachment, Constructivist,	Assess the quality of the interactions between teachers and their students	Three domains: Emotional Support, Classroom Organization, & Instructional Support	Versions available for toddler, pre-kindergarten, primary, & secondary grades	10 dimensions rated on Likert-type scale ranging from 1 (low) to 7 (high); 30-minute observe and
Early Childhood Classroom Observation Measure (ECCOM; Stipek & Byler, 2004)	Constructivist, Didactic	Evaluate academic instruction, social climate, & resources; assess constructivist (child-centered) and didactic (teacher-centered) instructional approaches	Constructivist (child-centered) subscales: Instruction, Management, Social Climate Didactic (teacher-centered) subscales: Instruction, Management, Social Climate	Ages 4 – 7 (preschool, kindergarten, & first grade)	32 items rated from 1 (practices are rarely seen) to 5 (practices predominate)

Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms & Clifford, 1980; Harms, Clifford, & Cryer, 1998)	NAEYC DAP	Measures global quality of early childhood center-based programs	Seven subscales: Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interaction, Program Structure, Parents and Staff	Ages 2.5 – 5 years (preschool & kindergarten)	43-item Likert-type scale ranging from 1 (inadequate) to 7 (excellent); 2.5 – 3 hour observations (preferable)
Preschool Classroom Implementation Rating Scale (PCI; Frede& Miller, 1990)	Constructivist	Measures general quality factors for a cognitive-developmental classroom	Twelve subscales: Room Arrangements, Routine, Planning, Work/Free Play, Clean-up, Recall, Small Group, Outside, Circle, Teacher/Child Interactions, Classroom Management, Team Evaluation & Planning	Ages: 3 – 6 (preschool & kindergarten)	52 items rated as not observed, not evident, evident, or optimal; Authors suggest observers spend one full day in a classroom

TABLE 2
Reliability and Validity Evidence for Observation Scales of Early Childhood Classroom Quality

		Reliability Evidence			Validity Evidence		
Observation Scale	Sources	Interrater (+/-1)	Internal Consistency (α)	Test-Retest	Concurrent	Predictive	Structural
Assessment Profile Research Edition II, 1998	Abott-Shim & Sibely 1998	.83 - .91	.83 - .91[b]	-	.64 - .74[d]	.42[j], .54[k]	CFA indicated 5 first-order factors & 1 second-order factor
CCIS, 2007, 2010	Carl, 2007, 2010	.88 - .93[a]	.94	-	.67 – .75[d]	.62[l]	-
CLASS, 2008	Pianta, La Paro, & Hamre, 2008	.53 - 1.00	.76 - .89	.18 - .62[a]	.45 - .63[d]	-.35[m]	CFA indicated 3-factor model
ECCOM, 2004	Stipeck & Byler, 2004	.74 - .92[a]	.73 - .98	-	.30[e]; .37[f]; .41[g]	.49[n], .58[o], .67[p]	CFA indicated separate 1-factor models for constructivist and didactic scales[r]
ECERS-R, 1998	Harms et al., 1998	.71	.71 - .88	.69[c]	.60[h], .68[i]	.49[q]	CFAs indicated 1-, 2- and 3-factor models
PCI, 1990	Barnett et al., 1988; 2008	.94 - 1.00	.89	.93[d]	.60[e]	-	-

Note. [a]Intraclass correlation coefficient (ICC). [b]Item response theory. [c]Pearson correlation (*r*). [d]ECERS, ECERS-R. [e]Higher-order thinking skills. [f]Basic math skills. [g]Basic Literacy Skills. [h]ECERS-E Literacy subscale. [i]CLASS Pre-K Emotional Support domain. [j]Story Retell. [k]Print Concepts. [l]Keystone Stars Quality Rating. [m]Time off task. [n]Mathematics standard (*β*). [o]Reading fluency (*β*). [p]Letter-sound recognition (*β*). [q]Peabody Picture Vocabulary Test. [r]CFA conducted on Finnish and Estonian teachers.

To help facilitate the review of evidence for each measure, the guiding theoretical framework, purpose, domains assessed, age range, and rating format of each observation scale are displayed in Table 1. In addition, published reliability and validity evidence for each scale is reported in Table 2.

