Early Childhood Research Quarterly 44 (2018) 242–256

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Factor structure and validity of the Early Childhood Environment Rating Scale – Third Edition (ECERS-3)

Diane M. Early[a,*], John Sideris[b], Jennifer Neitzel[a], Doré R. LaForett[a], Chelsea G. Nehler[a]

[a] Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill, United States
[b] Chan Division of Occupational Science and Occupational Therapy, University of Southern California, United States

ARTICLE INFO

Article history:
Received 31 May 2017
Received in revised form 9 April 2018
Accepted 17 April 2018

Keywords:
Classroom quality
Child care
Preschool
Factor analysis
Validity
Measurement

ABSTRACT

The Early Childhood Environment Rating Scale – Third Edition (ECERS-3) is the latest version of one of the most widely used observational tools for assessing the quality of classrooms serving preschool-aged children. This study was the first assessment of its factor structure and validity, an important step given its widespread use. An ECERS-3 observation was conducted in 1063 preschool classrooms in three states. In a subset of those classrooms ($n = 119$), Classroom Assessment Scoring System – Pre-K (CLASS Pre-K) and child assessment data were also collected. Analyses of the ECERS-3 suggested that a single factor does not adequately capture item variability. Of the solutions tested, the four-factor (Learning Opportunities, Gross Motor, Teacher Interactions, and Math Activities) provided the best combination of statistical support and theoretical utility. In general, the ECERS-3 Total Score and the four factors were moderately correlated with the three domains of the CLASS Pre-K. ECERS-3 Total Score, Learning Opportunities, and Teacher Interactions were positively related to growth in executive function, as were all three domains of the CLASS Pre-K. However, all significant associations were small, and most tested associations between ECERS-3 scores and children's growth, and between CLASS Pre-K and children's growth, were not significant. Results are discussed in terms of their implications for measuring preschool quality.

1. Introduction

The Early Childhood Environment Rating Scale – Third Edition (ECERS-3; Harms, Clifford, & Cryer, 2015), published in 2015, is the latest version of one of the most widely used observational tools for assessing the quality of classrooms serving preschool-aged children in the United States and around the world. It is designed to measure all domains of quality, including physical space, groupings, materials, instruction, health, and safety. The ECERS-3's predecessor, the Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 2005) is a major component of quality measurement in nearly all of the Quality Rated and Improvement Systems (QRIS) across the United States (Administration for Children and Families, 2015; Tout et al., 2010) and has been used in most major national studies of early childhood, including Head Start Family and Child Experiences Survey (FACES; Moiduddin, Aikens, Tarullo, West, & Xue, 2012) and the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B; National Center for Education Statistics, n.d.). The ECERS-R has also been used extensively by researchers to evaluate state-funded pre-k programs (Early et al., 2007). Many researchers, evaluators, and QRIS are either transitioning to or considering a transition to ECERS-3. The current study provides the first large-scale exploration of the ECERS-3's factor structure and its concurrent, divergent, and predictive validity.

1.1. Push for increased early childhood program quality

Early childhood classroom quality has received increased national attention in the past two decades because advocates and practitioners see it as a key strategy for encouraging school readiness and narrowing the achievement gap (Klein & Knitzer, 2006; National Association for the Education of Young Children, 2009; National Education Association, 2008). The term *quality* in early childhood education is used to describe the structural and process features of the environment that promote learning and development. Structural quality refers to distal factors in early childhood settings such as staff:child ratio, group size, teacher education and training, and staff wages and benefits. These factors are often considered as necessary supports to promote process quality, but are not sufficient on their own to optimize children's learning

* Corresponding author at: FPG CDI, CB #8180, University of North Carolina, Chapel Hill, NC 27599-8180, United States.
E-mail address: diane_early@unc.edu (D.M. Early).

https://doi.org/10.1016/j.ecresq.2018.04.009

(Burchinal, 2017). Process quality, on the other hand, refers to the proximal interactions among children, adults, and the environment that are thought to directly impact children's growth (Burchinal, 2017; Vandell & Wolfe, 2000). Classrooms that are high in process quality present children with rich opportunities to interact with adults, peers, and materials. All versions of the ECERS, and the other widely used early childhood classroom quality measure, the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), are designed to measure process quality.

Many state and local governments have created QRIS to boost both structural and process quality in early childhood programs (Administration for Children and Families, 2014; Tout et al., 2010) with the ultimate goal of improving child outcomes (Elicker & Thornburg, 2011). QRIS assign ratings to programs based on multiple factors and provide supports to improve quality, such as materials, professional development, funds for infrastructure improvements, or increased subsidies for low-income children whose families select higher rated programs.

On a national level, the U.S. Department of Education and the U.S. Department of Health and Human Services partnered to improve early childhood program quality and access through their Race to the Top-Early Learning Challenge grants (RTT-ELC; U.S. Department of Education, 2013). With this initiative, twenty states received funds to build on the strengths of their existing early childhood programs (e.g., child care, Head Start programs, publicly funded pre-k) while also working to reduce inefficiency, improve quality, and deliver a coordinated set of services. As with QRIS, the goal of this quality improvement effort is to promote school readiness (U.S. Department of Education, 2013).

State and national quality improvement initiatives typically combine several sources of information to gauge quality and monitor improvement, but many rely heavily on classroom observations, using tools like the ECERS-R or the CLASS. In fact, the term *quality* is often used interchangeably with ECERS-R or CLASS scores, with quality being informally defined as "whatever the ECERS-R or CLASS measures." La Paro, Thomason, Lower, Kintner-Duffy, and Cassidy (2012) made this point about the ECERS-R, but the same issue has arisen with the CLASS as more systems adopt that tool.

1.2. Research linking program quality and child outcomes

The tautology of conflating the definition of quality with the tools used to measure it poses a problem for the field because the measures of quality do not map neatly onto the expected outcomes for young children. It is unclear, however, if the problem lies in the tools or in the underlying idea that quality and growth in early academic and social skills are linked.

Several reviews and position papers back an association between classroom quality and growth in academic and social skills (Shonkoff & Phillips, 2000; Vandell, 2004; Yoshikawa et al., 2013); however, the findings are mixed, and when the findings are significant the associations are generally small (Burchinal, 2017; Burchinal, Kainz, & Cai, 2011). For instance, in a study designed to inform QRIS star-rating cutpoints, Le, Schaack, and Setodji (2015) found a non-linear relation between ECERS-R Total Score and children's cognitive skills and a linear relation with social skills. On the other hand, Gordon, Fujimoto, Kaestner, Korenman, and Abner (2013), and Mashburn et al. (2008) found no associations between ECERS-R Total Score and early academic skills.

As an alternative to using the Total Score in predictive models, some researchers factor analyzed the ECERS-R, but this strategy has also produced mixed and small associations with children's outcomes. These factors were labeled *Activities/Materials* and *Language/Interactions* (Cassidy, Hestenes, Hedge, Hestenes, & Mims, 2005) or *Provisions for Learning* and *Teaching and Interactions* (Clifford et al., 2005). On one hand, Auger, Farkas, Burchinal, Duncan, and Vandell (2014) found that *Provisions for Learning* was significantly associated with growth in vocabulary ($d = .14$), but Howes et al. (2008) and Weiland, Ulvestad, Sachs, and Yoshikawa (2013) did not. Moreover, Howes et al. (2008) did find that the *Teaching and Interactions* factor was significantly related to vocabulary growth ($d = .08$), but Auger et al. (2014) and Weiland et al. (2013) did not.

These mixed or weak associations are not simply a problem with the ECERS-R. In a meta-analysis of the relationship between CLASS and children's outcomes, Perlman et al. (2016) found a few small associations with executive function and social skills (pooled correlation coefficients between .06 and .09), but no associations with traditional early academic skills such as vocabulary, literacy, and math. These mixed findings and weak associations have led researchers to question whether the current tools for measuring quality are inadequate, if there are weaknesses in the assessments used to evaluate children's outcomes, or if early childhood classroom quality in fact plays a limited role in children's academic growth (Burchinal et al., 2011).

Some state QRIS, program evaluators, and researchers are turning to the ECERS-3 in hopes that it will have stronger associations with children's outcomes. Indeed, in the introductory materials to the ECERS-3, the authors note that this new edition was designed to "improve the prediction of child outcomes while maintaining the emphasis on the importance of a wide range of developmental outcomes in children" (Harms et al., 2015, p. 2). The current study aims to further the discussion about quality measurement and answer some pressing questions about the validity of the ECERS-3.

1.3. Similarities and differences between ECERS-R and ECERS-3

The ECERS-R and ECERS-3 share many common features. Both tools cover the broad range of children's developmental needs, including cognitive, social–emotional, physical, health, and safety, and they share a common scoring strategy and structure. For both tools, trained observers who have demonstrated that they can reliably use the tool spend several hours observing in early childhood classrooms. During this time, observers respond to a series of yes/no *indicators* that are anchored to a particular place on a 7-point *item*. For each item, the observer applies rules to the pattern of yes/no indicators to determine a score, which are labeled as *inadequate* (1), *minimal* (3), *good* (5), and *excellent* (7).

However, there are substantial differences between the tools. The ECERS-3 places much more emphasis on the role of the teacher in helping children develop cognitive and social skills. The indicators at the upper end of the scoring continuum focus on staff providing higher level learning opportunities that help young children develop advanced skills related to math and literacy, which were not included within the ECERS-R. For example, on the ECERS-3 Blocks item, Indicator 20.7.3 requires that "Staff point out the math concepts that are demonstrated in a way that interests children (Ex: discuss 'more' and 'less,' relationships in size or shape: 'Look, these two squares make a rectangle, just like this one.'; number of blocks; measurement)." A classroom could attain the highest score on the ECERS-R Blocks item without any staff interaction.

The ECERS-3 also places less emphasis on materials. For example, many of the items in the ECERS-R Learning Activities subscale (e.g., fine motor, art, blocks) were focused on how materials were organized, their general condition, and whether they were accessible for children to use throughout the observation. The ECERS-3 still includes indicators related to the provision of materials in these items; however, additional indicators have been added that focus on how staff interact with children while they are using the materials in the item (e.g., indicator 17.3.2 "Staff help solve problems with sharing the materials and have children clean up properly").

The full ECERS-R contained 43 items, but many users scored only the first 37 items, excluding items that measured the quality of the provisions for parents and staff. The ECERS-3 contains 35 items and omits the parents and staff items altogether. Items that appear in both the ECERS-R and the ECERS-3 include: *Health Practices*, *Space for Gross Motor Play*, and *Staff-Child Interactions*. New 7-point items were added to ECERS-3 such as *Individualized Teaching and Learning*, *Understanding Written Numbers*, and *Becoming Familiar with Print*. In the introduction to the ECERS-3 (Harms et al., 2015), the authors note that some indicators were revised, added, or moved, based on the psychometric work done by their colleagues at Frank Porter Graham Child Development Institute as well as Gordon et al. (2013).

