Published in final edited form as:

Early Child Res Q. 2018 ; 42: 158–169. doi:10.1016/j.ecresq.2017.10.001.

Using the Early Childhood Environment Rating Scale-Revised in High Stakes Contexts: Does Evidence Warrant the Practice?

Claude Messan Setodji,

RAND Corporation, 4570 5th Avenue, Suite 600, Pittsburgh, PA 15213, Tel: (412) 683 2300 (ext. 4920), Fax: (412) 802 4966, setodji@rand.org

Diana Schaack, and

University of Colorado Denver

Vi-Nhuan Le

NORC at the University of Chicago

Abstract

Increasingly, states establish different thresholds on the Early Childhood Environment Rating Scale–Revised (ECERS–R), and use these thresholds to inform high-stakes decisions. However, the validity of the ECERS–R for these purposes is not well established. The objective of this study is to identify thresholds on the ECERS–R that are associated with preschool-aged children’s social and cognitive development. Applying non-parametric modeling to the nationally-representative Early Childhood Longitudinal Study Birth Cohort (ECLS–B) dataset, we found that once classrooms achieved a score of 3.4 on the overall ECERS–R composite score, there was a leveling-off effect, such that no additional improvements to children’s social, cognitive, or language outcomes were observed. Additional analyses found that ECERS–R subscales that focused on teaching and caregiving processes, as opposed to the physical environment, did not show leveling-off effects. The findings suggest that the usefulness of the ECERS–R for discerning associations with children’s outcome may be limited to certain score ranges or subscales.

Keywords

QRIS; thresholds; ECERS–R; non-parametric models

During the 1990s, there was an uptick in attention paid to the quality of the care and education that young children experienced in their child care settings in the United States. This attention was driven in part by several multi-state studies that measured preschool quality using the Early Childhood Environment Rating Scale (ECERS; Harms, Clifford, & Cryer, 1980), and reported a national child care quality crisis, particularly for lower-income children (Kagan & Cohen, 1997; Helburn, Culkin, Morris, Mocan, Howes, et al., 1995;

Correspondence to: Claude Messan Setodji.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Loeb, Fuller, Kagan, & Carrol, 2004; Whitebook, Phillips, & Howes, 1993). This body of research also demonstrated weak, but positive, associations between the quality of preschool classrooms, as measured by the ECERS, and a number of developmental benefits for preschool-aged children (Peisner-Fienberg & Burchinal, 1997). Preschool quality, as measured by the ECERS, was also shown to be positively associated with children's academic achievement at the early elementary grades (Peisner-Feinberg, Burchinal, Culkin, Clifford, Howes, et al., 2001).

As a result, states began developing child care accountability and quality improvement initiatives, many of which were undergirded by the ECERS-R (the revised version of the ECERS). Presently, quality rating and improvement systems (QRIS) are the most prominent early care and education (ECE) reform effort in the United States, now being implemented in 41 states (Tout, Starr, Stoli, Moodie, Kirby, et al., 2010), and featured as a required initiative in the federal Race-to-the-Top Early Learning Challenge grants (U.S. Department of Education, 2016). QRIS establish ECE program, classroom, and practitioner quality standards, set thresholds or quality levels on these standards, and measure and monitor the extent to which classrooms meet quality levels. QRIS then provide an overall, summary program quality rating made available to families to assist in their ECE decision-making. Although all states construct their QRIS differently, approximately two thirds of states currently use the ECERS-R as part of their QRIS (QRIS Compendium, n.d.; Administration for Children and Families, 2013).

Currently, 37 states' QRIS attach financial incentives to a program's overall quality rating, the majority of which includes the ECERS-R as a key component (National Center on Early Childhood Quality Assurance, 2017). These financial incentives can include awarding bonuses to teachers based on different thresholds that they have met on the classroom assessment, awarding different levels of payment for children receiving child care subsidies based on a program's rating, and providing programs that meet a certain threshold of quality with improvement grants (Hamilton, Bates, Mitchell, & Workman, 2015; Mitchell, 2012; QRIS Compendium, n.d.). In some states, families are also awarded different levels of tuition support based on the rating of the preschool classroom they selected for their child (Schaack, Tarrant, Boller, & Tout, 2012). States provide these incentives tied to higher ECERS-R scores under the assumption that as classrooms meet higher thresholds of quality on the measure, better child outcomes will follow (Zellman & Perlman, 2008).

Yet the ECERS was not originally developed for such high stakes purposes but instead was developed in 1980 as a checklist to help ECE programs prepare for the National Association of the Education of Young Children (NAEYC) Program Accreditation (Frank Porter Graham Child Development Institute, 2003). The definition of quality adopted by the ECERS-R is thus consistent with both the NAEYC program accreditation standards as well as with the Child Development Associate requirements that focuses on the professional knowledge teachers need to facilitate high quality classrooms (Nord et al., 2007). Revised in 1998, the ECERS-R is currently constructed with 43 items organized into seven subscales, including Space and Furnishings (8 items), Personal Care Routines (6 items), Facilitation of Children's Language-Reasoning (4 items), Learning Activities (10 items), Teacher Interactions (5 items), Program Structure (4 items), and Supports for Parents and Staff (6

items). Two appealing features of the ECERS-R that contributes to it being one of the most widely used ECE quality measures in state policy are its ease of use and its comprehensiveness. First, the ECERS-R provides specific information about dimensions of quality in which programs are low scoring, and QRIS coaches can use this information to target quality improvement efforts, grants, and professional development to low scoring areas to help programs improve their ECERS-R scores. Second, as a global measure of quality, it is comprehensive in scope, and assesses both structural and process aspects of quality.

Recently, researchers, practitioners, and QRIS designers have begun to raise concerns about the ECERS-R, especially for use in high stakes contexts (Gordon, Fujimoto, Kastner, Korenman, & Abner, 2013; Zellman, Perlman, Le, & Setodji, 2008). Its critics believe that the ECERS-R does not focus enough on aspects of teaching processes that promote conceptual development across learning domains, or on the types of caregiving behaviors that promote secure and trusting relationships with children that facilitate adaptive social-emotional functioning that enable children to engage in learning (Sabol & Pianta, 2014). Instead, those critics believe that the ECERS-R places too much emphasis on environmental quality (Cassidy, Hestenes, Hansen, Hedge, Hestenes, et al., 2005). Concerns have also been raised about the empirical evidence available on the ECERS-R to justify its use in high-stakes contexts (Gordon et al., 2013). Developed as a self-assessment measure intended to provide feedback to programs about their quality, it is unknown whether the ECERS-R can be used to support high-stakes decisions.

Purpose of this Study

The goal of this study is to evaluate the potential utility of the ECERS-R by examining the associations between ECERS-R scores and children's cognitive and social-emotional outcomes, with an emphasis on assessing whether the associations are limited to certain ECERS-R score ranges or on particular ECERS-R subscales. More specifically, we address the following research questions:

1. Are there thresholds on the ECERS-R that are related to preschool children's cognitive and social-emotional outcomes?
2. Do the thresholds change when considering the ECERS-R subscales compared with the total ECERS-R score?