Assessment Profile for Early Childhood Programs: Research Edition II (Assessment Profile; Abbott-Shim & Sibley, 1998). The Assessment Profile is a global measure of the quality of an early childhood learning environment and teaching practices (Abbott-Shim & Sibley, 1998; Lambert, 2003). The Assessment Profile is aligned with NAEYC DAP (1997) standards (Lambert, 2003; Quality Assist, 2012). In addition, the framework of the Assessment Profile appears to reflect the developmental systems model learning perspective, which applies general systems and ecological theories to the classroom environment (Copple & Bredekamp, 2009; Pianta, 1999), and it emphasizes the interaction between the child, the teacher, and the environment. As shown in Table 1, the scale contains five primary classroom domains (i.e., Learning Environment, Scheduling, Curriculum Methods, Interacting, and Individualizing; Abbott-Shim & Sibley, 1998). The Assessment Profile, which can be used in classrooms with children ages 3 to 7 years, contains 60 dichotomous (*yes/no*) items.

The Assessment Profile Research Edition originally was developed using Item Response Theory (IRT; Abbott-Shim & Sibley, 1992). Average interrater reliability estimates and internal consistency coefficients fell in the acceptable range (Abbott-Shim, Lambert, & McCarty, 2000; Abbott-Shim & Sibely, 1992). The revised Assessment Profile Research Edition II was standardized on 2,820 classrooms across the U.S. (Abott-Shim & Sibely, 1998). With the revision of the Assessment Profile, second-order factor analysis was conducted to assess structural validity, and results indicated that the five observed scales stemmed from a single underlying construct of global classroom quality (Abbott-Shim et al., 2000). Concurrent validity analysis between the Assessment Profile and the ECERS yielded moderate to strong correlations (Abbott-Shim, 1991; Wilkes, 1989; see Table 2). Predictive validity evidence was identified in a 2006 study in which scores from the Assessment Profile were used to divide a sample of teachers into high, medium, and low quality classrooms, and the results indicated that children in the high quality classrooms scored significantly higher than lower quality classrooms on tests of print concepts and story retell (Gallagher & Lambert, 2006).

No published studies of long-term stability, however, are currently available for scores from the Assessment Profile: Research Edition II.

Child-Caregiver Interaction Scale, Revised Edition (CCIS; Carl, 2007, 2010**).** The CCIS is a global observation rating measure of teacher interactions with children (Table 1). The framework for the CCIS was developed to address limitations of the Arnett Caregiver Interaction Scale (CIS; Arnett, 1989). The theoretical framework of the CCIS is based upon child-caregiver attachment theory (Ainsworth & Bowlby, 1991; Bretherton, 1992) and socialization practices during parent-child interactions (Baumrind, 1991). In addition, ecological theory (Bronfenbrenner & Morris, 1998), constructivism (Kozulin, 1986), and research regarding early brain development were used to guide the development of the scale (Carl, 2007, 2010). The CCIS also reflects NAEYC DAP 2009 recommendations for developmentally appropriate practices (Carl, 2010). The CCIS consists of three primary interaction domains: Emotional/Interactional, Cognitive/Physical, and Social/Connections Within a Wider World. The

CCIS can be used in classrooms with children ranging from infancy to elementary. There are a total of 14 items on the scale that are rated from 1 (*inadequate care*) to 7 (*excellent care*).

An acceptable level of interrater reliability for CCIS scores was reported from training sessions in Carl's (2007) dissertation (Table 2). Internal consistency coefficients also were in the acceptable range (Carl, 2007). A strong concurrent relationship was observed between the CCIS and the "Interactions" subscale of the ECERS-R and an overall moderate correlation with the ECERS-R. Predictive validity analyses indicated that Keystone Star scores (state-awarded quality rating for early childhood classrooms) were found to be positive predictors of scores on the CCIS (Carl, 2007; Halle & Vick, 2007). Information regarding recent revisions to the CCIS indicated that construct validity was established by a panel of experts who reviewed the scale (Carl, 2010).

No published evidence has been reported regarding interrater reliability of CCIS scores during actual classroom observations, and there are no published data regarding the test-retest reliability of scores. In addition, structural analysis of CCIS has not been reported in the published literature to date.