Finally, as noted in the introduction to the ECERS-3 (Harms et al., 2015), some definitions, rules, and procedures were revised, such as how observers determine whether materials are accessible and how long the observation lasts. Although the method of combining indicator responses into item scores is the same in the two versions, the authors now recommend that observers score all yes/no indicators, instead of just scoring those needed to obtain the item scores. Additionally, the ECERS-3 targets classrooms where most children are from 3 through 5 years of age (rather than from 2½ through 5). A final significant change to the scoring procedures is that the staff interview items have been eliminated. Instead, assessors now rely solely on observation.

1.4. Research questions

The current study seeks to answer three questions about the ECERS-3:

1. What is the factor structure of the ECERS-3 items? Are the items measuring a single, underlying construct of quality or is the ECERS-3 better described as a multidimensional tool?
2. How are the ECERS-3 Total Score and subscales created from the factor analysis related to the three domains of the CLASS Pre-K (Emotional Support, Classroom Organization, and Instructional Support)?
3. How are the ECERS-3 Total Score and subscale scores related to growth in children's early academic, executive function, and social skills?

2. Method

Data for this study were collected during the 2015–2016 academic year in three states (Georgia, Pennsylvania, and Washington) through partnerships between the research team and the agency or organization responsible for that state's QRIS data collection. Each of these agencies had already been collecting ECERS-R and CLASS Pre-K data for several years. Two of the three states had been involved in the pilot testing of the ECERS-3. The data collection effort included two components: (1) a large sample of classrooms that we refer to as the *ECERS-3 Only* sample used to answer research question 1, and (2) a smaller sample of classrooms, that we call the *In-Depth* sample, where CLASS Pre-K and child outcomes data were collected in addition to ECERS-3 to answer research questions 2 and 3. Each is discussed in more detail below.

2.1. ECERS-3 Only data collection

2.1.1. Classroom sample

Across the three states, the ECERS-3 Only sample includes 944 classrooms serving preschoolers in 743 center-based early childhood programs. Each state partner was asked to conduct ECERS-3 visits in a minimum of 260 classrooms, and each state exceeded that goal. The ECERS-3 Only sample is not representative of early childhood programs in these states; however, its large size provides sufficient statistical power to address the first research question. We elaborate further on the limitations of this sample in the Discussion.

Georgia's QRIS, called Quality Rated, transitioned to the ECERS-3 during the data collection for this study. Participation in Quality Rated is voluntary, but the state is making efforts to encourage all programs to take part. When a program applies for Quality Rated, one-third of its preschool classrooms are randomly selected for an observation, and the classrooms in this sample were already scheduled to be observed using ECERS-3 as part of their rating. On the day of the observation, teachers were asked if they were willing to have their scores included in the research study and 291 out of 295 (99%) agreed. The final ECERS-3 Only sample in Georgia included these 291 classrooms in 206 programs.

Pennsylvania's QRIS, Keystone STARS, was planning its transition to ECERS-3 during data collection. As part of that planning, all programs that were due to renew their current ERS Facility Score to support maintaining a STAR 3 or 4 rating during the study period had their current score extended for two years and received diagnostic assessments, including an ECERS-3. One preschool classroom in each program was observed. These observations were not optional, but the teachers were asked if they were willing to have their scores included in the research study and 322 out of 359 (90%) agreed. To increase the variability in quality, Pennsylvania also recruited 68 classrooms in 67 STAR 1 and 2 programs. Those visits were optional. The state did not maintain complete records on the number of STAR 1 and 2 programs contacted, but their internal notes indicate that roughly 50% agreed. The final ECERS-3 Only sample in Pennsylvania included 390 classrooms (68 STAR 1 and 2 and 322 STAR 3 or 4) in 389 programs.

Our partner in Washington primarily recruited programs that were already participating in Early Achievers, their QRIS. Washington was not yet transitioning to ECERS-3 so the sample was specifically recruited for this study. Recruitment took place during institutes and trainings and through direct appeals to current and new QRIS participants. Their sample of 148 programs resulted from contacting 487 programs (30% consent). All preschool classrooms within each program were observed, resulting in a final sample of 263 classrooms in 148 programs.

2.1.2. Measures and procedures

All ECERS-3 data were collected on tablet computers using the Branagh Information Group's (BIG) ECERS-3 program. Each state's team was trained in the use of the tablet program by BIG. Data were transmitted to BIG's secure server, where it was deidentified and transmitted to the research team. Following ECERS-3 guidelines, all yes/no indicators were scored, regardless of if they were required to attain an item score. Observations typically lasted 3 h, except in a few part-day preschools where the program itself lasted less than 3 h ($n = 21$, 2.0%), in which case the entire program was observed as recommended by the Scale's authors. All observations included at least one meal or snack and most (80%) included at least 10 min of outside/gross motor time. Nap time was generally not observed, unless it occurred during the first 3 h of the day.

Aside from the ECERS-3 itself, the only other data collected from ECERS-3 Only sample were descriptive details about the program, classroom, and teachers, such as funding (i.e., public pre-k, Head Start) and teacher race/ethnicity. BIG added additional screens to the tablet system for this project to capture this information. Teachers present on the day of the observation were asked to provide answers to these questions, which were entered directly into the tablet computers.

2.1.3. Data collector training and reliability

Each state had designated a state anchor who had been trained, prior to this study, by one of the ECERS-3 authors. Additionally,

the study had a project wide anchor, who had also been trained by one of the ECERS-3 authors and attained inter-rater reliability of over 85% within one scale point across all items on two consecutive visits. As data collection got underway, the three state anchors met with the project-wide anchor for three days to conduct reliability visits and ensure that the Scale was being administered in a standard way in all states. They made two visits as a group of four and a final visit in pairs. After each visit, they met to determine consensus scores, which is the group's final determination of the correct score for each item. The percent of items on which each anchor's scores were within one scale point of the consensus scores was considered her reliability score for that visit. Across the three days, the reliability scores for the state anchors and project-wide anchor averaged 96% (*range* = 91–100%). The same four individuals met again about half way through the data collection process to re-establish project-wide reliability over the course of four days of observations. Across the four days, reliability scores for the state anchors and project-wide anchor averaged 98% (*range* = 91–100%).

The three state anchors were responsible for ensuring that all data collectors in their states were reliably collecting data. Because these agencies were collecting as part of their state's QRIS process, each already had reliability procedures and standards in place, and the study did not modify the state rules. All three states had some procedures in common. Each had multiple anchors (state anchors and other designated individuals who were highly trained) who oversaw the training and testing of the data collectors, and determined consensus and reliability scores as described above. In all three states, a reliability score of 85% within one scale point across all items was considered the minimal acceptable reliability. The states varied slightly in how anchors were trained and the number of reliable visits each data collector needed before being allowed to collect data independently; however, all had to attain at least 85% agreement within one scale point on two consecutive visits.

To ensure that the team members remained reliable with one another across the three states, data collectors made 86 visits in pairs during data collection and agreed on consensus scores. On average, 91% (*SD* = 6) of their original scores were within one scale point of the consensus score.

2.2. In-Depth data collection

2.2.1. Classroom sample

In addition to the ECERS-3 Only data collection, each state partner was asked to recruit 40 center-based programs, and randomly select one preschool classroom in each of those programs for participation in the in-depth portion for the study. As with the ECERS-3 Only sample, the partners were responsible for identifying and recruiting their own programs for this portion of the study, with some support and guidance from the research team. States were asked to use whatever past data they had available (e.g., CLASS scores, ECERS-R scores, star ratings) to recruit programs that were likely to vary with regard to quality. As with the ECERS-3 Only sample, the In-Depth classroom sample is not representative of a particular population.

The final In-Depth sample included 119 classrooms in 119 programs. Across the three states, the overall response rate for the In-Depth sample was 64% (119 participating out of 186 contacted). As programs declined the invitation to participate, additional programs were contacted until we enrolled the target of 120 programs. The final sample includes 119 classrooms because one program withdrew after data collection began and it was too late to recruit a replacement. We do not have information about programs that declined, so those who agreed cannot be compared to those who did not.

2.2.2. Child sample

Children within each classroom in the In-Depth sample were selected for participation. Data collectors aimed to select five children at random from all eligible children in the room. Eligible children were those who (1) had parental permission, (2) were between 3 and 5 years of age, (3) spoke English at home, according to their parent,[1] and (4) were present on the day of the pretest. If there were fewer than five eligible children, all eligible children were selected. Overall, 64% of parents provided permission. In the fall of 2015, 575 children participated in the pretest. Of those, 491 (85%) participated in the posttest. Of the 84 children who did not participate in the posttest, almost all were no longer enrolled in the participating classroom ($n = 78$); the remainder were absent during multiple visits to the classroom ($n = 5$) or did not want to take part ($n = 1$).

The sample description and current analyses include all children who participated in pre- and/or posttest. We compared children who took part in both pre- and posttest test data collection to those who participated at pretest only on several key variables. They were similar on most variables, except that those who left the study were younger ($t = -2.95$, $p < .05$); had more behavior problems at pretest according to their teachers ($t = 3.41$, $p < .01$), and had lower preliteracy skills at pretest ($t = -2.62$, $p < .05$). The two groups were not significantly different on parental years of education, race, social skills, expressive language, or math skills at pretest.

On average, each of the 119 classrooms in the In-Depth portion of the study had 4.83 (range = 2–5) children with valid pretest data and 4.17 (*range* = 1–5) children with valid data at both pre- and posttest. Child demographic information was collected using a brief parent questionnaire distributed with permission forms. The form indicated that it should be completed by the adult who "takes the most care of and knows the child the best." This was typically the mother (86%) or father (8%).

2.2.3. Classroom observation measures and procedures

All classrooms in the In-Depth sample received ECERS-3 observations, which were conducted by the same data collectors and following the same procedures as described for the ECERS-3 Only sample. Program, teacher, and classroom demographic information were collected along with the ECERS-3, again following the same procedures as the ECERS-3 Only portion of the study.