Psychometric Properties of the ECERS-R

In light of the fact that the use of the ECERS-R has expanded and is now being widely used as an accountability measure on which high-stakes decisions are made, there is a growing body of literature that has begun to examine the psychometric properties of the ECERS-R. A number of studies, for example, have subjected the measure to factor analytic techniques to examine its dimensionality. Several studies have observed the ECERS-R to be unidimensional (Holloway et al., 2001; Perlman, Zellman, & Le, 2004), or two dimensional consisting of factors that tap into the physical environment/materials and teacher interactions (Cassidy, Hestenes, Hansen, Hedge, Hestenes, et al., 2005; Sakai, Whitebook, Wishards, &

Howes, 2003), but no studies have found evidence of the seven scales described in the ECERS-R.

Using item response theory, Gordon and colleagues (2013) also demonstrated evidence of individual item multidimensionality, resulting in rating category disorder on 32 of the 36 items they examined. Within the ECERS-R, individual items are composed of multiple binary indicators. Training procedures for the ECERS-R indicate that raters should stop scoring an item once a classroom has not met a particular indicator. As a result of this scoring convention, when an observer assigned a low score on one of the items due to a classroom not meeting a particular indicator on the item, the observer often missed scoring other indicators within the item as higher quality. This may mean that the level of quality needed for classrooms to earn a score of 6, for example, could be less than the level of quality needed to earn a score of 5 (Gordon et al. 2013). This type of research has raised some concerns about how the ECERS-R is constructed. A small body of research has also examined the association of the ECERS-R against other measures of developmentally appropriate practices, classroom structural quality indices, and children's developmental outcomes. In this research, small to moderate correlations with overall ECERS-R scores have been found with measures of instructional quality such as the Classroom Scoring and Assessment System (CLASS; Mashburn, Pianta, Hamre, Downer, Barbarin, et al., 2008) and the Early Language and Literacy Classroom Observation (ELLCO; Smith & Dickinson, 2002). In a recent meta-analysis of child care quality studies conducted in both the U.S. and in international settings, Vermeer and colleagues (2016) found strong, positive correlations between the ECERS-R and the teacher sensitivity sub-scale on the Caregiver Interaction Scale (Arnett, 1989). However, with respect to classroom structural quality, weak correlations have been detected between the ECERS-R and teachers' education level and classroom ratios (Early et al., 2007; Gordon et al., 2013; Zellman et al., 2008). No significant associations were detected between ECERS-R scores and classroom group sizes in a meta-analysis using 17 studies (Vermeer et al., 2016). Mixed evidence has also been found when examining the associations between the ECERS-R and developmental outcomes for young children. Some studies have reported weak but positive linear associations with preschoolers' receptive and expressive language skills, applied problem-solving skills, and some indices of social-emotional development (Early, Bryant, Pianta, Clifford, Burchinal, Ritchie, et al., 2006; Mashburn et al., 2008). Other studies have failed to find significant associations between the ECERS-R and these dimensions of children's developmental outcomes (Sabol & Pianta, 2014; Zellman, et al., 2008).

Quality Thresholds on the ECERS-R

One aspect of the ECERS-R that has been understudied is the existence of thresholds or cut-points along the ECERS-R that may differentiate among the different levels of children's developmental functioning. As discussed by Burchinal, Vandergrift, Pianta, and Mashburn (2010), thresholds are of particular interest to researchers and policymakers because they can inform efficient allocation of resources. They note:

Most of the literature has examined linear associations, yielding findings that higher quality is better and lower quality is worse (Vandell, 2004), but identification

of thresholds in the association between quality and child outcomes has been a goal of researchers and policy makers for several reasons. A primary goal has been to identify levels in the association between quality and child outcomes at which the linear association begins to asymptote or level off, above or below which there is little evidence of increases in learning associated with increases in quality. A threshold that indicated that the quality-outcome association levels off above a given level of quality would suggest that policies should focus on improving quality up to that threshold level, but improving quality above that point may not be necessary for improving child outcomes. Policy to address this goal would invest in lower or average quality classrooms while leaving classrooms with quality scores above the threshold alone. In contrast, it is possible a threshold could define the minimum level at which a positive association between quality and outcomes is observed. In this scenario, there may be no detected relation between quality and outcome gains until quality reached a certain point on the scale... This form of threshold effect would suggest that it is especially important to ensure that children experience at least the minimum level of quality child care in order for those experiences to be related to improved child outcomes. It would point perhaps to not allowing vouchers to pay for care that was below the threshold, while also incentivizing teachers above the threshold to continue to improve (Burchinal et al., 2010, p.167).

It is important to note that when state policymakers select cut-points on their QRIS, the points are intended to represent meaningful levels of quality along the score distribution. For example, many state policymakers adopt a lower threshold on the ECERS-R between a score of 3.0 and 3.5 because the lower threshold denotes a minimum level of quality to which they believe children should be exposed (QRIS Compendium, n.d.). Similarly, although all programs are encouraged to strive for the highest ECERS-R score of 7.0, most states award maximum QRIS points to programs that receive an ECERS-R score between 5.0 and 5.5 because these scores denote “good enough” levels of quality (QRIS Compendium, n.d.). Thus, in setting cut-points on their QRIS, state policymakers are implicitly assuming that the chosen cut-points are not arbitrarily-functioning thresholds, but reflect either a minimum level of quality that needs to be reached for the well-being of children, or reflect an asymptotic leveling-off, such that additional improvements in quality are not associated with commensurate increases in child outcomes. Despite the fact that many states have set thresholds on the ECERS-R, and are using these thresholds to distribute resources to classrooms and target professional development efforts, little research has been conducted to determine whether the selected cut-points are empirically functioning as intended. Given the potential high-stakes consequences attached to these thresholds, professional standards recommend examining the validity of the thresholds so that states can construct QRIS in ways that are empirically defensible and that help them meet their goals (American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014).

Early studies examining the existence of cut-points on the unrevised version of the ECERS created thresholds based on the distribution of the data, or on the hypothesized anchors determined by the authors. For instance, Howes, Phillips, and Whitebook (1992) classified

classrooms into one of four quality categories based on the distribution of ECERS scores in their sample. Their classifications ranged from inadequate (scores of 1.00–2.90), barely adequate (scores of 3.00–3.90), good (scores of 4.00–4.90) to very good (scores of 5.00–7.00). They found significant associations between ECERS scores and the quality of peer relationships, such that children in classrooms scoring above 4.00 had better relationships with their peers. Similarly, Burchinal, Peisner-Feinberg, Bryant, and Clifford (2000) combined data across three large-scale child care studies and using the quality thresholds defined by the developers of the ECERS, classified classrooms into “poor” (scores of 1.00–2.99), “average” (scores of 3.00–4.99), or “good” (scores above 5.00) quality groups. They found that children whose classrooms were classified into the poor-quality group showed more behavioral problems and demonstrated lower verbal, reading, and mathematical problem-solving skills than did children whose classrooms were classified into the average- or good-quality groups.

However, as more and more states are using the revised version of the ECERS in policy contexts, and as the ECERS-R has become a part of most teacher’s vernacular and influences how they organize their classroom structure, more current research on the functioning of the ECERS-R, and specifically on its quality thresholds are needed. The most recent research on quality thresholds on the ECERS-R departs from previous methodological approaches in that the analyst does not have to specify a threshold *a priori*. Instead, it is possible for the analyst to empirically determine the threshold using generalized additive mixed models, a non-parametric method that allows for the detection of a full range of thresholds that may be masked by the selection of predetermined thresholds. Using this non-parametric approach on data from Colorado’s QRIS, Le, Schaack, and Setodji (2015) found the existence of two thresholds on the ECERS-R, one at 3.40 and the second at 4.60. Within this score range, the associations between classroom quality, as measured by the ECERS-R, and children’s receptive and expressive language and applied mathematical problem solving skills were significantly positive. However, between the ECERS-R score range of 1.0 and 3.40, and beyond an ECERS-R score of 4.60, there were no associations between the ECERS-R and children’s language or mathematics skills.