The Classroom Assessment Scoring System, Pre-Kindergarten (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008). The CLASS Pre-K is an observation system intended to examine the quality of the interactions between teachers and their students(La Paro, Pianta, & Stuhlman, 2004). Toddler, elementary and secondary versions of CLASS are also available. Teaching quality on all versions of CLASS is assessed in terms of three major domains consisting of Emotional Support, Classroom Organization, and Instructional Support (Table 1). The primary theoretical focus of the CLASS framework is the developmental systems model of early learning (Pianta, 1999). However, within each domain attachment, behavioral, constructivist, and metacognitive theories are evident (Hamre, Pianta, Mashburn, & Downer, 2007). The CLASS Pre-K consists of 10 items coded on a scale of 1 (*low*) to 7 (*high*). As many as six cycles can be completed in one CLASS Pre-K observation, but the authors recommend a minimum of two cycles. One CLASS cycle consists of a 20-minute observation and 10 minutes of coding scores.

Interrater reliability for scores on CLASS Pre-K has varied with indices falling in the unacceptable to acceptable ranges across studies (Hamre, Mashburn, Pianta, & LoCasale-Crouch, 2008; Pianta et al., 2008; Sandilos & DiPerna, 2011). Internal consistency score reliabilities for 2, 3, and 4 cycles of CLASS for preschool and third grade classrooms range from partially acceptable to acceptable (Pianta et al., 2008). Stability analyses within a school year indicate that the Emotional Support domain demonstrates the highest levels of stability (Curby, Grimm, & Pianta, 2010). Classroom Organization and Instructional Support exhibit lower levels of stability, with Instructional Support consistently demonstrating the lowest score stability over time (Curby et al., 2010; Pianta et al., 2008). Structural validity of the CLASS framework has been tested in several studies in the United States and Finland (Downer et al., 2012; Hamre et al., 2007; Pakarinen et al., 2010). The three-factor structure (i.e., Emotional Support, Classroom Organization, Instructional Support) has demonstrated the best fit across validation studies (Downer et al., 2012; Hamre et al., 2007; Pakarinen et al., 2010). Concurrent validity analyses have yielded moderate correlations with the ECERS-R (Pianta et al., 2008; see Table 2). Predictive validity evidence has indicated that scores on the Emotional Support domain demonstrate a moderate negative relationship with the amount of time children are observed

being off-task (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009) and a positive relationship with growth in sound awareness skills (Curby, Rimm-Kaufman, & Ponitz, 2009).

Limited evidence of moderate to strong predictive validity of CLASS scores with social-emotional and academic outcomes across grade levels was identified. In addition, research regarding the CLASS should continue to assess the stability of scores over time, as certain domains (e.g., Instructional Support) have exhibited low levels of reliability in previous research (Curby et al., 2010).

Early Childhood Classroom Observation Measure (ECCOM; Stipek & Byler, 2004). The ECCOM originally was developed for educational-quality research. The theoretical framework for the scale is based on both constructivist (child-centered; Kozulin, 1986) and didactic (teacher-centered) theories of learning (Halle & Vick, 2007; Stipek & Byler, 2004). The ECCOM measures classroom quality through Instruction, Social Climate, and Management subscales (Table 1). Preschool through first grade classrooms may be observed with the ECCOM. A total of 32 items on the ECCOM are rated from 1 (*practices are rarely seen*) to 5 (*practices predominate*).

The psychometric properties of the ECCOM have been examined in the United States, Finland, and Estonia. Interrater agreement for the ECCOM, both in the United States and abroad, fell in the acceptable range, and internal consistency indices ranged from partially acceptable to acceptable across subscales (Halle & Vick, 2007; Lerkkanen et al., 2012; Stipek & Byler, 2004; see Table 2). Concurrent validity analyses indicated a moderate positive relationship between the ECCOM constructivist (child-centered) subscale and teachers' ratings of students' higher-order thinking skills, as well as a moderate positive relationship between the didactic (teacher-centered) subscale and teachers' ratings of basic literacy and math skills in their classroom (Stipek & Byler, 2004). Predictive validity analyses indicated that first grade teachers who received higher ECCOM ratings on both instructionally- and emotionally-supportive child-centered practices had a higher percentage of students who met end-of-year standards in letter-sound recognition, reading fluency, and mathematics (Perry, Donohue, & Weinstein, 2007). Structural validity of ECCOM has been examined with a sample of Finnish and Estonian classrooms, and the results indicated that separate one-factor models (child-centered & teacher-centered) best fit the data (Lerkkanen et al., 2012).