In addition to the ECERS-3, a CLASS Pre-K (Pianta et al., 2008) observation was conducted in each classroom in the In-Depth sample. The CLASS Pre-K is an observational tool for measuring teacher-child interactions. It is made up of 10 dimensions, organized into three domains. The Emotional Support domain includes the dimensions of positive climate, negative climate, teacher sensitivity, and regard for student perspectives. The Classroom Organization domain includes behavior management, productivity, and instructional learning formats. The Instructional Support domain includes concept development, quality of feedback, and language modeling. Each dimension is rated from 1 to 7 with 1 or 2 indicating the classroom is *low* on that dimension; 3, 4, or 5 indicating that the classroom is in the *mid-range*; and 6 or 7 indicating the classroom is *high* on that dimension. Observers rate the classrooms and teachers on the 10 dimensions roughly every 30 min throughout the observation morning. For this project, we aimed to collect six 30-min observation cycles in each room; however, in some cases

[1] Parents were asked on the permission form to indicate all languages the child spoke at home. In general, children whose parent had indicated they did not speak English at home were excluded so that the sample would not include children whose English skills were too limited to be assessed in English. Four exceptions were made for children whose teacher indicated their English skills were strong, despite their parent indicating that they spoke only Spanish at home.

only four (n = 8; 7%) or five cycles (n = 23; 19%) were completed. At the start of each of the six CLASS Pre-K cycles, data collectors noted the number of children and staff present.

State partners hired and oversaw their own CLASS Pre-K data collection teams. All CLASS Pre-K observers were trained and certified as reliable by Teachstone, the CLASS Pre-K publishers. Additionally, prior to collecting data for this project, data collectors made visits in teams to ensure they were reliable with one another. Each data collector made at least two such visits and scored within one scale point of one another on at least 80% of the items in each of the three domains (Emotional Support, Classroom Organization, and Instructional Support). Further, two CLASS Pre-K data collectors were present for 11 of the 119 CLASS Pre-K visits and their scores were compared to ensure team members remained reliable with one another. Across all items, on average, 95% (SD = 6) of their scores were within one scale point of one another.

The CLASS Pre-K and ECERS-3 observations took place between November 2015 and April 2016 and were generally conducted on the same day (81%) by two independent observers. When they could not be scheduled on the same day, they were almost always within three days of one another (18%); in one case (1%) they were 11 days apart. CLASS Pre-K observers followed the authors' guidelines as provided in the CLASS Pre-K manual regarding when to arrive and leave, how long to stay, and which activities to observe. Thus, although the two independent ECERS-3 and CLASS Pre-K observers were typically present on the same day, they did not necessarily observe all of the same activities. For example, whereas CLASS Pre-K observers do not observe during outdoor free time, ECERS-3 observers typically do, so the two observers would not have been together during those times.

2.2.4. Child assessment measures and procedures

The goal was to conduct all pretest child assessments between the fourth and eighth week after the classroom opened for the school year. This goal was met for 112 of the 119 classrooms (94%). In the remaining classrooms, pretest assessments took place in the ninth (4%) or tenth (2%) week.

Posttest child assessments were conducted during a six-week window, starting eight weeks before the end of the school year. For year-round programs, and those that closed in July or August, the posttest window started six weeks prior to the close of the public schools in that area. If a child who was assessed in the fall was absent on the day of the spring assessment, the data collector returned on a later date to assess that child whenever possible. If a child was no longer enrolled at the time of posttest, she or he was eliminated from the study. On average, posttest data were collected 6.92 months (SD = 0.43, *range* 5.98–7.92) after pretest.

State partners hired and oversaw their own child assessment teams; however, they were trained by the research team and followed procedures designed by the research team to select and assess children. Prior to fall data collection, a member of the research team traveled to each state and spent two days training the state-level teams on the child assessment battery. Following that training, assessors were asked to spend between 40 and 60 h practicing the battery alone, with adults, or with children. At the end of the practice, the same research team member returned to each state and watched individual assessors complete the full battery with a child and signed-off on their correct administration. When problems were identified, the individual was given additional instruction and support, followed by additional practice and testing. Once data collection began, assessors submitted their first two assessments to the research team via overnight mail so they could be reviewed immediately to provide feedback to the assessor. The child assessment battery included the following measures.

2.2.4.1. Woodcock–Johnson IV (WJ IV, selected subtests; Schrank, McGrew, & Mather, 2014). The WJ IV is a set of individually administered, nationally normed tests for measuring general intellectual ability, specific cognitive abilities, oral language, and academic achievement. Three subtests were administered: Picture Vocabulary, Letter-Word Identification, and Applied Problems. The Picture Vocabulary subtest measures children's expressive language by asking them to name a series of increasingly complex images such as cat, zipper, and doorknob. The Letter-Word Identification subtest assesses children's preliteracy skills by asking children to identify first single letters, then progressively more complex words. On the Applied Problems subtest, children are asked to complete math-related tasks such as showing two hands, counting objects, and adding or subtracting small numbers. The WJ IV technical manual (McGrew, LaForte, & Schrank, 2014) reports reliabilities for 4-year olds of .94 for Picture Vocabulary, .97 for Letter-Word Identification, and .93 for Applied Problems. For concurrent validity, they report a correlation with the Differential Abilities Scale-II (DAS II) Verbal Ability of .78 for Picture Vocabulary and .71 for Letter-Word Identification. No concurrent validity statistics were reported for the Applied Problems subscale.

2.2.4.2. Head-Toes-Knees-Shoulders (HTKS; Ponitz, 2008). HTKS is an executive function assessment that is administered one-on-one with a young child. During the assessment, children are asked to play a game in which they must do the opposite of what the experimenter says. For example, the experimenter instructs children to touch their head, but instead of following the command, the children are supposed to do the opposite and touch their toes. If children pass the head/toes part of the task, they complete an advanced trial where the knees and shoulders commands are added. The HTKS task has been conceptualized by its authors as a measure of inhibitory control (children must inhibit the dominant response of imitating the examiner), working memory (children must remember the rules of the task), and attentional focusing (children must focus attention to the directions being presented by the examiner). According to the authors, HTKS demonstrated positive correlations with parent ratings of attentional focusing and inhibitory control and teacher ratings of behavior regulation in a sample of kindergartners. Fall HTKS scores predicted higher achievement and self-regulation as rated by the teacher in the spring (Ponitz, McClelland, Matthews, & Morrison, 2009). The developers report Cronbach's alphas between .92 and .94. (McClelland et al., 2014). Using guidance from one of the HTKS authors (M. McClelland, personal communication, May 20, 2016), we included all practice items (n = 17) and test items (n = 30) to create a total score. Item scoring ranges from 0 to 2, and the possible range for the total score is 0−94.

2.2.4.3. Devereux Early Childhood Assessment Preschool Program, 2nd Edition (DECA-P2, LeBuffe & Naglieri, 2012). This teacher report of children's social skills asks teachers to respond to a series of 38 statements regarding the child's behavior over the past four weeks, using a 5-point scale that ranges from *never* to *very frequently*. It provides two scores: (1) Total Protective Factors (27 items), measuring the child's initiative, self-regulation, and ability to maintain positive connections with others, and (2) Behavioral Concerns (11 items), measuring the extent to which the child displays behavioral challenges that might require referral or intervention. It is nationally standardized and the authors report test–retest reliability of .95 for the Total Protective Factors and .80 for Behavioral Concerns. In the current sample, the internal reliability (Cronbach's alpha) was .95 for Total Protective Factors and .87 for Behavior Concerns. It was scored using the author guidance to create nationally standardized T-Scores.

2.3. Analysis plan

The results section starts with descriptive information regarding the 1063 classrooms in which ECERS-3 observations were conducted (ECERS-3 Only and In-Depth samples combined), the 119 classrooms in the In-Depth sample, and the 575 children in the child outcomes portion of the study.

Three main sets of analyses follow the descriptive information: (1) tests of the structure of the scale, including a single ECERS-3 Total Score, the six subscales suggested by the authors, and potential new factors for the ECERS-3 items derived from confirmatory and exploratory factor analyses; (2) analysis of concurrent and divergent validity in which ECERS-3 scores are compared to CLASS Pre-K scores; and (3) analysis of predictive validity, in which hierarchical linear modeling (HLM) is used to test ECERS-3 scores as predictors of children's growth in early academic, executive function, and social–emotional skills. Parallel HLMs are presented in which CLASS Pre-K scores are used.

3. Results

3.1. Sample description

Table 1 presents and compares descriptive information about the classrooms and teachers in the ECERS-3 Only sample, the In-Depth Sample, and the two groups combined (all classrooms).

Table 1
Classroom and teacher sample.

	ECERS-3 Only (n = 944)	In-Depth (n = 119)	All classrooms (n = 1063)
Number of programs	743	119	862
Unannounced visit**	85.64%	32.77%	79.66%
Mean (SD) children present	14.12 (4.50)	14.75 (4.20)	14.19 (4.47)
Mean children (SD) per adult (ratio)	6.95 (2.35)	7.41 (2.54)	7.00 (2.38)
Age of most children at start of school year			
Mostly 3-year-olds**	34.30%	18.49%	32.51%
Mostly 4-year-olds**	55.81%	70.59%	57.49%
Equal number of 3- and 4-year-olds	9.89%	10.92%	10.00%
Classroom auspice			
Head Start	14.96%	8.55%	14.23%
State-funded pre-K*	37.75%	47.86%	38.84%
In a public school	18.41%	11.02%	17.57%
Lead teacher race/ethnicity			
Asian	2.44%	1.68%	2.35%
Black/African American	18.11%	11.76%	17.40%
Hispanic/Latino	4.45%	3.36%	4.33%
White	68.22%	77.31%	69.24%
Mixed race/Other	3.17%	3.36%	3.20%
Missing/refused	3.60%	2.52%	3.48%
Lead teacher education			
High school diploma or less	10.17%	13.45%	10.53%
Some college	11.33%	5.88%	10.72%
Associate's	15.47%	21.85%	16.18%
Bachelor's	42.37%	42.86%	42.43%
Graduate work or degree	14.30%	14.29%	14.30%
Missing/refused	6.36%	1.68%	5.83%

Notes: The ECERS-3 Only and In-Depth samples were compared on all variables using Chi Square and significant differences are noted. Classrooms can belong to multiple classroom auspice categories. Classrooms with blended funding were counted as Head Start and/or state-funded pre-K if any enrolled children were funded with those sources. *In a public school* refers to a physical location in public school building where older children were also attending.

* $p < .05$.

** $p < .01$.

Table 2
Child sample (n = 575).