These findings indicate that, on average, children in classrooms with an ECERS-R score of 1.0 showed similar outcomes to children in classrooms with an ECER-R score of 2.5 or 3.4. For children in classrooms with ECERS-R scores between 3.4 and 4.6, increases in ECERS-R scores were associated with increases in children’s outcomes. Beyond an ECERS-R score of 4.6 there was a leveling-off effect, where additional improvements in the ECERS-R scores were not associated with improvements in child outcomes. These results suggest that although the ECERS-R can show positive associations with children’s developmental outcomes, its usefulness at detecting associations may be limited to certain score ranges, and it may cease to reliably discern associations once a mid-level rating has been reached.

Explaining the Leveling-Off Effects

It is possible that the leveling-off effect observed in the aforementioned study of the ECERS-R may be an artifact of the composition of the instrument (Setodji, Le, & Schaack, 2013), which is more focused on environmental quality than on process quality (Cassidy et

al., 2005; Douglas, 2004; Sabol & Pianta, 2014). That is, because the adequacy of the physical environment is hypothesized to be more distally related to child outcomes than are actual caregiving and instructional processes (NICHD Early Childhood Research Network, 2002), the leveling-off effect observed in the Le et al. (2015) study may be due to the relatively greater emphasis of the ECERS-R on environmental quality than on process quality. This result also suggests that the leveling-off effect may be more pronounced on the subscales that emphasize the adequacy of the physical environment than on the subscales that emphasize specific care and instructional practices. From this perspective, we would expect the more process-oriented subscales, namely Interaction, Learning Activities, and Language-Reasoning, that assess such teacher behaviors as responding positively to children's cues and needs, facilitating learning activities across instructional domains, and engaging in book reading and language play, to be less likely to show leveling-off effects than the subscales that emphasize the availability of materials and the safety of the physical environment (i.e., Space and Furnishings and Personal Care Routines).

Gordon et al (2013) also demonstrated that ECERS-R items and subscales may consist of a mixture of both process and environmental quality, with a focus on environmental quality at the lower end of the score distribution and process quality at the upper end. The Program Structure subscale is indicative of this type of construction. As an illustrative example, this scale includes an item about children's free play. At the lower end of the score distribution for this item, classrooms are assessed on whether toys, games, and equipment are accessible to children during free play. The emphasis on the availability of materials at the lower end of the score distribution can be contrasted with an emphasis on teacher behaviors at the upper end of the score distribution, where programs are assessed on the frequency that teachers facilitate and extend children's play by making suggestions and scaffolding their play (Child Care Resources, 2006). Thus, although not as process-oriented as some other ECERS-R subscales, the Program Structure subscale may also be less likely to show leveling-off effects.¹

This study has been designed to build on Le et al.'s (2015) previous study of thresholds on the ECERS-R conducted in one state policy context. Specifically, this study examines the existence of quality thresholds along the ECERS-R in a nationally-representative sample of classrooms and children and explores whether quality thresholds are similar across ECERS-R subscales.

Methods

Sample

We analyzed data from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), which was a nationally-representative study that followed children over time and collected information on children's early life experiences starting from birth through kindergarten entry. The ECLS-B target population were all children born in 2001 excluding children whose mothers were less than 15 years of age when giving birth and children who were adopted prior to the 9-month assessment. The ECLS-B study used birth certificates obtained

¹We do not discuss the Parents and Staff subscale because it was not administered in our study.

from the National Center for Health Statistics vital statistics system as a sampling frame to identify eligible children. Children were sampled using a multistage sampling design such that in the first sampling stage, units were sampled by geographical region, median household income, proportion of minority population, and urban versus non-urban area. In the second stage, units were stratified by race/ethnicity, birth weight, and twin status. Asian Americans, twins, and children with low birth weights were oversampled in order to allow for more precise estimates of these subgroups. Approximately 14,000 parents were initially contacted, and active consent was used to solicit participation in the study. Approximately 10,700 parents of children born in 2001 participated in the first wave of data collection when the children were approximately 9 months of age (please note, according to the reporting rules set forth by National Center for Educational Statistics, all sample sizes must be rounded to the nearest 50). More information about the sampling design can be found in Bethel et al. (2005).

The ECLS-B consisted of four waves of data collection. The first wave of data collection took place when children were approximately 9 months of age, the second wave was collected when children were approximately two years old, the third wave at approximately four years old, and the final wave took place at kindergarten entry. The study used surveys, interviews, observations, and direct assessments to collect multiple indicators about children's health, development, care, and education from the children themselves, their parents, and their child care teachers and directors.

For the purposes of our study, we were interested in the subset of four-year-old children who had attended center-based care at least 10 hours a week during the four-year old assessment, and who had been randomly selected by the ECLS-B team to be part of an in-depth observational study of their child care arrangement. These 1250 children represented the analytic sample for this study.

In our study, children, on average, were approximately 53 months of age, and were enrolled in preschool an average of 24.5 hours per week (standard deviation 12.2 and range 10 to 65). The sample was evenly split between males and females, with nearly two-thirds of children living in families who were classified within the three lowest strata of the socioeconomic composite. With respect to race/ethnicity, 44% of children were Caucasian, 24% were African-American, 17% were Hispanic, 6% were Asian, and the remaining 9% were of another race/ethnicity. Approximately 16% of children came from families whose primary language at home was not English, and 9% were identified as receiving special education services.

With respect to program characteristics, nearly 19% were public school-based prekindergarten classrooms, 16% were community-based prekindergarten classrooms, 12% were preschool classrooms in full-day child care centers, 27% were Head Start classrooms, and 23% were preschool classrooms operating in nursery school programs. In addition, 58% were for profit programs, 82% were licensed, and 51% had been accredited by a local, state, or national organization. Approximately 28% of programs were sponsored by another organization, with the most common sponsoring organizations being public schools (16%), Head Start (9%), and the state or local government (8.5%).

Classroom Quality Measures

Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998)—The ECLS-B data collection team used the ECERS-R to assess the global quality of preschool-aged child care classrooms. The ECERS-R consisted of 43 items, organized into seven subscales described earlier. The ECLS-B team deleted the six items comprising the Parents and Staff subscale because the items were not directly related to children's classroom experiences. The ECLS-B team scored the remaining 37 items on a 7-point Likert scale, with 1.0 representing inadequate quality and 7.0 representing excellent quality. The overall ECERS-R scale score was computed by averaging all the item scores to yield a score between 1.0 and 7.0. The subscale scores were computed in an analogous manner, with the item scores on each subscale averaged to yield a score between 1.0 and 7.0. The internal consistency estimate for the ECERS-R was very high at 0.95 (Snow et al, 2007).

Children's Cognitive and Social-emotional Measures

Cognitive outcomes—The ECLS-B dataset included three measures of developmental ability in language, literacy, and mathematics (see Snow et al., 2007 for a full description of each measure). The *language test* assessed children's receptive language skills by asking children to point to a picture of an orally spoken vocabulary word. The *literacy test* assessed children's phonological awareness, letter-sound knowledge, letter recognition, print conventions, and word recognition. The *mathematics test* was a two-stage, ability-adaptive measure. In the first stage, children took a core set of items assessing relative size/quantity, pattern matching, number recognition, and counting. Children who performed particularly poorly or particularly well on the core set of items received an additional set of items. The poorly performing children took a supplemental test assessing counting and shapes, while the highly-performing children took a supplemental test assessing word problems and number sentences. The internal consistency estimates for the cognitive measures were high, ranging from 0.81 to 0.89 (Najarian et al., 2010).