No evidence of test-retest reliability has been reported in published literature regarding the ECCOM. Additionally, no structural validity evidence for the use of the ECCOM with classrooms in the United States was identified through searches of published literature.

Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998). The ECERS-R is a widely used instrument that assesses characteristics of preschool, kindergarten, and child-care programs. The framework of the original ECERS (Harms & Clifford, 1980) was based on research regarding developmentally appropriate practices at the time when the licensing and accreditation process for ECE programs was first being established (Sakai, Whitebook, Wishard, & Howes, 2003). The 1998 revision of the ECERS-R utilized the NAEYC 1997 DAP guidelines as the primary conceptual framework (Harms et al., 1998). As a result, the ECERS-R assesses seven distinct components of the classroom: Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff (Table 1). The ECERS-R features 43 items

rated from 1 (*inadequate*) to 7 (*excellent*; Harms et al., 1998). A 2.5- to 3-hour observation period is recommended by the authors (Harms et al., 1998).

Reported interrater reliability for the ECERS-R fell in the partially acceptable to acceptable ranges (Harms et al., 1998; Denny, Hallam, & Homer, 2012). Internal consistency reliability estimates for scores on subscales also ranged from partially acceptable to acceptable (Harms et al., 1998). Average test-retest reliability fell in the marginally acceptable range (Clifford, 2005). Concurrent validity analyses indicated that the ECERS-R Interactions subscale demonstrated a high-moderate correlation with the CLASS Pre-K Emotional Support domain, and the ECERS-R Language and Reasoning subscale correlated moderately with the ECERS-Extension Literacy subscale (Denny et al., 2012). Predictive validity analyses yielded a moderate positive correlation with scores on the Peabody Picture Vocabulary Test (Harms et al., 1998), and a panel of experts reviewed the revised scale to examine construct validity (Zaslow et al., 2011). Factor analytic findings regarding the structural validity of the ECERS-R have been inconsistent, as different studies yielded one or two domains of classroom quality, as opposed to the seven distinct aspects of classroom quality suggested by the authors (Perlman, Zellman, & Vi-Nhuan, 2004; Snow & Van Hemel, 2008). Most recently, Gordon, Fujimoto, Kaestner, Korenman, and Abner (2013) identified a 3-factor solution (Table 2).

No evidence was found to support the current factor structure of the published version of the ECERS-R. Also, minimal research was found identifying moderate to strong associations between ECERS-R scores and child outcomes.

Preschool Classroom Implementation Rating Scale (PCI; Frede & Miller, 1990). The PCI originally was created as a measure of fidelity for the High/Scope Perry Preschool curriculum but subsequently was revised for use in all preschool and kindergarten programs (Frede & Miller, 1990; Halle & Vick, 2007). The framework for both the High/Scope Perry Preschool curriculum and the PCI was based on Vygotsky's work in constructivist philosophy (Kozulin, 1986) and Piaget's research on cognitive development (Piaget, 1952). The PCI is composed of 12 subscales (i.e., Room Arrangements, Daily Routine, Planning, Work/Free Play, Clean-up, Recall, Small Group, Outside, Circle, Teacher/Child Interactions, Classroom Management and Organization, Team Evaluation & Planning). The 52 items of the PCI are rated using a 4-point continuum: *not observed, not evident, evident, or optimal* (Table 1). The scale can be used in preschool and kindergarten classrooms, and the authors recommend spending a full day in a classroom to achieve the most accurate ratings (Frede & Miller, 1990).

Interrater reliability coefficients reported from training and real-time observations fell in the acceptable range (Barnett et al., 2008). The test-retest and internal consistency reliability estimates also fell in acceptable ranges (Barnett, Frede, Mobasher, & Mohr, 1988; Barnett et al., 2008). Regarding concurrent validity, a moderate relationship was identified between the PCI and ECERS-R (Barnett et al., 2008; see Table 2).

It should be noted that there is limited published empirical evidence of acceptable reliability and validity of PCI scores in non-constructivist-based classrooms; previous research indicated that classrooms with constructivist-based curricula (i.e., High/Scope Perry Preschool, Tools of the Mind) scored higher on the PCI than non-constructivist-based classrooms (Barnett et al., 1988; Barnett et al., 2008). No predictive or structural validity evidence has been reported in the published literature to date.