Variable	Value
Mean (SD) child age in months at pretest	51.99 (6.36)
Mean (SD) months between pretest and posttest	6.92 (0.43)
Gender	
Girl	50.78%
Boy	49.22%
Mean (SD) parent education (in years)	14.25 (2.33)
Family poverty	
Less than 100%	30.87%
Over 100% but less than 185%	22.33%
185% or higher	46.80%
Race/ethnicity	
Asian	1.94%
Black/African American	16.55%
Latino/Hispanic	7.39%
Native American	1.06%
White/Caucasian	56.51%
Multiple	16.55%
Language(s) child typically speaks at home	
English only	90.16%
Spanish only	0.70%
English and Spanish	5.27%
English and another language	3.87%
Individualized Education Plan (IEP)	5.49%

Notes: All information on this table (aside from months between pre- and posttest) was gathered from the child's parent, during the permission process. Poverty was calculated by comparing income and family size to federal poverty guidelines. Parents were asked for their highest educational attainment, and that value was converted to years of education as follows *8th grade or less* = 8; *some high school but no diploma* = 11; *high school diploma or equivalent* = 12; *some college* = 13; *technical training or certificate* = 13; *Associate's* = 14; *Bachelor's degree* = 16; *graduate degree* = 18.

Research Question 1, regarding the factor structure of the ECERS-3, was addressed using all classrooms. Research Questions 2 and 3, regarding the associations among ECERS-3, CLASS Pre-K, and children's outcomes, used the In-Depth sample. The average class size was small, over half the rooms were primarily for four-year olds, and about one-third were state-funded pre-K. Compared to the ECERS-3 Only sample, the In-Depth sample was less likely to have unannounced visits, less likely to serve mostly 3-year olds, more likely to serve mostly 4-year olds, and more likely to be part of a State-Funded Pre-K.

The demographic characteristics as reported by the child's primary caregiver during the permission process are presented in Table 2. On average, children in the sample were 52 months old at pretest and their primary parent (typically the mother) had over 14 years of education. Almost one-third of children were from families with incomes at or below the poverty line for their family's size and almost half were from families over 185% of the poverty line. Over half were White, with African American and children of multiple races making up the other larger groups. Almost all spoke only English at home.

3.2. Descriptive statistics for ECERS-3 items

Table 3 presents descriptive statistics for the 35 ECERS-3 items. All items were completed for all 1063 observations except Items 27 (*Appropriate use of technology; n* = 291) and 35 (*Whole-group activities for play and learning, n* = 1044). Those are the only two items on which Not Applicable (NA) is permitted. The range on all items was 1–7, except Item 20 (*Blocks*), where the maximum score attained was 6. Cronbach's alpha for the 35 items was .93, suggesting very high internal consistency. This is the same level of internal consistency reported by the scale authors in the ECERS-3 manual (Harms et al., 2015, p. 4). Elimination of any single item had no effect on alpha.

Table 3
ECERS-3 item-level descriptive statistics and standardized factor loadings from four factor solution of the exploratory factor analyses ($n = 1063$).

Items	Mean	SD	F1	F2	F3	F4
17. Fine Motor	3.98	1.59	**.83**	−.10	.01	−.04
18. Art	3.43	1.48	**.73**	−.04	−.01	.12
26. Promoting acceptance of diversity	4.07	1.19	**.69**	−.04	−.04	−.14
15. Encouraging children's use of books	3.69	1.47	**.67**	−.08	.08	−.04
21. Dramatic play	3.14	1.66	**.61**	.01	−.03	.14
22. Nature/science	2.54	1.17	**.53**	.12	.01	.22
34. Free play	4.06	1.51	**.53**	.09	.29	−.01
29. Individualized teaching and learning	4.32	1.70	**.49**	.00	.34	.13
4. Space for privacy	4.07	1.60	**.48**	.02	.07	.06
20. Blocks	2.23	1.26	**.46**	.05	−.06	.20
19. Music and movement	3.15	1.17	**.45**	.05	.14	.01
6. Space for gross motor play	3.18	1.42	.00	**.79**	−.08	.00
7. Gross motor equipment	2.80	1.68	.01	**.68**	.05	.05
28. Supervision of gross motor	4.11	1.74	−.05	**.45**	.39	−.04
32. Discipline	4.18	1.52	−.04	−.07	**.85**	.05
30. Staff-child interaction	4.97	1.84	.00	−.02	**.84**	−.09
31. Peer interaction	4.47	1.56	.03	.00	**.75**	.04
35. Whole-group activities for play and learning	3.80	1.50	−.08	.00	**.72**	.08
13. Encouraging children to use language	4.20	1.54	.05	−.04	**.70**	.14
33. Transitions and waiting times	3.90	1.92	−.02	.03	**.66**	.05
9. Toileting/diapering	3.21	1.41	.00	.04	**.52**	−.18
14. Staff use books with children	3.38	1.69	.08	−.03	**.49**	.02
10. Health practices	3.06	1.40	.11	.14	**.43**	−.18
12. Helping children expand vocabulary	3.65	1.42	.03	.01	**.43**	.36
24. Math in daily events	2.99	1.43	.10	.04	.24	**.49**
23. Math materials/activities	2.29	1.34	.29	.00	.08	**.48**
25. Understanding written numbers	1.73	1.15	.18	.03	.02	**.47**
1. Indoor space	4.55	1.55	.08	.04	.35	−.19
2. Furnishings for care, play and learning	4.05	1.10	.20	.03	.28	−.15
3. Room arrangement for play and learning	3.42	1.45	.27	.07	.21	.06
5. Child-related display	3.24	1.37	.12	−.04	.24	.18
8. Meals/snacks	3.15	1.29	.17	.16	.14	.01
11. Safety practices	4.03	1.72	.08	.14	.37	−.32
16. Becoming familiar with print	3.19	1.24	.14	−.02	.28	.29
27. Appropriate use of technology	3.14	1.86	–	–	–	–
Mean scores	3.53	0.80	3.52	3.36	3.90	2.34
Standard deviations			1.01	1.27	1.07	1.04

Notes: $n = 1063$ for except items 27 (*Appropriate use of technology;* $n = 291$) and 35 (*Whole-group activities for play and learning,* $n = 1044$). We labeled the factors: Learning Opportunities (F1), Gross Motor (F2), Teacher Interactions (F3), and Math Supports (F4). The subscale scores used in subsequent analyses are the unit weighted means of items that load .40 or higher (bold). Appropriate Use of Technology was excluded from the EFA because it was scored Not Applicable in most observations.

3.3. Confirmatory factor analysis

The ECERS-3 is typically scored by calculating a simple mean of the 35 items under the implicit assumption that it is a unidimensional instrument. To determine if it is actually uni- or multidimensional, we used a confirmatory factor analysis (CFA) parameterized with all items loading onto a single factor, and with factor variance fixed to 1 and factor mean fixed to 0.[2] Assessment of model fit used standard criteria: RMSEA < .05 for good fit and <.10 for weak fit; CFI > .9 for good fit (e.g., Hu & Bentler, 1999; Sivo, Fan, Witta, & Willse, 2006). The CFA indicated that model fit was weak (*RMSEA* = .081, *CFI* = .727, *Chi-Square* [560] = 4429.08, $p = .0000$).

Next, we fixed all factor loadings to 1, which is equivalent to taking the simple mean of items rather than using the factor loadings to score the instrument. Fixing the factor loading at 1 further reduced model fit (*RMSEA* = .093, *CFI* = .612, *Chi-Square* [594] = 6106.28, $p = .0000$) and the reduction in fit was significant (*Chi-Square* [34] = 1677.20, $p = .0000$), suggesting that if a one-factor model is used, a mean score is even less reliable than a factor score. Finally, we evaluated the six subscales presented by the authors in the tool itself: Space and Furnishings, Personal Care Routines, Language and Literacy, Learning Activities, Interaction, and Program Structure. The CFA indicated that this model fit was also weak (*RMSEA* = .104, *CFI* = .548, *Chi-Square* [561] = 6979.28, $p = .0000$). Note that we did not attempt to replicate the 2-factor structure that some researchers have found using the ECERS-R (e.g., Cassidy et al., 2005) because the items and indicators are quite different in the two versions of the tools. Attempting to force ECERS-3 items into the ECERS-R factors would have required omitting most of the items that are new to ECERS-3 or making assumptions about their loadings that went well beyond the original factor analysis work.

3.4. Exploratory factor analysis

The CFA results indicated that a single factor cannot adequately capture sources of variance among the items. Exploratory factor models were estimated using a maximum likelihood extraction with an oblique rotation. We chose to exclude the Appropriate Use of Technology item given its very low response rate. The scree plot suggested a 2- or 4-factor solution. We ultimately selected the 4-factor solution and our rationale is detailed below.

[2] Concerns regarding whether Likert response items should be treated as categorical or continuous have been raised (see Carifio & Perla, 2007 for a summary and exploration). We replicated our CFA models using categorical methods (i.e., estimation of asymptotic covariance matrices as part of the analysis in MPLUS); results replicated nearly exactly with no consequences for the interpretation of the results. We choose to retain the models that treated scores as continuous.

Analysis of the scree plot included parallel analysis (Horn, 1965). Parallel analysis is the comparison of the scree plot from the observed data to a scree plot of data that were randomly generated with the same number of items and subjects, but with no latent structure. The point at which the plots cross indicates the optimal number of factors. In our data, the plots crossed between four and five factors suggesting that the inclusion of up to four factors would provide a meaningful structure, but five or more factors are just a function of random noise. Further, the 4-factor solution satisfies Thurstone's (1954) simple structure criterion, whereas the 2-factor does not. Simple structure requires that a given item is unambiguously related to a specific factor and that each factor is composed of a relatively unique set of items. In our data, it was clear that the 2-factor solution did not meet these criteria; there were six items with non-zero loadings for both factors.

Next, we examined statistical measures of model fit. The 4-factor solution was the only one to achieve traditional criteria (4-factor: *RMSEA* = .046, *CFI* = .927, $X^2(461) = 1486.21$, $p = .0000$; versus 2-factor: *RMSEA* = .064, *CFI* = .848, $X^2(494) = 2612.20$, $p = .0000$), and review of the items indicated that the 4-factor solution was more easily interpreted than the 2-factor solution. It is worth noting that the 2-factor solution suggested by this EFA was quite different from the Activities/Materials and Language/Interactions factors identified by Cassidy and colleagues (2005) using the ECERS-R. Although the loadings on this first ECERS-3factor were related to Activities, which is consistent with Cassidy and colleagues, the items on the second factor included most, but not all, of the health, safety, language, interaction, and supervision items. This second factor was particularly difficult to name because it combined a wide range of items.

Our cutoff for selecting an item as meaningful for interpreting a given factor was a standardized factor loading of .40. This is a strict criterion that does result in the loss of several items for each factor, but also aids in providing uniquely identified and interpretable factors. The factor loadings are presented in Table 3. The loadings themselves can be read as correlations between the item and the factor. Each factor should be understood as something shared by all of the items, and the loadings are a measure of the strength of those relationships.