Social-emotional outcomes—Children's social-emotional skills were assessed in two ways. The first way was through their child care teachers' ratings. The ECLS-B study team administered selected items from the Preschool and Kindergarten Behavior Scales (Merrell, 2003) and Social Skills Rating System (Gresham & Elliott, 1990) to assess key constructs including prosocial skills, temperament, approaches toward learning, and problem behaviors. Using factor analysis, we created the following scales by averaging the responses on each item comprising the scale: (a) a 4-item *approaches towards learning* scale, which assessed children's attitudes towards learning, including their persistence and eagerness to learn ($\alpha = 0.84$); (b) a 6-item *social competence* scale, which measured the extent to which children had friends, were accepted by other children, or were empathetic to others ($\alpha = 0.82$); (c) a 4-item *restlessness* scale, which assessed the extent to which the child was overly active or fidgety ($\alpha = 0.87$); (d) a 2-item *internalizing* scale, which assessed the extent to which children worried or appeared depressed ($\alpha = 0.74$); and (e) a 2-item *externalizing* scale, which assessed the extent to which children had temper tantrums or was physically aggressive ($\alpha = 0.76$).

The second way that children's social-emotional skills were assessed was via direct assessment in which children and parents engaged in the Two Bags Task (Owen, Barfoot, Vaughn, Domingue, & Ware, 1996). The Two Bags Task was a standardized, semi-structured task in which children's and parents' interactions with each other were observed. The task consisted of a 10-minute session in which parents and children were instructed to interact with two bags. The first bag contained a book and the second bag contained materials for play. Parents were instructed to start with the first bag, then move to the second bag when they were ready. Parents' and children's interactions were videotaped and the ECLS-B study team coded their behaviors. Using a 7-point Likert scale, the team coded children's behaviors along three dimensions: quality of play, child engagement, and child negativity. *Quality of play* assessed children's sustained involvement with the play materials, including their attention to the objects, their self-direction, and complexity of play. *Child engagement* assessed the extent to which the child communicated positive regard to the parent, and initiated and maintained interactions with the parent. *Child negativity* assessed the extent that the child showed anger, hostility, or dislike toward the parent. Interrater reliability on the ratings was high, with the average percent agreement at nearly 91% for all of the scales (Najarian et al., 2010).

Model Covariates

Child-level characteristics—The ECLS-B researchers conducted interviews with parents (usually mothers) that yielded information related to the child's age, gender, ethnicity, primary language spoken at home, disability status, and the number of weekly hours their child attended child care. They also gathered information about parents' education level, occupation, and income, and this information was summarized in a five-category, ordinal-scaled socioeconomic composite measure of family social standing (see Najarian et al., 2010 for more information about the socioeconomic composite).

Prior cognitive development—To obtain a measure of cognitive and language development prior to preschool, we used the scores from a cognitive measure administered at two years of age. When children were two years old, trained ECLS-B researchers administered the Bayley Short Form—Research Edition (BMDSF-R; Bayley, 1993) mental development subscale. The BMDSF-R for two-year-olds was a 33-item measure designed to assess early literacy, language, abstract thinking, and problem solving skills. The test had a high internal consistency estimate of 0.94 (Andreassen, Fletcher, & Park, 2007).

Prior social-emotional skills—Child care teachers did not provide ratings of children's social-emotional skills at two years of age. Thus, we relied on the Two Bags Task administered when children were two years as a measure of prior social-emotional skills. The Two Bags Task was similar to that administered when children were four years old, but slightly different child-level scores were generated. Child engagement and child negativity were also assessed at this age, but instead of quality of play, sustained attention was assessed instead. Sustained attention measured the extent that children were focused and involved with the task. Interrater reliability for the scales was high, with the average percent agreement on the codes ranging from 93% to 97%.

Data Collection, Training and Procedures

All ECLS-B data collectors underwent 44 hours of training over a 7-day period in which they received comprehensive training in all aspects of fieldwork, including following general project procedures, administering the different components and instruments, interacting with children, and ensuring quality control. Training consisted of lectures, discussions, interactive activities, practice exercises, and mock interviewers. The data collectors also received refresher training throughout the data collection period.

To ensure reliability of the ECERS-R rating, the data collectors underwent two-day ECERS-R-specific training sessions, which consisted of four practice observations. Raters had to be within one point of the consensus score on at least 80% of the items in order to be considered reliable. Raters also underwent periodic reliability checks during the data collection period to prevent rater drift. A 3-hour ECERS-R observation was conducted once during the academic school year. Prior research on the ECERS-R suggests stable ratings over a ten-week time period (Clifford & Rosbach, 2005; Hofer, 2010).²

Data collectors also received almost 24 hours of training on how to administer the child assessments. During these sessions, they were taught about the behaviors of young children, assessing children with disabilities, building and maintaining rapport with children, and mandatory reporting procedures. Structured role-playing methods were also used, in which trainers acted as study participants and data collectors demonstrated their assessment administration techniques. Once data collectors passed their cognitive assessment certification test, which required a passing score of 75%, they conducted field assessments of children. To assess children, data collectors identified a 2.5-hour block of time during which the child was usually awake during the fall and winter of children's preschool year. They then administered the cognitive measures during a 30–45 minute window. (For more information on data collection and procedures, see Snow et al., 2007).

Analytic Approach

Generalized additive modeling (GAM)—To identify thresholds, we used GAM, a statistical model that facilitates the identification of thresholds by relaxing the assumption that the association between quality and outcomes is linear. Our GAM model took the following functional form:

$$\text{Outcome}_i = \mu + f_1(X_{1i}) + \dots + f_p(X_{pi}) + g(\text{ECERS} - R_i) + \varepsilon_i$$

such that X_{pi} represents the vector of child-level covariates, $f_1 \dots f_p$ and g represent unknown, nonlinear functions that are estimated non-parametrically, and ε_i is a random error term that is normally distributed with common variance. We plotted the raw values of the ECERS-R scores against the “smoothed” values from $g(\text{ECERS} - R_i)$, which resulted in a GAM plot that showed a point-by-point estimate of the association between the ECERS-R and each outcome. To identify thresholds, we looked for inflection points within the GAM plots, which were indicative of changes in the strength of associations between the ECERS-

²The ECERS-R generally lacks information about its test-retest reliability (Zaslow et al., 2009).

R and outcomes. A threshold on the ECERS-R would be denoted on the GAM graph as a flat or a slight negative slope, followed by a sharp positive slope (see Setodji, Le, & Schaack, 2013 for more information on this method).

Piecewise regression—A limitation of the GAM analysis is that it does not identify a specific cut-point as a threshold. Instead investigators need to use their judgment to identify which cut-points on the GAM graph should serve as a threshold. Because investigators can potentially identify a range of different values as thresholds, studies have recommended that investigators use piecewise regression to test whether the slopes within the regions demarcated by the chosen thresholds were significantly different from zero (Setodji, Le, & Schaack, 2013). To identify a minimum threshold of quality that needs to be surpassed before significant associations can be observed, investigators would look for score regions in which the associations between the ECERS-R and outcomes were not significant, followed by score regions where the associations were statistically positive. To identify an asymptotic threshold representing a leveling-off effect, investigators would look for score regions in which the association between the ECERS-R and outcomes were statistically positive, followed by score regions where the associations were no longer significant.