STRENGTHS AND LIMITATIONS OF REVIEWED CLASSROOM OBSERVATION SCALES

Across the six observation scales that met criteria for inclusion in this review, several overarching strengths emerged. First, many of these scales used key developmental theory and/or prior empirical research to inform the scale development process. The most commonly represented frameworks were constructivist learning theory, attachment theory, developmental-systems model, and NAEYC DAP guidelines. Second, all of the scales place some emphasis on caregiver/teacher and child interactions as a focal point of the observation; a practice that is theoretically justified by prior empirical research (Baumrind, 1991; Pianta, 1999). Moreover, the inclusion of emotional climate and adult-child interactions are strong characteristics for observation scales, as teacher-child relationships have been identified as an important aspect of early academic success (La Paro & Pianta, 2000; Pianta, La Paro, Payne, Cox, & Bradley, 2002).

Third, interrater agreement and internal consistency reliability were reported for all of the scales, and indices fell primarily in partially-acceptable to acceptable ranges. In addition, moderate to strong concurrent validity evidence also has been demonstrated for all of the scales. Finally, on a practical level, many of the reviewed scales can be used in a variety of preschool environments (i.e., home care, child care, center-based) and up through the early elementary grades (Kindergarten, 1st grade, 2nd grade); though additional evidence for use across grades/age ranges is still needed for some measures. The scales also allow for a range of time during which an observation can be conducted (e.g., 30 min – 3 hours).

Beyond these collective strengths, there are some common limitations shared across multiple measures included in the review. For example, three of the scales were published nearly 15 years ago (i.e. Assessment Profile, ECERS-R, PCI). The evolution of early childhood theory, research, and practice during this time period may have implications for the validity of interpretation of results from these measures. In addition, as shown by the variability in factors assessed across measures, there is a lack of consensus on an operational definition of "classroom quality." While some observation systems place emphasis on daily scheduling, material resources, and physical structure of the environment (i.e., ECERS-R, PCI), other scales focus largely on specific areas of quality such as interpersonal interactions and/or didactic techniques (i.e., ECCOM, CCIS, Assessment Profile). Of the systems included in this review, the CLASS Pre-K and ECCOM appear to demonstrate the most balance among socio-emotional, instructional, and behavioral elements of high quality classrooms.

With regard to psychometric properties, each scale is lacking published evidence regarding at least one type of reliability or validity. Evidence of test-retest stability and structural validity could not be located for many of the observation scales. Specifically, test-retest data were missing for three of the scales (i.e., Assessment Profile, CCIS, ECCOM), factor analytic methods were not used to examine the internal structure of two scales (i.e., CCIS & PCI), and structural results were inconsistent for the ECERS-R. For several of the measures, interrater reliability coefficients were based on data collected during observer training as opposed to actual data from real-time observations (i.e., PCI, CCIS). In addition, there was limited evidence indicating strong predictive relationships between scores on the reviewed scales to outcome variables, as measures often yielded moderate correlations with academic or socio-emotional outcomes.

RECOMMENDATIONS FOR PRACTIONERS

The goal of this review was to examine observation scales that can be used to assess global quality in early childhood classrooms. Recommendations for practice are grouped into three categories reflecting aspects of the early childhood classroom environment that practitioners may wish to examine.

Material Resources and Physical Structure. The ECERS-R and PCI directly examine concrete material resources in the classroom. The reliability of scores on both scales indicates that ratings are fairly consistent across observations, which may be related to the objective nature of many of the items (e.g., material *present* or *not present*). The inconsistent results of structural analyses of the ECERS-R and the lack of structural evidence for the PCI are significant limitations for both scales. However, research continues to be conducted on the ECERS-R and there is substantial concurrent validity evidence linking scores on ECERS-R subscales to those on similar measures of classroom quality. The ECERS-R may be best used as a measure of physical quality and quantity of materials in the classroom, as it does not measure instructional practices. The PCI may be most effective in capturing high quality teaching and instructional resources in constructivist-based classrooms (Barnett et al., 2008).

Teacher-Child Relationship. Although all of the scales address, to some extent, the teacher-child relationship, the CCIS and the CLASS Pre-K appear to most comprehensively assess this key aspect of classroom quality. Both the CCIS and CLASS Pre-K were developed within the past 5 years, and they provide domains focusing on emotional and social interactions between caregivers and students. However, additional research is needed regarding the psychometric properties of the CCIS, as no other published data have been identified since the 2007 validation study. Conversely, although CLASS Pre-K has the most recent publication date of the scales included in this review, it is one of the more thoroughly researched measures. In particular, studies of structural validity in different countries and with varying samples have consistently supported the presence of three primary domains (Downer et al., 2012; Hamre et al., 2007; Pakarinen et al., 2010).