We decided to further explore the factor structure of the scale with the 4-factor solution and named the factors Learning Opportunities (e.g., fine motor, art, blocks), Gross Motor (e.g., space for gross motor, gross motor equipment), Teacher Interactions (e.g., staff–child interactions, discipline), and Math Supports (e.g., math in daily events, math materials/activities). For the remainder of the analyses, we use Total Score (i.e., traditional mean score of all items) and the simple means of the items that load at .40 or above on each of the four factors as predictors. We continue to analyze the Total Score despite its weak psychometric properties because it is the typical way the tool is used and, therefore, understanding how it relates to CLASS Pre-K and children's development is important. We present the average of the items that load on the four factors because they are more easily understood by a broad audience than the factor scores, do not vary depending on the sample, and could be easily employed by other users.

Before electing to use the mean scored subscales, we estimated two models based on the subset of items identified in the EFA as loading on the four factors. The first of these models was a factor scored model where all loadings were left unconstrained (RMSEA = .059, CFI = .910, $X^2(293) = 1372.32$). The second was a mean scored model where all of the factor loadings were constrained to 1 (RMSEA = .077, CFI = .832, $X^2(315) = 2325.15$). Although both models fit moderately well, using the mean scoring does come at a cost. Model fit for the mean score model of the four factors is significantly worse than a factor scored model (X^2 difference [22] = 952.83, $p < .001$). However, the scores produced by the two scoring methods are almost perfectly correlated ($r = .99$ between the mean score and factor score for the first three factors and $r = .95$ for the fourth factor), and the relative fit indices suggest very little difference between the scoring methods (*AIC* = 89,412.99, *BIC* = 89,831.38 and *AIC* = 90,322.82, *BIC* = 90,630.89 for factor and mean scored, respectively). So, this decision is unlikely to affect our results and we present robustness checks below.

Table 4
Correlations among ECERS-3 Total Score and subscales ($n = 1063$).

	A	B	C	D	E
A. ECERS-3 Total Score	1.00				
B. ECERS-3 Learning Opportunities	0.86	1.00			
C. ECERS-3 Gross Motor	0.43	0.22	1.00		
D. ECERS-3 Teacher Interactions	0.89	0.59	0.36	1.00	
E. ECERS-3 Math Supports	0.65	0.59	0.15	0.51	1.00

Notes: $p < .001$ for all values.

Table 5
ECERS-3 and CLASS Pre-K descriptive statistics for classrooms in the In-Depth portion of the study.

	n	Mean	SD	Min	Max
ECERS-3 Total Score	119	3.40	0.76	1.66	5.15
ECERS-3 Learning Opportunities	119	3.51	0.95	1.18	5.36
ECERS-3 Gross Motor	119	3.07	1.31	1.00	6.00
ECERS-3 Teacher Interactions	119	3.64	1.02	1.02	5.73
ECERS-3 Math Supports	119	2.32	1.04	1.00	6.67
CLASS Pre-K Emotional Support	118	5.68	0.79	3.63	7.00
CLASS Pre-K Classroom Organization	118	5.28	1.05	2.75	6.92
CLASS Pre-K Instructional Support	118	2.26	0.69	1.00	5.06

See Table 4 for correlations among the ECERS-3 Total Score and derived subscales. The Total Score is strongly correlated with Learning Opportunities and Teacher Interactions factors. The Gross Motor factor is the least strongly associated with the other subscales or the Total Score. Note that seven items did not strongly load on any of these factors. They are listed at the bottom of Table 3.

3.5. Descriptive information about in-depth classrooms and child assessments

The remainder of the analyses are limited to the classrooms in the In-Depth sample (i.e., those with CLASS Pre-K and child assessments in addition to ECERS-3). Table 5 presents the descriptive statistics for these classrooms and Table 6 presents fall and spring descriptive statistics for the child assessments. CFA models using just the In-Depth sample indicated that the factor loadings for the one, two, and four factor solutions were very similar (within the standard error) to the full sample of 1063 classrooms used in the analyses presented above.

For the Woodcock Johnson, we present W scores, which are linked to normative data and allow for predicting an individual's proficiency at any level of task difficulty (Jaffe, 2009). Specifically, W scores are on an equal-interval scale and are well-suited for measuring growth. Unlike standard scores which remain constant with normative growth, increases in W score reflect actual growth on the indicator measured. Thus, we anticipate that W scores will increase as the child matures.

3.6. Associations between ECERS-3 and CLASS Pre-K

To assess the concurrent and divergent validity of the ECERS-3, we calculated Spearman correlations between the various ECERS-3 scores and the three domains of the CLASS Pre-K, after limiting the sample to the 95 cases (80% of the total) where the two observations were made on the same day. Spearman correlations were used because both ECERS-3 and CLASS are on ordinal, rather than

Table 6
Descriptive statistics for child level data.

	Fall			Spring		
	n	Mean	SD	*n*	Mean	SD
DECA Total Protective Factors T-score	533	50.63	9.64	454	53.61	9.71
DECA Behavioral Concerns T-score	533	48.48	10.49	454	48.58	10.51
HTKS Total Score	572	17.47	24.48	487	33.62	32.13
WJ IV Picture Vocabulary W Score	575	455.86	13.98	491	462.65	12.31
WJ IV Letter-Word W Score	575	328.09	25.49	491	345.57	28.02
WJ IV Applied Problems W Score	575	401.10	23.54	490	415.44	19.22
WJ IV Picture Vocabulary Std Score	573[a]	100.55	13.11	489[a]	100.16	11.91
WJ IV Letter-Word Std Score	575	92.30	12.87	491	92.17	12.59
WJ IV Applied Problems Std Score	573[a]	93.92	15.41	489[a]	96.13	13.87

[a] Two children scored below the floor for standard scores.

Table 7
Spearman correlations among ECERS-3 and CLASS Pre-K scores when the two were on the same day ($n = 95$).

		CLASS Pre-K		
		Emotional Support	Classroom Organization	Instructional Support
ECERS-3	Total Score	0.44***	0.44***	0.33***
	Learning Opportunities	0.36***	0.30**	0.22*
	Gross Motor	0.33***	0.26*	0.12
	Teacher Interactions	0.40***	0.34***	0.39***
	Math Supports	0.31**	0.28**	0.24*

* $p < .05$.
** $p < .01$.
*** $p < .001$.

ratio, scales. As seen in Table 7, the three domains of the CLASS Pre-K are significantly, but modestly, associated with the ECERS-3 Total Score and most of the ECERS-3 subscales. The association between ECERS-3 Gross Motor and CLASS Pre-K Instructional Support is non-significant.

3.7. HLM predicting child outcomes from ECERS-3

For most outcomes, we tested associations between ECERS-3 and children's early social and academic outcomes using longitudinal three-level HLMs with time (level 1) nested in child (level 2) nested in classroom (level 3). The models included random intercepts at both level 2 and level 3; both intercepts were significant in all models ($p < .05$). Fixed effects included time (pretest or posttest), ECERS-3 scores, and the interaction of the two. The interaction is the key test. It indicates the amount of change over time that is associated with ECERS-3 scores. Information collected from children who only participated at pretest contributes to estimates of the intercepts. Data from children with pre- and posttest scores contribute to estimates of the intercepts and slopes. Results from these HLMs appear in Table 8. In the interest of efficiency, we present only the parameters for the interaction of time and ECERS-3 scores. The parameter estimates were standardized so that they represent the amount of growth on the dependent variable, in standard deviations, associated with a one standard deviation change ECERS-3.

Review of the regression diagnostics (residuals plots, *Q*–*Q* plots) suggested that model assumptions were met for all child outcome variables with the exception of the HTKS, on which the Fall scores were highly skewed; of a possible 94, one-fifth (22%) of the children scored a zero and just over half (51%) scored a four or less. Spring scores were more normal. For the HTKS outcome, we ran the models as pretest controlled regressions; the Fall scores are included as covariates of quality and Spring scores alone are the outcomes. The nesting of children within classroom still required the estimation of a two-level HLM with children at level one and classroom at level two. Regression diagnostics for these models indicated that the residuals were reasonably normally distributed, although with a moderate positive skew (all models had excess kurtosis between .00 and .07, with skew between .81 and .86).

Looking first at the ECERS-3 Total Score, which is the typical way in which ECERS-3 scores are calculated, these findings indicate that it was only significantly associated with growth in children's executive function skills (as measured by HTKS, see Table 8). There was also a marginal association between Total Score and growth in preliteracy (as measured by WJ IV Letter-Word). ECERS-3 Total Score was not associated with growth in social skills, expressive vocabulary, or math skills.

The Learning Opportunities subscale was significantly associated with growth in executive function and math skills (as measured by WJ IV Applied Problems), and marginally associated with growth in preliteracy. The associations between the Gross Motor subscale and outcomes were not significant, although there was a marginal association with preliteracy. The Teacher Interactions subscale was related to executive function, but not to social–emotional or early academic skills. The Math Supports subscale was significantly related to protective social–emotional skills as measured by the DECA and marginally related to preliteracy. Surprisingly, Math Supports was not related to children's math skills, as measured by the WJ IV Applied Problems. All significant associations were small, with effect sizes ranging from .06 to .08, indicating that a one-standard deviation change in the ECERS-3 was associated with less than .10 standard deviation growth in children's outcomes.

3.8. Robustness checks

To ensure that these findings were robust with regard to analytic decisions, two types of robustness checks were conducted. First, we re-ran the same models outlined above, including numerous covariates to account for ways in which children and classrooms might vary systematically, including state (dummy coded as two dichotomous variables), Head Start or not, family income relative to federal poverty levels for family size (three levels, represented as two dichotomous variables), primary parent's years of education, child gender, child age at pretest, and time between fall and spring assessments in months. All covariates and ECERS-3 scores were mean centered prior to analysis. Missing data were managed through missing data replacement (Allison, 2001) using the EM algorithm through SAS *Proc MI* (SAS Institute, 2002–2012). Following accepted practice (see, for example, Schafer & Graham, 2002), 20 data sets were imputed. Following statistical analysis, results of model effects and posthoc comparison were compiled using SAS *Proc MIAnalyze* (SAS Institute, 2002–2012). When these covariates were added to the models, the pattern of significance was identical to the models presented in Table 8. Further, the parameter estimates were almost identical in magnitude. This led us to conclude

Table 8
Standardized parameter estimates (standard errors) for interaction of time by ECERS-3 scores as predictors of social–emotional and academic skills.