Composite measures—In order to assess the association between the ECERS-R and the different constructs, we created three composite scores. The first measure was a *Cognitive Composite*, which was an aggregate measure of the language, literacy, and mathematics scores. We subjected the cognitive measures to a principal components analysis and used the component loadings as scoring coefficients for the cognitive composite. Loadings for this composite indicated unidimensionality, with loadings ranging from 0.80 to 0.91. We repeated the process for the social-emotional measures, such that we subjected the eight social-emotional scales to a principal components analysis, then used the components loadings to create an aggregate *Social-emotional Composite*. The loadings for this component also supported a single factor, with component loadings ranging from 0.48 to 0.81. Finally, we used principal components analysis on both the cognitive and social-emotional scales to create a single composite that we called the *Developmental Functioning Composite*.

We conducted the GAM and piecewise analysis on each of the composite measures, as well as on the individual cognitive and social-emotional scales comprising the composite measures. This strategy allowed us to determine whether the thresholds derived from the composite measures were similarly identified on the individual scales comprising those measures.

Model details—As noted earlier, our covariates included child-level background characteristics as well as a child's prior cognitive and/or social measure scores collected when they were two years old. To facilitate interpretations, we reverse scored the restlessness, internalizing, externalizing, and child negativity scales so that higher scores represented more favorable outcomes (e.g., higher restlessness scores indicate lower levels of restlessness). We also standardized the outcome variables to have a mean of zero and variance of one. To account for the complex sampling design, we adjusted the standard errors via the Huber-White method (Freedman, 2006).

Results

Table 1 provides the descriptive statistics for the ECERS-R and outcome measures used in our study. The mean ECERS-R score was 4.52, with a standard deviation of approximately 1 point. Using the guidelines adopted by other researchers where scores below 3.00 are considered “poor,” scores between 3.00 and 4.99 are considered “average,” and scores above 5.00 are considered “good” (Burchinal et al., 2000; Helburn et al., 1995), the mean ECERS-R score in our study would be considered average. The mean scores on most of the ECERS-R subscales were also average, except for Interaction and Program Structure, where the mean scores fell within the good range.

Thresholds on the Cognitive and Social-emotional Outcome Measures

Composite measures—Figure 1 provides the GAM plots for the Cognitive and Social-emotional Composites in relation to the ECERS-R. In the plots, the dotted line represents the non-parametric association between the ECERS-R and the outcome measures under a GAM analysis, and the solid line represents the association between the ECERS-R and the outcome measures under a linear regression analysis. Note that the point estimates of the “effect size” axis on this plot represented an adjusted scaled level of the outcome for a particular ECERS-R value that allows for the comparison of one ECERS-R value to another. Thus, a negative value does not imply a negative association between the ECERS-R and child outcomes, but instead denotes a weaker effect.

Despite the modest correlation between the two composites ($r = 0.30$), the GAM plots are surprisingly similar. Both plots suggest a ceiling threshold existed at an ECERS-R score of 3.4, such that the association between the ECERS-R and outcomes appeared to be strongly positive between an ECERS-R range of 1 and 3.4, but flat beyond an ECERS-R score of 3.4. Figure 3 provides the GAM plot for the Developmental Functioning Composite. Not surprisingly, the GAM plot for the single combined Developmental Functioning Composite mirrored the plots observed for the Cognitive Composite and the Social-emotional Composite (see Figure 2), such that there was a threshold effect at an ECERS-R score of 3.4. Approximately, 14.5% of the children in our sample were enrolled in centers with an ECERS-R score lower than 3.4.

Table 2 provides the results of the linear regression and piecewise regression analyses. Although the linear regression analyses suggested there were no associations between any of the composite scores and the ECERS-R, the piecewise regression identified significant associations, at least within specific score ranges. Namely, between the ECERS-R score range of 1 and 3.4, there were significant increasing associations between the ECERS-R and the Social-emotional and Developmental Functioning Composites, but outside of this score range, the associations were null.

Cognitive outcomes—Looking specifically at the measures that comprised the Cognitive Composite, only the mathematics measure showed a significant association with the ECERS-R between an ECERS-R score range of 1 and 3.4; the language arts and literacy measures did not show a significant association with the ECERS-R within this score range. However, visual inspection of the GAM plot for the literacy scores alone (see Appendix A)

indicated that the association between literacy performance and the ECERS-R was strongest within the region of 1 and 3.4, but then leveled off after an ECERS-R score of 3.4 was reached. Thus, although not statistically significant, the literacy measure showed suggestive evidence supporting the notion that an ECERS-R score of 3.4 represented a ceiling threshold.

Social-emotional outcomes—With respect to the individual social-emotional measures comprising the Social-emotional Composite, the regression results indicated that the majority of the measures showed threshold effects at a score of 3.4. Namely, the approaches towards learning, restlessness, externalizing, negativity, and engagement scales showed significant associations with the ECERS-R between the score range of 1 and 3.4, and then a leveling off effect thereafter.

Associations with the ECERS-R Subscales

We conducted additional analysis to determine whether the leveling-off effect that was observed on the total ECERS-R score would be similarly observed on the ECERS-R subscales, or on the individual achievement and social-emotional outcome measures that comprised the Developmental Composite. Although the vast majority of subscale-outcome combinations showed a leveling-off effect at the highest score levels (e.g., a score of 5.0 or 5.5), there were some specific subscales and outcomes that did not show a leveling-off effect. Figure 3 shows the GAM plots for the cognitive outcomes that did not show a leveling-off effect, and Figure 4 shows the corresponding plots for the social-emotional outcomes. As shown in Figure 3, depending on the specific subscale and outcome under consideration, at a subscale score of approximately 5.0 or 5.5, we observed a surge in the strength of association between the Interaction, Learning Activities, or Program Structure subscales and the receptive vocabulary and/or literacy measures. For the social-emotional outcomes (see Figure 4), there were positive, linear associations between the internalizing and restlessness outcomes with Program Structure, and between externalizing behaviors with Learning Activities. (Recall that the internalizing, externalizing, and restlessness scores have been reverse scored). Taken together, the results suggest that there may not necessarily be diminishing associations between classroom quality and outcomes, particularly when the classroom quality measures emphasize the teaching and caregiving processes and activities that more directly promote children's cognitive or social-emotional development.

Discussion

As part of their efforts to improve the quality of ECE in the United States, QRIS developers are currently grappling with where to set the thresholds or cut-points on the ECERS-R. Professional recommendations suggest that the validity of the thresholds be examined to determine whether the thresholds are empirically functioning in the manner as intended (AERA, APA, & NCME, 2014). Well-functioning thresholds should either denote the minimum level of quality that needs to be exceeded before significant associations with child outcomes are observed, or the asymptotic point at which the association between quality and outcomes level off (Burchinal et al., 2010). However, the literature is largely absent about the empirical functioning of specific cut-points on the ECERS-R (see Le, et al.,

2015 for an exception). Ideally, these thresholds should be consistently identified across different constructs. For example, a cut-point that has been identified as a leveling-off threshold on cognitive outcomes but not similarly identified on the social outcomes has less utility than a cut-point that serves as a leveling-off threshold on both the cognitive and social outcomes. To date, few studies have compared whether the same threshold appears on both cognitive and social outcomes.