Academic Instruction and Didactic Practices. Several of the scales reviewed in this study assess teachers' instructional practices. The Assessment Profile, CLASS Pre-K, and ECCOM all examine various aspects of academic instruction, and have demonstrated moderate relationships with academic outcomes. The three measures appear to present global and theoretically supported views of instructional practices; however, there are unique characteristics of each scale that should be considered in the selection process. In a review of the Assessment Profile, Snow and Van Hemel (2008) identified the dichotomous format of the scale as a limitation, as teachers may receive credit for specific instructional practices, but the frequency and type of instruction would remain unclear due to the truncated method of scoring (Snow & Van Hemel, 2008). With regard to the ECCOM, the Individualizing subscale requires that educational documents be examined (e.g., referrals, assessments, and parent-teacher communication), which is an important consideration as this information may be difficult to obtain in some observation settings (Snow & Van Hemel, 2008). One critique of the Instructional

Support domain of CLASS Pre-K has been the absence of evaluation of major academic areas (i.e., reading, math, science, etc.; Snow & Van Hemel, 2008). Thus, it is important that users carefully consider what aspects of instruction they wish to observe before selecting a scale.

DIRECTIONS FOR FUTURE RESEARCH & DEVELOPMENT

This review provides information about early childhood observation scales that are currently available for use in primary classrooms, as well as some insight regarding future research needed to substantiate the use of all reviewed measures. Several conclusions can be drawn from this review. First, there are multiple components of classroom quality that can be measured, so it is important to examine the constructs emphasized by each scale (e.g., interactions, materials, etc.). Second, practical aspects of each scale should be considered within the context of the classroom and constraints of the observation before selecting a specific measure for use (e.g., length of observation, appropriate grade level, types of information/documentation needed). Third, several of the scales need to be updated to reflect current research and theory regarding early childhood education. Fourth, there are both strengths (e.g., internal consistency, concurrent validity) and gaps (e.g., structure, stability) in the psychometric evidence for many of the scales. Finally, the reviewed scales primarily have demonstrated moderate relationships with academic and socio-emotional outcomes, but all of the measures would be substantiated by further research linking observational teacher-quality data to student achievement and socio-emotional development.

Data from observation systems can be used to promote effective instruction for children. However, early childhood educators need to be confident that recommendations resulting from assessments of classroom quality are valid and informed by empirically sound data collection methods. Thus, it is important for educational practitioners and researchers to be cognizant of the strengths and limitations of each observation scale. An assessment of "global classroom quality" may be most accurate and comprehensive when observational ratings are considered in conjunction with other classroom-based data (e.g., student surveys, achievement data; Bill & Melinda Gates Foundation, 2012). When selecting a scale, practitioners should take time to consider the measure that is most appropriate for their needs based on the measure's content, theoretical foundation, and empirical evidence.

REFERNCES

Abbott-Shim, M. (1991). Quality care: A global assessment. Unpublished manuscript. Georgia State University.

Abbott-Shim, M., Lambert, R., & McCarty, F. (2000). Structural model of Head Start classroom quality. *Early Childhood Research Quarterly, 15*, 115-134. doi:10.1016/S0885-2006(99)00037-X

Abbott-Shim, M., & Sibely, A. (1992). *Assessment Profile for Early Childhood Programs: Research Edition I.* Atlanta, GA: Quality Assist.

Abbott-Shim, M., & Sibley, A. (1998). *Assessment Profile for Early Childhood Programs: Research Edition II.* Atlanta, GA: Quality Counts, Inc.

AERA, APA, NCME (1999). *The standards for educational and psychological testing*. Washington DC: AERA Publications.

Ainsworth, M. D. S., & Bowlby, J. (1991). An ethnological approach to personality development. *American Psychologist, 46*, 331-341. doi:10.1037//0003-066X.46.4.333

Arnett, J. (1989) Caregivers in day-care centers: does training matter? *Journal of Applied Developmental*

Psychology, 10, 541-552.