Dependent variable	ECERS-3				
	Total Score	Learning Opportunities	Gross Motor	Teacher Interactions	Math Activities
DECA Total Protective T-score (n = 533)	0.01 (0.04)	0.02 (0.04)	0.02 (0.04)	0.00 (0.04)	**0.08 (0.04)***
DECA Behavioral Concerns T-score (n = 533)	−0.03 (0.03)	−0.03 (0.04)	−0.05 (0.04)	−0.03 (0.03)	−0.02 (0.03)
HTKS Total Score (n = 572)	**0.06 (0.03)***	**0.08 (0.03)****	−0.03 (0.04)	**0.06 (0.03)***	0.03 (0.04)
WJ IV Picture Vocabulary W Score (n = 575)	0.00 (0.03)	−0.02 (0.03)	0.00 (0.03)	0.01 (0.03)	0.00 (0.03)
WJ IV Letter-Word W Score (n = 575)	0.05 (0.03)†	0.05 (0.03)†	0.05 (0.03)†	0.01 (0.03)	0.04 (0.02)†
WJ IV Applied Problems W Score (n = 575)	0.03 (0.03)	**0.08 (0.03)***	−0.01 (0.03)	0.01 (0.03)	−0.02 (0.03)

Notes: Significant associations appear in bold. For all outcomes other than HTKS, each cell represents a separate 3-level HLM in which time (pre- vs. posttest) is nested within child, which is nested within classroom, and the parameter estimates presented are for the interaction of time by ECERS-3. For HTKS, each cell is a 2-level HLM, in which child is nested in classroom, and pretest score is controlled. The parameter estimates presented for HTKS are for the effect of ECERS-3. For all models, the parameter estimates have been standardized so that they represent the amount of growth on the dependent variable, in standard deviations, associated with a one standard deviation change on the CLASS Pre-K.
* $p < .05$.
** $p < .01$.
† $p < .10$.

that our primary findings were robust to classroom and child level differences.

As a second type of robustness check, the four factor scores were used in place of the four mean scores and, again, the pattern of findings was largely similar to those presented in Table 8. The only difference was that the Factor 4 Math Supports score was not related to total protective social emotional skills and was marginally ($p < .10$) positively related to executive function. These slight differences led us to conclude that using the means of items, rather than factor scores, did not substantially undermine the subscales' utility.

3.9. HLM predicting child outcomes from CLASS Pre-K

Finally, to understand how the ECERS-3 associations with children's social and academic outcomes were similar or different from another widely used tool, we estimated the same models as presented in Table 8 using the three CLASS Pre-K domains as predictors. As with the ECERS-3, the parameter estimates were standardized so that they represent the amount of growth on the dependent variable, in standard deviations, associated with a one standard deviation change on the CLASS Pre-K.

As seen in Table 9, all three CLASS Pre-K domains were significantly associated with growth in executive function, and Classroom Organization was significantly associated with growth on the DECA-P2 Total Protective Factors. The magnitude of the significant associations was small and similar to those observed for the ECERS-3. Each standard deviation increase on the CLASS was associated with a .08–.10 standard deviation growth in outcome. None of the CLASS Pre-K domains was associated with growth in children's early academic skills. When these CLASS Pre-K models were re-run with the covariates described earlier, the direction and significance of all findings were unchanged and the parameter estimates were similar.

4. Discussion

This study represents the first attempt to evaluate the factor structure and validity of the newly published ECERS-3. Understanding its strengths and weaknesses is critically important as state QRIS, professional development providers, and early childhood researchers start to adopt this tool.

4.1. Summary of findings

The results suggested that a single factor does not adequately capture item variability. Expanding the solution to allow for two or four factors resulted in adequate, although still weak, model fit. Of the solutions tested, the four-factor provided the greatest balance of statistical support and theoretical utility. The ECERS-3 Total Score, along with the three of the four subscales were significantly, moderately correlated with all three domains of the CLASS Pre-K, providing evidence of the ECERS-3′s concurrent validity.

The ECERS-3 Total Score was associated with growth in executive function skills. The Learning Opportunities subscale was associated with growth in executive function and math skills. The Teacher Interactions subscale was related to growth in executive function, and the Math Activities subscale was associated with

Table 9
Standardized parameter estimates (standard errors) CLASS Pre-K domain scores as predictors of social–emotional and academic skills.

Dependent variable	CLASS Pre-K		
	Emotional Support	Classroom Organization	Instructional Support
DECA Total Protective T-score (n = 528)	0.03 (0.04)	**0.09 (0.04)***	0.02 (0.04)
DECA Behavioral Concerns T-score (n = 528)	−0.05 (0.04)	−0.03 (0.04)	−0.01 (0.04)
HTKS Total Score (n = 567)	**0.10 (0.03)****	**0.09 (0.03)****	**0.08 (0.04)***
WJ IV Picture Vocabulary W Score (n = 570)	0.00 (0.03)	0.03 (0.03)	0.02 (0.03)
WJ IV Letter-Word W Score (n = 570)	−0.01 (0.03)	0.01 (0.03)	−0.01 (0.03)
WJ IV Applied Problems W Score (n = 570)	0.01 (0.03)	0.00 (0.03)	−0.02 (0.03)

Notes: Signficant associations appear in bold. For all outcomes other than HTKS, each cell represents a separate 3-level HLM in which time (pre- vs. posttest) is nested within child which is nested within classroom, and the parameter estimates presented are for the interaction of time by CLASS Pre-K domain score. For HTKS, each cell is a 2-level HLM, in which child is nested in classroom, and pretest score is controlled. The parameter estimates presented for HTKS are for the effect of CLASS Pre-K. For all models, the parameter estimates have been standardized so that they represent the amount of growth on the dependent variable, in standard deviations, associated with a one standard deviation change on the CLASS Pre-K.
† $p < .10$.
* $p < .05$.
** $p < .01$.

growth in social skills. These associations with children's outcomes provide some evidence of the tool's predictive validity, but the associations are small and not domain-specific. Each of the three CLASS Pre-K domains was significantly associated with growth in executive function. Additionally, the Classroom Organization domain of the CLASS Pre-K was related to growth in social skills. As with the ECERS-3, all associations between CLASS Pre-K and children's growth were small.

4.2. Four subscales

Although Cronbach's alpha for the 35 items of the ECERS-3 was acceptable, indicating that the items were positively correlated with one another, subsequent analysis suggested that multiple correlated factors were preferable, particularly the four-factor solution. The Learning Opportunities subscale included items such as fine motor, art, and dramatic play and described the extent to which children have access to a variety of materials and experiences during open-ended activities for an extended period. The Gross Motor subscale included only the three gross motor items (space for gross motor, gross motor equipment, and gross motor supervision). The Teacher Interactions subscale, which included items such as discipline and staff-child interactions, as well as the language items, such as encouraging children to use language, described the extent to which teachers are actively engaged with children and encouraging learning. The fourth subscale, Math Supports, included the three items specifically related to math instruction and activities.

The Learning Opportunities and Teacher Interactions subscales are similar to subscales found in previous work on the ECERS-R (Cassidy et al., 2005; Sakai, Whitebook, Wishard, & Howes, 2003). In both the ECERS-R and the ECERS-3, these subscales differentiate between the types of materials and experiences offered from the ways that teachers organize the day and provide support for learning. The Math Activities and Gross Motor subscales are new. The ECERS-R included only one math item, so identifying a Math factor from the ECERS-R would not have been possible. The ECERS-3 tool includes three math items, each requiring a high level of intentionality on the part of the teacher. The fact that they form their own factor indicates that math instruction is addressed differently from other types of learning activities, such as nature/science and fine motor. Likewise, the indicators within the Gross Motor items are more specific in the ECERS-3 than they were in the ECERS-R, providing more exact definitions of time, hazards, and adult roles. This increased specificity appears to have enabled the estimation of these two narrower constructs.

It is noteworthy that seven of the 34 ECERS-3 items that were included in the factor analyses (recall that Appropriate Use of Technology was excluded due to low response) do not appear on any of the four subscales. Of these seven, four measure the physical space (indoor space; furnishing for care, play, and learning; room arrangement for play and learning; child related display), two address personal care (meals/snacks; safety practices), and one is becoming familiar with print. The fact that none of these loaded at .40 or higher indicates that these items are not associated with one another strongly enough to create their own factor, nor are they associated with any of the four factors described in this paper strongly enough to load onto them. This does not necessarily mean that they are unimportant to the measurement of quality broadly. In fact, six of the seven did load at .40 or higher in the single factor solution, which is also the Total Score, representing the most common application of the tool. This indicates that although these items do not fit our more granular approach to measuring quality, they may still be associated with quality as measured by ECERS-3 in general. Deeper exploration into how these items relate to one another and contribute to how children develop is needed. Future research should explore more complex models such as those with a higher order general quality factor.

4.3. Concurrent and divergent validity

Correlations among the ECERS-3 Total Score, the four derived ECERS-3 subscales, and the three domains of the CLASS Pre-K provide evidence for concurrent and divergent validity of the ECERS-3. The ECERS-3 subscale most strongly associated with CLASS Pre-K was Teacher Interactions, which provides evidence for convergent validity since CLASS Pre-K is intended as a measure of teacher–child interactions (Pianta et al., 2008). The lower and non-significant correlations between the ECERS-3 Gross Motor subscale and the CLASS Pre-K domains provide evidence for divergent validity. CLASS Pre-K is not designed to measure gross motor activities, and CLASS Pre-K observers do not code during outdoor free play. These results indicate that the Gross Motor subscale is capturing an aspect of quality that is unique to ECERS-3.

Correlations between ECERS-3 Total Score and CLASS Pre-K are slightly smaller than those seen in previous research using ECERS-R and CLASS Pre-K. For instance, Denny, Hallam, and Homer (2012) reported correlations between ECERS-R Total Score and the three subscales of the CLASS Pre-K ranging from .58 to .61 in a sample of 114 child care classrooms in Tennessee. Using an older version of the CLASS, La Paro, Pianta, and Stuhlman (2004) reported that the correlation between ECERS-R and CLASS Emotional Support was .52 and between ECERS-R and CLASS Instructional Support was .40, in a sample of 224 state-funded prekindergarten classrooms in six states.

4.4. Links between quality and children's outcomes

As with the past research using ECERS-R and CLASS Pre-K (Burchinal et al., 2011; Burchinal, Zaslow, & Tarullo, 2016), the associations between ECERS-3 and children's outcomes are either nonsignificant or small, raising general questions about the field's tools for measuring quality. There are several possible explanations for these weak associations, and it is possible that several of them are true simultaneously.

One possibility is that the tools and strategies for measuring quality are too imprecise. The field may have a general understanding of quality, but we do not yet know exactly how to translate that understanding into a measurement system. Zaslow, Burchinal, Tarullo, and Martinez-Beck (2016) argue that our knowledge of high-quality instruction is improving and involves "engaging activities, small and large group instruction, and sequenced presentation of instructional materials that allow for deep learning. . ." (p. 80). No single tool exists for measuring all important aspects of quality; however, Zaslow and her colleagues note that combining interaction-specific and domain-specific quality measures may be necessary.