ECERS-R Associations with Child Outcomes

Using a representative sample of children and classrooms across state policy contexts, we found that there was a threshold at an ECERS-R score of 3.4, such that there was a statistically significant positive association between the score range of 1.0 and 3.4, and a null association outside of this range with children's concurrent outcomes. We note that the leveling-off effect at a score of 3.4 observed in this study is lower than the leveling-off effect observed at a score of 4.6 in the Le et al. (2015) study. The lower threshold found in this study may be due to differences in the policy context. The Le et al. (2015) study examined data only from Colorado, which was one of the first states in the nation to develop and implement a QRIS. Classrooms in the sample had been participating in quality improvement activities for several years prior to data collection. By way of contrast, the present study included nationally-representative data from all states nationwide, few of which had a QRIS that had been in place as long as Colorado's QRIS. This may explain why the mean ECERS-R score in the Le et al. study (2015) was nearly a point higher than the average ECERS-R rating observed in the present study. Because higher scores were more common in the Le et al. (2015) study, it is not surprising that the leveling-off effect was observed at a higher score point as well.

In the present study, it is possible that a threshold was observed at a score of 3.4 because of the nature of the ECERS-R content. Between a score range of 1.0 to 3.0, the ECERS-R focuses on either the absence of positive teacher behaviors, negative teacher behaviors, or environmental constructs that may constrain learning and development. Once a score of 3.0 is reached, items begin to focus more on the presence of factors in the classroom environment that support learning and development. Thus, a score of 3.4 represents a transitional score from "poor" quality to "average" quality. Notably, the cut-score of 3.4 was observed across the Cognitive Composite as well as the Social-emotional Composite. We also found that the ECERS-R was generally unrelated to child outcomes when linear models were used, but significant associations were observed when nonlinear models were used.

Our study suggests that the ECERS-R might no longer be useful for assessing the association between quality and child outcomes beyond a mid-level score of 3.4. Once a score of 3.4 is reached, the ECERS-R as a composite measure becomes less effective at distinguishing between children's outcomes, as an "average" score on the ECERS-R has the same predicted association with children's outcomes as a "good" score. However, this result does not necessarily imply that scores in the "good" range are unrelated to child outcomes, as our study assessed only a subset of important child developmental outcomes, and there are other child outcomes (e.g., persistence, peer relationships, health outcomes, or general knowledge) that may show significant associations with higher ECERS-R scores.

Because higher quality is expected to be associated with better outcomes (Vandell, 2004), the finding that there is a leveling-off effect at a mid-level rating of 3.4 points to a possible limitation of the ECERS-R for high-stakes contexts. As a global measure of quality, the ECERS-R may not be optimally measuring the seven main components of early childhood environments articulated by the developers (Perlman et al., 2004). It may comprehensively evaluate physical environmental features, but provides less information regarding effective teaching practices (Cassidy et al., 2005). As noted by Douglas (2004), the ECERS-R is limited in its focus on the quality of teacher-child and peer interactions in the classroom and on the quality of instructional support provided to children across learning domains. Thus, the finding that scores above a mid-level rating of 3.4 are not related to children's outcomes may be a reflection of the content of the ECERS-R, and its overall emphasis on environmental quality (Cassidy et al., 2005). In other words, the over-emphasis on the physical environment in the overall ECERS-R score may be obfuscating associations that exist between the process-oriented aspects of the scale and child outcomes.

Leveling-off Effects and the ECERS-R Subscales

For the more process-oriented ECERS-R subscales, instead of a leveling-off effect, we observed a linear association between the subscales and child outcomes, or a surge in the magnitude of association between the subscales and outcomes at a score of 5.0 or 5.5 (see Figures 3 and 4). For example, given the process-oriented nature of the Learning Activities subscale, it is not surprising to observe surges in the magnitude of association with receptive vocabulary and with literacy after a score of 5.0, or to observe positive linear associations with social outcomes such as externalizing behaviors. Similarly, for the Program Structure subscale, which was constructed so that items on the higher end of the score distribution are focused on teacher behaviors and the quality of their interactions with children (Child Care Resources, 2006), we observed a surge in the strength of the association with receptive vocabulary at a score of 5.5, and no leveling-off effects with internalizing behaviors and restlessness. The fact that a positive linear association or late surge continued to be observed on a subset of items measuring aspects of process quality is consistent with the notion that the ECERS-R may not have had a sufficient number of items that capture educational practices that are most strongly related to children's developmental outcomes. Thus, it is possible the limited focus of the ECERS-R on teaching processes and practices accounted for the leveling-off effects for the overall ECERS-R as well as with most of its subscales. This also means that the usage of a summary assessment tool such as the ECERS-R versus more targeted tools (e.g., its subscales) is a critical issue in classroom-related assessment. For these reasons, researchers have developed the ECERS-Extension, or ECERS-E (Sylva, Siraj-Blatchford, & Taggart, 2003) as a more cognitively-oriented supplement to the ECERS-R to correspond to a newly-adopted preschool curriculum in England.

Notably, despite being process-oriented subscales, Interaction and Language-Reasoning showed leveling-off effects with almost all the cognitive and social outcomes included in our study. This may be a function of the number of items found on these scales. For example, the Language-Reasoning and Interaction subscales are comprised of four and five items, respectively, whereas the Learning Activities subscale was comprised of 10 items. Thus, the Learning Activities subscale may have had a greater number of items that assessed teacher

caregiving and instructional processes and was therefore more sensitive at capturing the specific care and instructional practices that are proximally related to children's outcomes.

Findings from a study conducted by Gordon and colleagues (2013) offer another explanation that may contribute to understanding the leveling-off effect that was observed in our study. Using the ECLS-B dataset, the authors found evidence of raters not adhering to the developers' rating instructions due to item multidimensionality, in which the level of quality needed for classrooms to earn a score of 6, for example, could be less than the level of quality needed to earn a score of 5. Due to differences in the nature of their content, it is possible that rating category disorder occurred more frequently for the Interaction and Language-Reasoning subscales than for the Learning Activities subscale, which is why there was less evidence of leveling off effects on the Learning Activities subscale than on the Interaction and Language-Reasoning subscales. At the lower score levels of the Learning Activities subscale, raters are asked to assess whether children have access to certain learning materials. By contrast, at the lower score levels on the Interaction and Language-Reasoning subscales, raters are asked to assess the quality of interactions, which require a greater degree of inference and interpretation. It is possible that the higher level of subjectivity on these latter subscales resulted in greater category disorder. We note that this explanation is speculative, and future studies should examine the frequency of rater category disorder by the aspect of quality being assessed.

It is also possible that as the ECERS-R has become commonplace in most ECE classrooms that teachers have learned the "rules" of the ECERS-R and are structuring their classrooms in ways that symbolically comport to it (Tarrant & Huerta 2015). In other words, it is possible that in the 1990s, when the ECERS-R was not a common classroom and policy tool, the environmental quality of a classroom could serve as a reasonable proxy for broader developmentally appropriate care and instructional practices. Thus, associations between quality, as measured by the ECERS-R, and child outcomes may have been more robust. With its inclusion in most states' QRIS, the ECERS-R has become a mainstream tool, and most classrooms are comporting their physical environment to ECERS-R standards. As quality has improved over time, associations between the ECERS-R and children's cognitive and social-emotional outcomes could be changing as well.