Barnett, S. W., Frede, E. C., Mobasher, H., & Mohr, P. (1988). The efficacy of public preschool programs and the relationship of program quality to efficacy. *Educational Evaluation and Policy Analysis, 10*, 37-49. doi:10.3102/01623737010001037

Barnett, S. W., Jung, K., Yarosz, D. J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly, 23*, 299-313. doi:10.1016/j.ecresq.2008.03.001

Baumrind, D. (1991). The influence of parenting style on adolescent competence and substance use. *Journal of Early Adolescence*, *11*, 56-95. doi:10.1177/0272431691111004

Bill & Melinda Gates Foundation (2012). Gathering feedback for teaching. *The MET Project.* Retrieved from www.gatesfoundation.org

Bretherton, I. (1992). The origins of attachment theory: John Bowlby and Mary Ainsworth. *Developmental Psychology, 28*, 759-775. doi: 10.1037//0012-1649.28.5.759

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes. In W. Damon & R. M Lerner (Eds.), *Handbook of child psychology*: *Theoretical models of human development* (5th ed., pp. 993-1028). New York: John Wiley & Sons.

Carl, B. (2007). *Child Caregiver Interaction Scale*. (Doctoral dissertation, Indiana University of Pennsylvania). Retrieved from http://dspace.lib.iup.edu:8080/dspace/bitstream /2069/53/1/Barbara%20Carl.pdf

Carl, B. (2010). *Child Caregiver Interaction Scale, Revised Edition.* [Online PowerPoint presentation, Early Childhood Training Institute]. Retrieved from ecti.hbg.psu.edu /ucpc/docs/CCIS2010-Overview.ppt

Clifford, R. (2005). Structure and stability of the Early Childhood Environment Rating Scale. In H. Schohenfeid, S. O'Brien, & T. Walsh (Eds). *Questions of quality*. Dublin, Ireland: St. Patrick's College.

Copple, C., & Bredekamp, S. (Eds.). (2009). *Developmentally appropriate practice in early childhood programs: Serving children from birth through age 8* (3rd ed.). Washington, DC: National Association for the Education of Young Children.

Curby, T. W., Grimm, K., J., & Pianta, R. C. (2010). Stability and change in early childhood interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25,* 373-384. doi:10.1016/j.ecresq.2010.02.004

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101,* 912-925. doi:10.1037/a0016647

Denny, J. H., Hallam, R., & Homer, K. (2012). A multi-instrument examination of preschool classroom quality and the relationship between program, classroom, and teacher characteristics, *Early Education and Development, 23,* 678-696. 10.1080/10409289.2011.588041

Downer, J. T., Lopez, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly, 27,* 21-32. doi: 10.1016/j.ecresq.2011.07.005

Frede, E. C., & Miller, A. K. (1990). *Preschool Classroom Implementation Rating Instrument: High/Scope manual*. Unpublished manuscript.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston: Pearson Education.

Gallagher, P. A., & Lambert, R., G. (2006). Classroom quality, concentration of special needs, and child outcomes in Head Start. *Exceptional Children, 73*, 31-52.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development, *Developmental Psychology, 49,* 146-160. doi: 10.1037/a0027899

Halle, T., & Vick, J. E. (2007). *Quality in early childhood care and education settings: A compendium of measures*. Washington, DC: Prepared by Child Trends for the Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from http://www.childtrends.org/Files/Child_Trendsm2007_12_10_FR_CompleteCompendium.pdf

Hamre, K. B., Mashburn, A. J., Pianta, R. C., & LoCasale-Crouch , J. (2008). *Classroom Assessment Scoring System, Pre-K: Technical appendix*, Baltimore, MD: Paul H. Brookes Publishing Co.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 early childhood and elementary classrooms.* Retrieved from http://www.icpsr.umich.edu/icpsrweb/PREK3RD/resources/507559.jsp

Harms, T., & Clifford, R. M. (1980). *Early Childhood Environment Rating Scale*. New York, NY: Teachers College Press.