Likewise, there may be a mismatch between the broad nature of the quality measures and the more narrowly defined constructs of language, literacy, and math. Tools are emerging that measure instructional quality in domains such as math (e.g., Sarama & Clements, 2007) and literacy (e.g., Holland Coviello, 2005). There also is emerging evidence that more narrowly focused tools show stronger associations with outcomes in those same areas (Purpura, Hume, Sims, & Lonigan, 2011; Purpura, Logan, Hassinger-Das, & Napoli, 2017; Zaslow et al., 2016).

Another possibility is that child outcomes are not measured precisely enough. Even if quality is well measured, short child assessment batteries that attempt to cover a broad range of skills may simply not be accurate enough to detect associations with quality. They may also reify the well-documented concerns in using norm-referenced standardized tests with preschoolers that

were designed to span early childhood through adolescence, or even adulthood. These concerns include insufficient floors, steep item gradients, norming bands that are too wide to capture rapid developmental changes during this age period, and difficulty in achieving truly representative norming samples (Nagle, 2007; Willis & Dumont, 2003). The fact that, in the current study, we found links between Math Activities and growth in children's social skills, but not between Math Activities and growth in math skills as measured by the Applied Problems subtest, provides some evidence that our outcomes are not measured with enough precision. Indeed, Purpura and colleagues have argued that early math assessments actually capture general skills like critical thinking and comprehension (Purpura et al., 2017) or language development (Purpura et al., 2011). This would explain why the Applied Problems subtest was linked to Learning Opportunities in the current study. An additional consideration is that the pre- and posttest batteries were less than 7 months apart on average. This may not be enough time, or the batteries might not be sensitive enough, for developmental changes to emerge.

Increasing the precision with which child outcomes are measured is an important task because it has significant implications for determining the relationships between young children's preacademic skills and program quality features. However, the field is also faced with other measurement issues that have emerged in recent years. For example, we must also continue to explore factors both within and outside learning environments, such as child-level factors (e.g., IQ, resiliency), access to early childhood mental health services, family supports, and family resources that may contribute significantly to the development and success of young children (Zaslow et al., 2010). Early childhood program quality may play only a small role in children's achievement.

Weaknesses in the way these tools are employed for measuring classroom quality may lead to low reliability, providing another possible explanation for the weak associations. For instance, truly understanding classroom quality may require multiple days of observation (Mashburn, Downer, Rivers, Bracket, & Martinez, 2014). Further, the reliability standards typically used for ECERS-R, ECERS-3, and CLASS—including those used in this study—rely on percent agreement within one scale point. In the current sample, the standard deviation for the Total Score was only .80, so individual raters may produce scores that are more than a standard deviation apart and still be considered "reliable."

Identifying potential problems with current tools and data collection strategies may help users understand the limitations of measures, but it does not provide solutions for users such as QRIS administrators, professional development providers, and intervention evaluators. All potential solutions, including increasing the number of days of observation and strengthening reliability standards, involve significant additional resources. Narrowly focusing on specific domains might increase our predictive ability, but would either mean forgoing an understanding of quality in multiple domains or using several different tools, likely requiring multiple days of observation.

Instead, we recommend that users think carefully about why they are measuring quality and identify reasonable expectations for the administration of quality measurement tools. If the goal is to identify which programs generally provide children with safe, stimulating, enjoyable early learning experiences, and which programs need more support to provide for foundational quality, there is evidence that ECERS-3 is useful for such purposes. It is significantly correlated with other tools, it shows some associations with children's growth, and has been well received by early childhood professionals including early childhood leaders in the states where the data for the current study were collected. If the goal, however, is to promote specific academic outcomes, then those need to be clearly defined and significant resources should be devoted to measuring both the quality of instruction and children's outcomes in those areas.

4.5. Importance of executive function

Although for many domains the pattern of findings indicates small or no associations between ECERS-3 (or CLASS) and children's growth, the associations with executive function are fairly consistent. As with the academic domains, associations with executive function are small and likely suffer measurement error. The consistency in the pattern of associations, however, implies that they are more than noise and therefore merit additional discussion.

As noted above, the broad nature of ECERS-3 and CLASS may make them poorly suited for predicting gains in the narrowly defined areas of math and literacy. Executive function, on the other hand, is a broad set of skills that are necessary for school success, which may explain why it appears to be more closely tied to global classroom quality. Executive function is important because children with stronger executive function show greater gains on tests of early math, language, and literacy development during early childhood than peers with weaker executive function (Allan & Lonigan, 2011; Duncan et al., 2007). Likewise, children with strong executive function are better equipped to develop advanced social skills because they can regulate behaviors that ensure successful interactions with others (Lewis & Carpendale, 2009). These data provide evidence that multiple aspects of early childhood classroom quality, as measured by ECERS-3 and CLASS, promote this important set of skills.

Additionally, the specific content of the ECERS-3 and the CLASS may map especially well onto executive function. The items on the Learning Opportunities and Teacher Interactions subscales, as well all three CLASS Pre-K domains, emphasize children making decisions and developing independence within open-ended activities, teachers scaffolding of appropriate behaviors within the context of a predictable classroom routine, and teachers providing opportunities for peer interaction – all of which are critical to the development of executive function in young children (Graziano, Garb, Ros, Hart, & Garcia, 2016; Lonigan, Allan, & Phillips, 2017).

4.6. Limitations

This study relied on data collected by three agencies within participating states, in part for their own purposes. For that reason, the samples are not representative of any particular group of early childhood classrooms. This lack of representativeness opens up the possibility that a different sample might have yielded a different factor structure or different associations with children's outcomes. For instance, if this sample is systematically of higher or lower quality than the population, and if associations between ECERS-3 and outcomes are stronger at the higher or lower end of the scale, we may be falsely minimizing or exaggerating associations. This is the first large-scale study using the ECERS-3 and the findings require replication.

The modest size of the In-Depth sample is also a limitation. The In-Depth sample was large enough to detect small-to-moderate effects of both ECERS-3 and CLASS Pre-K on a handful of outcomes, but our analyses may be missing some very small effects. We are not certain, however, that such small effects would be of interest to the field.

Another limitation was that training and data collector reliability were overseen by the partnering agencies within each state who used their own methods and standards to determine inter-rater reliability. Further, each state relied on *consensus scoring* and percent of items that were within one scale-point of the consensus score, rather than stronger methods for measuring inter-rater reliability such as weighted kappas or intraclass correlations between

original (rather than consensus) scores (McHugh, 2012). In some ways, this is a strength of the study because it reflects the types of standards and variation present across the nation and, therefore, may mean that data are similar to what would be seen in any large-scale data collection effort using the ECERS-3. That said, it is possible that different data collectors had different interpretations of the scale, decreasing our ability to find associations with children's outcomes.

Finally, the effect sizes in this study are all small (Cohen, 1992). The effect sizes (standardized parameter estimates) for significant associations range from .06 to .10. One way to contextualize the size of research findings is to compare them to other research on similar topics or with similar outcome measures. Burchinal (2017) noted that typical effect sizes for associations between early care and measures of education quality and children's outcomes are typically less than .10 and often less than .05. Thus, the size of these effects is in line with past work, but there is no evidence that this revision improves prediction of children's outcomes, as intended by the ECERS-3 authors.

Another way to contextualize effect sizes is to compare them to other predictors that are considered important. NICHD Early Child Care Research Network (2016) used that approach and reported that the effect sizes for parenting quality in predicting 54-month outcomes were .34 for total language and .32 for preacademic skills or roughly triple the effect found for significant associations between child care quality and outcomes in the NICHD study or the current study.

Despite their small size, we believe these associations are meaningful given the short time frame between pre- and posttest; the error inherent in measuring children's outcomes at this age; their similarity to effect sizes seen in previous research linking quality to children's outcomes; and the large number of factors other than preschool quality that impact children's development.

Finally, although the four-factor solution was psychometrically stronger than the single factor, this work does not provide compelling evidence that users of the ECERS-3 should create or use these subscales, unless there is a project-specific or theoretical reason to do so. The subscales do not markedly improve prediction of children's outcomes, they do involve loss of some information, they have not been replicated in other samples, and they are less intuitive than the Total Score. For the time being, we recommend that most users follow the author's guidelines for scoring, and calculate a simple mean of all items.

4.7. Next steps

These data and analyses raise many important questions that we will endeavor to answer in the future. First, it is possible that alternate scoring using Item-Response Theory would yield stronger associations with outcomes. Although such a scoring system might be too complicated for regular use by local- and state-level evaluators, it could provide deeper understanding of the underlying structure of the ECERS-3 and how various conceptualizations of quality relate to children's growth. Second, whereas our findings indicate limited association with growth across all children, the links may be stronger for some children than others or under certain circumstances than others. To address this possibility, future analyses will investigate various moderators such as family income, race, and announced versus unannounced visits. We also plan to investigate the possibility that associations between ECERS-3 and children's outcomes are nonlinear, which Le et al. (2015) found using ECERS-R. Finally, users of these tools need information about how the ECERS-R and ECERS-3 relate to one another. Although our current data set does not include ECERS-R observations, some state and local agencies have collected such data. We are working with them to combine data and investigate how the two tools relate to one another.

4.8. Conclusions and implications

Researchers, policymakers, and professional development providers need reliable and valid tools for assessing early childhood classroom quality. This study partially supports ECERS-3 as such a tool, but also demonstrates the need for further tool development and additional psychometric work on this and others. Questions remain about which classroom quality constructs are most closely aligned to children's outcomes, how to best measure children's outcomes, how stable classroom quality is across the school year, and how reliable data collectors must be for observations to be linked to children's growth. Despite these questions, continuing to focus on quality is critical as a means of ensuring that all children have safe, warm, and stimulating early childhood experiences.

Author note

The authors of this paper are past and present colleagues of the ECERS-3 authors and have worked closely with Dr. Richard Clifford, the ECERS-3 second author. Dr. Clifford provided feedback when the authors were applying for the Institute for Education Science grant that funded this project and facilitated relationships between the research team and the states that collected the data. As is typical for ECERS-3 data collection, Drs. Harms and Cryer were compensated for their time in training staff to conduct ECERS-3 observations. However, none of the ECERS-3 authors played a role in data collection, analysis decisions, or writing of this paper. The Branagh Information Group, which provides the software to collect ECERS-3 data electronically, was a subcontractor on this grant. Their role involved deidentifying electronic data and providing it to the research team, but they had no role in analysis decisions or writing of this paper.