Research Implications

Our study suggests that the ECERS-R shows positive associations to child outcomes up to a mid-level score of 3.4, beyond which it may lose its usefulness at assessing quality in relation to those outcomes. It may be necessary to use a more process-oriented measure once a threshold of 3.4 is met, as the ECERS-R may not be effective at capturing differences in relation to child outcomes for classrooms at the higher end of the score distribution. Using multiple measures of quality is consistent with the practice of some states' QRIS, which include both the ECERS-R and a more process-oriented measure (such as the CLASS) in their rating systems (QRIS Compendium, n.d.). Future research should explore ways in which the ECERS-R can be combined with a more process-oriented quality measure to provide a more comprehensive picture of quality in a classroom.

Another avenue for future research is the use of nonlinear models or quantile regressions as means to explore associations between quality measures and child outcomes. The literature has frequently reported weak or null associations between quality and child outcomes, and the results from our linear models were consistent with this body of literature. However, GAM provided a more nuanced picture, revealing significant associations between quality and child outcomes in specific score regions. Indeed, Torquati, Raikes, Welch, Ryoo, and Tu (2012) found that many associations between quality and children's developmental outcomes are nonlinear. They examined the association between the ECERS-R and 17 social-emotional and cognitive outcomes. Of the 11 outcomes that showed significant associations with the ECERS-R, only two outcomes showed significant linear associations; the remaining nine outcomes showed significant nonlinear associations. Thus, GAM may not only help QRIS developers understand the empirical functioning of their chosen quality cut-points, it may also help to uncover statistically significant and meaningful associations that are obscured by linear models. Overall, we recommend QRIS developers to continue to explore the validity of the thresholds adopted in their accountability systems through GAM as well as other statistical means.

Study Limitations

Our study has several limitations. First, the ECLS-B data used in this study was not designed as a randomized control trial to assess the impact of the ECERS-R on children's social, cognitive, and language outcomes, and therefore, the results cannot support causal inferences about the thresholds and relationships described in this study. While we have attempted to minimize any self-selection bias by including measures of prior cognitive development and social-emotional skills and other measures of child demographics, the results may nonetheless reflect non-random sorting of children into classrooms. Second, the use of a complex modeling tool such as GAM in policy contexts has its limitations. Because GAM requires visual inspection, there can be differences in the interpretations of the GAM plots. Although cut points identified by visual inspection were validated empirically through other means (such as piecewise regression), there is still the possibility that the GAM graphs will support different but equally viable alternative thresholds (Setodji, Le, & Schaack, 2013). Although these thresholds may lead to slightly different quality classifications of classrooms, it is important to recognize that any cut point will introduce artificial variation between centers that are on the cusp of two different levels. Finally, our study examined a limited set of cognitive and social outcomes. Although the results were robust across both set of outcomes, it is unclear whether the same types of relationships would be observed with other types of outcomes, such as those relating to executive functioning or metacognition. Future studies should explore whether our findings can be replicated with a wider range of outcomes using other independent datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Administration for Children and Families. Use of ERS and other program assessment tools in QRIS. Washington, DC: Authors; 2013. Available at https://qrisguide.acf.hhs.gov/files/QRIS_Program_Assess.pdf
- American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. The standards for educational and psychological testing. Washington, DC: Authors; 2014.
- Andreassen, C., Fletcher, P. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) psychometric report for the 2-year data collection (NCES 2006–045). Washington DC: National Center for Education Statistics; 2006.
- Arnett J. Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*. 1989; 10(4):541–552. [http://dx.doi.org/10.1016/0193-3973\(89\)90026-9](http://dx.doi.org/10.1016/0193-3973(89)90026-9).
- Bethel, J., Green, J.L., Nord, C., Kalton, G. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) methodology report for the 9-month data collection (2001–02) (NCES 2005–147). U.S. Department of Education. Washington, DC: National Center for Education Statistics; 2005.
- Bayley, N. Bayley Scales of Infant Development. 2. San Antonio, TX: Psychological Corporation; 1993.
- Burchinal MR, Peisner-Feinberg ES, Bryant DM, Clifford RM. Children's social and cognitive development and child care quality: Testing for differential associations related to poverty, gender, or ethnicity. *Applied Developmental Science*. 2000; 4(3):149–165.
- Burchinal MR, Vandergrift N, Pianta R, Mashburn A. Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*. 2010; 25:166–176.
- Cassidy DJ, Hestenes LL, Hegde A, Hestenes S, Mims S. Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly*. 2005; 20(3):345–360.
- Child Care Resources. ITERS-R and ECERS-R: Breaking down the subscales. Monroe, North Carolina: Author; 2006.
- Clifford, R., Rosbach, HG. Structure and stability of the Early Childhood Environment Rating Scale. In: Schonfeld, H.O'Brien, S., Walsh, T., editors. *Questions of quality*. Dublin, Ireland: The Centre for Early Childhood Development & Education, The Gate Lodge, St. Patrick's College; 2005. p. 12-21.
- Douglas, F. A critique of ECERS as a measure of quality in early childhood education and care. In: Schoenfeld, H.O'Brien, S., Walsh, T., editors. *Questions of quality*. Dublin Ireland: St. Patrick's College; 2004.
- Early D, Bryant D, Pianta R, Clifford R, Burchinal M, Ritchie S, Barbarin O. Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly*. 2006; 21:174–195.
- Frank Porter Graham Child Development Institute. Early developments. The University of North Carolina at Chapel Hill; Chapel Hill, NC: 2003.
- Freedman DA. On the so-called Huber sandwich estimator and robust standard errors. *The American Statistician*. 2006; 60(4):299–302.
- Gordon RA, Fujimoto K, Kaestner R, Korenman S, Abner K. 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*. 2013; 49(1):146–160. [PubMed: 22468567]
- Gresham, FM., Elliot, SN. Social Skills Rating System (SSRS). Circle Pines, MN: American Guidance Service; 1990.
- Hamilton, D., Bates, J., Mitchell, A., Workman, S. QRIS incentives and program quality: A Discussion of state approaches and the effectiveness of incentives. Paper presented at the annual meeting of the QRIS National Meeting; National Harbor, MD. 2015.
- Harms, T., Clifford, R., Cryer, D. The Early Childhood Rating Scale - Revised. New York, NY: Teachers College Press; 1980; 1998.