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

Kozulin, A. (1986). The concept of activity in Soviet psychology: Vygotsky, his disciples and critics. *American Psychologist, 41,* 264-274. doi:10.1037//0003-066X.41.3.264

Lambert, R. G. (2003). Considering purpose and intended use when making evaluations of assessments: A response to Dickinson. *Educational Researcher, 32*, 23-26. doi: 10.3102/0013189X032004023

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70,* 443-484. doi:10.1016/j.jsp.2006.01.003

La Paro, K. M., Pianta, R., C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the pre-kindergarten year. *The Elementary School Journal, 104,* 409-426. doi: 10.1086/499760

Lerkkanen, M. , Kikas, E. , Pakarinen, E. , Trossmann, K. , Poikkeus, A., Rasku-Puttonen, H., Siekkinen, M., & Nurmi, J. (2012). A validation of the Early Childhood Classroom Observation Measure in Finnish and Estonian kindergartens. *Early Education and Development, 23,* 323-350. doi: 10.1080/10409289.2010.527222

Malone, L. M., Cabili, C., Henderson, J., Esposito, A. M., Coolahan, K.,…Boller, K. (2010). *Compendium of student, teacher, and classroom measures used in NCEE evaluations of educational interventions, Vol. II.* (NCEE 2010-4013). Washington, DC: U.S. Department of Education.

National Association for the Education of Young Children. (1997). *Developmentally appropriate practice in early childhood programs serving children from birth to age 8.* NAEYC position statement. Retrieved from http://www.naeyc.org/positionstatements/dap

National Association for the Education of Young Children. (2009). *Developmentally appropriate practice in early childhood programs serving children from birth to age 8.* NAEYC position statement. Retrieved from http://www.naeyc.org/positionstatements/dap

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2008).

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. (2010). A validation of the Classroom Assessment Scoring System in Finnish kindergartens. *Early Education and Development, 21,* 95-124. doi:10.1080/10409280902858764

Perlman, M., Zellman, G. L., & Vi-Nhuan, L. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly, 19,* 398-412. doi: 10.1016/j.ecresq.2004.07.006

Perry, K. E., Donohue, K. M., & Weinstein, R. S. (2007). Teaching practices and the promotion of achievement and adjustment in first grade. *Journal of School Psychology, 45*, 269–292. doi:10.1016/j.jsp.2007.02.005

Piaget, J. (1952). *The origins of intelligence in children.* New York: International Universities Press.

Pianta, R. C. (1999). Enhancing relationships between children and teachers. Washington, DC: American Psychological Association. doi: 10.1037/10314-000

Pianta, R. C., La Paro, K. M. & Hamre, B. K. (2008). *The Classroom Assessment Scoring System Manual, Pre-K*. Baltimore: Brookes Publishing Co.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relationship of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal, 102*, 225-238. doi:10.1086/499701

Quality Assist (2012). *The Assessment Profile*. Retrieved from http://www.qassist.com/pages/research-and-evaluation

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*, 958-972. doi:10.1037/a0015861

Sakai, L., Whitebrook, M., Wishard, A., & Howes, C. (2003). Evaluating the Early Childhood Environment Rating

Scale (ECERS): Assessing differences between the first and revised edition. *Early Childhood Research Quarterly, 18,* 427-445. doi:10.1016/j.ecresq.2003.09.004

Salvia, J. & Ysseldyke, J. E. (2007). *Assessment in Special and Inclusive Education* (10th ed.). New York: Houghton Mifflin Company.

Sandilos, L. E., & DiPerna, J. C. (2011). Interrater Reliability of the Classroom Assessment Scoring System Pre-K (CLASS Pre-K). *Journal of Early Childhood and Infant Development, 7,* 65-85.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). Jerome Sattler Publishers, Inc.

Snow, C. E., & Van Hemel, S. B. (2008). Early childhood assessment: Why, what, and how. *National Academy of the Sciences.* Retrieved from http://www.nap.edu/openbook.php? record_id=12446&page=1

Stipek, D., & Byler, P. (2004). The Early Childhood Classroom Observation Measure. *Early Childhood Research Quarterly, 19,* 375-397. doi:10.1016/j.ecresq.2004.07.007

Wilkes, D. (1989). *Administration, classroom program, sponsorship: Are these indices of quality care in day care centers?* (Doctoral Dissertation, Georgia State University). Retrieved from http://ezaccess.libraries.psu.edu/login?url=http://search.proquest.com.ezaccess.libraries.psu.edu/docview/303784213?accountid=13158

Zaslow, M., Martinez-Beck, I., Tout, K., Halle, T. (2011). *Quality measurement in early childhood settings*. Baltimore, MD: Paul H. Brookes Publishing Co.