This research was supported under a grant from the Institute for Education Sciences, U.S. Department of Education (R305A150109). However, those contents do not necessarily represent the policy of the Department of Education, and endorsement by the Federal Government should not be assumed. We are grateful for our partners in Georgia, Pennsylvania, Washington, and the Branagh Information Group, without whom this work would not have been possible. Special thanks to Syndee Kraus and Katie Hume who coordinated this multistate effort. And, we wish to thank the many early childhood teachers, parents, and children who participated in this project.

References

Administration for Children & Families. (2015). *QRIS resource guide*.. Retrieved from. https://https://qrisguide.acf.hhs.gov/index.cfm?do=resourceguide

Allan, N. P., & Lonigan, C. J. (2011). Examining the dimensionality of effortful control in preschool children and its relation to academic and socioemotional indicators. *Developmental Psychology*, *47*, 905–915. http://dx.doi.org/10.1037/a0023748

Allison, P. D. (2001). *Missing data (quantitative applications in the social sciences)*. Thousand Oaks, CA: Sage Publications.

Auger, A., Farkas, G., Burchinal, M. R., Duncan, G. J., & Vandell, D. L. (2014). Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental Psychology*, *50*(12), 2559–2571. http://dx.doi.org/10.1037/a0037995

Burchinal, M. (2017). Measuring early care and education quality. *Child Development Perspectives*, http://dx.doi.org/10.1111/cdep.12260

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–32). Baltimore, MD: Brookes Publishing Company.

Burchinal, M., Zaslow, M. J., & Tarullo, L. (Eds.), (2016). *Monographs of the society for research in child development* (Vol. 81) http://dx.doi.org/10.1111/mono.12248

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response

formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116. http://dx.doi.org/10.1080/08824096.2013.836937
Cassidy, D. J., Hestenes, L. L., Hedge, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly*, *20*, 345–360. http://dx.doi.org/10.1016/j.ecresq.2005.07.005
Clifford, R. M., Barbarin, O., Chang, F., Early, D. M., Bryant, D., Howes, C., . . . & Pianta, R. (2005). What is pre-kindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science*, *9*(3), 126–143. http://dx.doi.org/10.1207/s1532480xads0903_1
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155
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*(5), 678–696. http://dx.doi.org/10.1080/10409289.2011.588041
Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Brooks-Gunn, J. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428
Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., & Zill, N. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, *78*(2), 558–580. http://dx.doi.org/10.1111/j.1467-8624.2007.01014.x
Elicker, J., & Thornburg, K. R. (2011). *Evaluation of quality rating and improvement systems for early childhood programs and school-age care: Measuring children's development*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services (Research-to-Policy, Research-to-Practice Brief OPRE 2011-11c).
Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for assessments of child care quality and its relation to child development. *Developmental Psychology*, *49*(1), 146–160. http://dx.doi.org/10.1037/a0027899
Graziano, P. A., Garb, L. R., Ros, R., Hart, K., & Garcia, A. (2016). Executive functioning and school readiness among preschoolers with externalizing problems: The moderating role of student–teacher relationship. *Early Education and Development*, *27*(5), 573–589. http://dx.doi.org/10.1080/10409289.2016.1102019
Harms, T., Clifford, R., & Cryer, D. (2005). *Early Childhood Environment Rating Scale* (rev. ed.). New York, NY: Teachers College Press.
Harms, T., Clifford, R., & Cryer, D. (2015). *Early Childhood Environment Rating Scale* (3rd ed.). New York, NY: Teachers College Press.
Holland Coviello, R. (2005). *Language and literacy environment quality in early childhood classrooms: Exploration of measurement strategies and relations with children's development*. State College, PA: Pennsylvania State University.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *32*, 179–185.
Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D. M., Clifford, R. M., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*, 27–50.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
Jaffe, L. E. (2009). *Development, interpretation, and application of the W score and the relative proficiency index*. Rolling Meadows, IL: Riverside Publishing (Woodcock–Johnson III Assessment Service Bulletin No. 11).
Klein, L., & Knitzer, J. (2006). *Effective preschool curricula and teaching strategies*. NY: National Center for Children Living Poverty.
La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409–426. Retrieved from. http://www.jstor.org/stable/3202821
La Paro, K. M., Thomason, A. C., Lower, J. K., Kintner-Duffy, V. L., & Cassidy, D. (2012). Examining the definition and measurement of quality in early childhood education: A review of studies using the ECERS-R from 2003 to 2010. *Early Childhood Research and Practice*, *14*(1). Retrieved from. http://ecrp.uiuc.edu.libproxy.lib.unc.edu/v14n1/laparo.html
Le, V. N., Schaack, D. D., & Setodji, M. C. (2015). Identifying baseline and ceiling thresholds within the Qualistar Early Learning Quality Rating and Improvement System. *Early Childhood Research Quarterly*, *30*, 215–226.
LeBuffe, P. A., & Naglieri, J. A. (2012). *The Devereux Early Childhood Assessment for Preschoolers, Second Edition (DECA-P2) assessment, technical manual, and user's guide*. Lewisville, NC: Kaplan.
Lewis, C., & Carpendale, J. I. (2009). Introduction: The links between social interaction and executive function. *New Directions in Child and Adolescent Development*, *2009*, 1–15. http://dx.doi.org/10.1002/cd.232
Lonigan, C. J., Allan, D. M., & Phillips, B. M. (2017). Examining the predictive relations between two aspects of self-regulation and growth in preschool children's early literacy skills. *Developmental Psychology*, *53*(1), 63–76. http://dx.doi.org/10.1037/dev0000247
Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power and efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*, 146–155. http://dx.doi.org/10.1007/s11121-012-0357-3
Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., & Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development*, *79*(3), 732–749. http://dx.doi.org/10.1111/j.1467-8624.2008.0115
McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The Head-Toes-Knees-Shoulders task. *Frontiers in Psychology*, *5*, 3–14. http://dx.doi.org/10.3389/fpsyg.2014.00599
McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual. Woodcock–Johnson IV*. Rolling Meadows, IL: Riverside.
McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.
Moiduddin, E., Aikens, N., Tarullo, L., West, J., & Xue, Y. (2012). *Child outcomes and classroom quality in FACES 2009*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families (OPRE Report 2012-37a). Retrieved from. http://www.acf.hhs.gov/sites/default/files/opre/faces_2009.pdf
Nagle, R. (2007). Issues in preschool assessment. In B. Bracken, & R. Nagle (Eds.), *Psychoeducational assessment of preschool children* (pp. 29–48). New York: Routledge.
National Association for the Education of Young Children. (2009). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8: A position statement of the National Association for the Education of Young Children*. Washington, DC: National Association for the Education of Young Children. Retrieved from http://www.naeyc.org/positionstatements/dap
National Center for Education Statistics. (n.d.). *Early Childhood Longitudinal Program (ECLS): Data collection procedures*. Retrieved from http://nces.ed.gov/ecls/birthdataprocedure.asp.
National Education Association. (2008). *Early childhood education and school readiness: A policy brief*. Washington, DC: National Education Association.
NICHD Early Child Care Research Network. (2016). Child-care effect sizes for the NICHD study of early child care and youth development. *American Psychologist*, *6*(2), 99–116.
Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PLOS ONE*, *11*(12) http://dx.doi.org/10.1371/journal.pone.0167660
Pianta, R., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Pre-K version*. Baltimore, MD: Brookes Publishing Company.
Ponitz, C. C. (2008). *Head-Toes-Knees-Shoulders*. Charlottesville, VA: Center for the Advanced Study of Teaching and Learning, Social Development Lab.
Ponitz, C. C., McClelland, M. M., Mathews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology*, *45*, 605–619. http://dx.doi.org/10.1037/a0015365
Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, *110*(4), 647–658. http://dx.doi.org/10.1016/j.jecp.2011.07.004
Purpura, D. J., Logan, J. A. R., Hassinger-Das, B., & Napoli, A. R. (2017). Why do early mathematics skills predict later reading? The role of mathematical language. *Developmental Psychology*, *53*(9), 1633–1642. http://dx.doi.org/10.1037/dev0000375
SAS Institute. (2002–2012). *SAS/STAT version 9.4 [computer software]*. Cary, NC: SAS Institute.
Sakai, L. M., Whitebook, M., Wishard, A., & Howes, C. (2003). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing the differences between the first and revised versions. *Early Childhood Research Quarterly*, *18*, 427–445. http://dx.doi.org/10.1016/j.ecresq.2003.09.004
Sarama, J., & Clements, D. (2007). *Classroom Observation of Early Mathematics–Environment and Teaching*. Buffalo: State University of New York. Unpublished instrument.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 147–177. http://dx.doi.org/10.1037/1082-989x.7.2.147
Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock–Johnson IV*. Rolling Meadows, IL: Riverside Publishing Company.
Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academy Press.
Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The Search for optimal cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, *74*, 267–288.
Thurstone, L. L. (1954). An analytical method for simple structure. *Psychometrika*, *9*(3), 173–182. http://dx.doi.org/10.1007/BF02289182
Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *Compendium of quality rating systems and evaluations*. Washington, DC: Office of Planning, Research, and Evaluation.
U.S. Department of Education. (2013). *Race to top—Early learning challenge*.. Retrieved from http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/index.html
Vandell, D. L. (2004). Early child care: The known and the unknown. *Merrill-Palmer Quarterly*, *50*(3), 387–414. http://dx.doi.org/10.1353/mpq.2004.0027

Vandell, D. L., & Wolfe, B. (2000). *Child care quality: Does it matter and does it need to be improved?* Retrieved from https://aspe.hhs.gov/system/files/pdf/174171/report.pdf

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*(2), 199–209. http://dx.doi.org/10.1016/j.ecresq.2012.12.002

Willis, J., & Dumont, R. (2003). *Top ten problems with normed achievement tests for young children*. Retrieved from. http://alpha.fdu.edu/psychology/ten_top_problems_with_normed_ach.htm

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, & Zaslow, M. J. (2013). *Investing in our future: the evidence base on preschool education*. Washington, DC: Society for Research in Child Development & Foundation for child Development. Retrieved from. https://www.fcd-us.org/assets/2016/04/Evidence-Base-on-Preschool-Education-FINAL.pdf

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L., & Burchinal, M. (2010). *Quality dosage, thresholds, and features in early childhood settings: A review of the literature, OPRE 2011-5*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zaslow, M., Burchinal, M., Tarullo, L., & Martinez-Beck, I. (2016). V. Quality thresholds, features, and dosage in early care and education: Discussion and conclusions. *Monographs of the Society Research in Child*, *81*, 75–87. http://dx.doi.org/10.1111/mono.12240