- Helburn, SW., Culkin, ML., Morris, JR., Mocan, NC., Howes, C., Phillipsen, LC., Bryant, DM., Clifford, RM., Cryer, D., Peisner-Feinberg, ES., Burchinal, M., Kagan, SL., Rustici, J. Cost, quality, and child outcomes in child care centers. Denver, CO: Department of Economics, Center for Research in Economic and Social Policy, University of Colorado, Denver; 1995.
- Hofer KG. How measurement characteristics can affect ECERS-R scores and program funding. *Contemporary Issues in Early Childhood*. 2010; 11(2):175–191.
- Holloway S, Kagan S, Fuller B, Tsou L, Carroll J. Assessing child-care quality with a telephone interview. *Early Childhood Research Quarterly*. 2001; 16:165–189.
- Howes C, Phillips D, Whitebook M. Thresholds of quality: Implications for the social development of children in center based care. *Child Development*. 1992; 63:449–60. [PubMed: 1611946]
- Kagan, SL., Cohen, N. Solving the quality crisis: A vision for America's child care system. Yale University Bush Center; New Haven: 1997.
- Le V, Schaack D, Setodji CM. Identifying baseline and ceiling thresholds within the Qualistar Early Learning Quality Rating and Improvement System. *Early Childhood Research Quarterly*. 2015; 30(B):215–226. [PubMed: 25530668]
- Loeb S, Fuller B, Kagan SL, Carrol B, Carroll J. Child care in poor communities: Early learning effects of type, quality, and stability. *Child Development*. 2004; 75(1):47–65. [PubMed: 15015674]
- Mashburn AJ, Pianta R, Hamre BK, Downer JT, Barbarin O, Bryant D, Burchinal M, Clifford R, Early D, Howes C. Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development*. 2008; 79:732–749. [PubMed: 18489424]
- Merrell, KW. *Preschool and Kindergarten Behavior Scales*. 2. Boston, MA: Houghton Mifflin Harcourt; 2003.
- Mitchell, A. Financial incentives in quality rating and improvement systems: Approaches and effects. Washington, DC: Alliance for Early Childhood Finance; 2012.
- Najarian, M., Snow, K., Lennon, J., Kinsey, S. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool–kindergarten 2007 psychometric report (NCES 2010–009). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education; 2010.
- National Center on Early Childhood Quality Assurance. Financial incentives in QRIS. Washington, DC: Author; 2017.
- NICHD Early Child Care Research Network. Child care structure→process→outcome: Direct and indirect effects of child care quality on young children's development. *Psychological Science*. 2002; 13:199–206. [PubMed: 12009038]
- Owen MT, Barfoot B, Vaughn A, Domingue G, Ware AM. 54-month parent-child structured interaction qualitative rating scales. NICHD Study of Early Child Care Research Consortium. 1996
- Peisner-Feinberg E, Burchinal M. Relations between preschool children's child-care experiences and concurrent development: The cost, quality, and outcomes study. *Merrill-Palmer Quarterly*. 1997; 43:451–477. 1997.
- Peisner-Feinberg ES, Burchinal MR, Clifford RM, Culkin ML, Howes C, Kagan SL, Yazejian N. The relation of preschool quality to children's cognitive and social developmental trajectories through second grade. *Child Development*. 2001; 72(5):1534–1553. [PubMed: 11699686]
- Perlman M, Zellman GL, Le V. Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly*. 2004; 19:398–412.
- QRIS Compendium. QRIS Compendium State Profile Reports. (n.d.). Available at <http://qriscompendium.org/profile-report>
- Sabol TJ, Pianta RC. Do standard measures of preschool quality used in statewide policy predict school readiness? *Education, Finance and Policy*. 2014:116–164.
- Sakai LM, Whitebook M, Wishard A, Howes C. Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing differences between the first and revised editions. *Early Childhood Research Quality*. 2003; 18(4):427–445.

- Schaack, D., Tarrant, T., Boller, K., Tout, K. Quality rating and improvement systems: Alternative approaches to understanding their impact on the early learning system. In: Kagan, SL., Kaurez, K., editors. *Early childhood systems: Transforming early learning*. New York, NY: Teacher's College Press; 2012. p. 71-86.
- Setodji CM, Le V, Schaack D. Using generalized additive modeling to empirically identify thresholds within the ITERS in relation to toddlers' cognitive development. *Developmental Psychology*. 2013; 49(4):632–645. [PubMed: 22642314]
- Smith, M., Dickinson, D. *Early language & literacy classroom observation*. Baltimore, MD: Brookes; 2002.
- Snow, K., Thalji, L., Derecho, A., Wheelless, S., Lennon, J., Kinsey, S., Rogers, J., Raspa, M., Park, J. *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), preschool year data file user's manual (2005–06) (NCES 2008–024)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education; 2007.
- Stevenson-Garcia, J., Brenneman, K., Frede, E., Weber, M. *Preschool rating instrument for science and mathematics (PRISM)*. New Brunswick, NJ: National Institute for Early Education Research; 2010.
- Sylva, K., Siraj-Blatchford, I., Taggart, B. *Assessing quality in the early years: Early Childhood Environment Rating Scale-Extension (ECERS-E): Four curricular subscales*. Stoke-on Trent: Trentham Books; 2003.
- Sylva K, Siraj-Blatchford I, Taggart B, Sammons P, Melhuish E, Elliot K, et al. Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*. 2006; 21:76–92.
- Tarrant K, Huerta L. Substantive or symbolic stars: Quality rating and improvement systems from a new institutional lens. *Early Childhood Research Quarterly*. 2015; 30(1):327–338.
- Torquati, J., Raikes, H., Welch, G., Ryoo, JH., Tu, X. Testing thresholds of child care quality on child outcomes. Paper presented at the Nebraska Center for Research on Children, Youth, Families, and Schools; Lincoln, NE. 2012.
- Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., Boller, K. Prepared for the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Washington, DC: Child Trends; 2010. *Compendium of quality rating systems and evaluations*.
- U.S. Department of Education. Race-to-the-Top Early Learning Challenge grants. 2016. Retrieve from <http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/index.html>
- Vandell D. Early child care: The known and the unknown. *Merrill-Palmer Quarterly*. 2004; 50:387–414.
- Vermeer HJ, van IJzendoorn MH, Cárcamo RA, et al. Quality of child care using the environment rating scales: a meta-analysis of international studies. *International Journal of Early Childhood*. 2016; 48(1):33–60. DOI: 10.1007/s13158-015-0154-9
- Whitebook, M., Howes, C., Phillips, D. *Who cares? Child care teachers and the quality of care in America*. Oakland, CA: Child Care Employee Project; 1990.
- Zaslow, M., Forry, N., Weinstein, D., Nuenning, M., McSwiggan, M., Durham, M. *Selected observational measures for assessing the quality of early childhood classrooms: An annotated bibliography*. Washington, DC: Institute of Education Sciences; 2009.
- Zellman, G., Perlman, M. *Child-care quality rating and improvement systems in five pioneer states: Implementation issues and lessons learned*. Santa Monica, CA: RAND Corporation; 2008.
- Zellman, GL., Perlman, M., Le, V., Setodji, CM. Assessing the validity of the Qualistar Early Learning quality rating and improvement system as a tool for improving child-care quality. RAND Corporation; Santa Monica, CA: 2008.

Highlights

- Linearity assumptions between ECERS-R and child outcomes may not be accurate
- The ECERS-R is related to child outcomes only within the score range of 1 and 3.4
- The same ECERS-R threshold is observed across both social and cognitive outcomes

a- Relationship between ECERS-R and cognitive outcome
between ECERS-R and social outcome

b- Relationship

Figure 1.

Graphical Relationship between the ECERS-R and the Cognitive and Social-emotional Composites under GAM and Linear Regression Analyses

Figure 2.
Graphical Relationship between the ECERS-R and the Developmental Functioning Composite under GAM and Linear Regression Analyses

Figure 3.
Graphical Representation between the ECERS-R Subscales and Selected Achievement Outcome Measures under GAM and Linear Regression Analyses

a- Relationship between Program Structure and Internalizing
Program Structure and Restlessness

b- Relationship between

c- Relationship between Learning Activities and Externalizing

Figure 4.

Graphical Representation between the ECERS-R Subscales and Selected Social-emotional Outcome Measures under GAM and Linear Regression Analyses

Table 1

Descriptive Statistics for the Outcome Measures and ECERS-R subscales

Outcome	N	Mean	StdDev	Range
Classroom Quality				
ECERS-R (total)	1250	4.52	1.03	1.25–6.97
Furnishing and display	1250	4.71	1.06	1.38–7.00
Interaction	1250	5.47	1.47	1.00–7.00
Learning activities	1250	3.95	1.12	1.00–7.00
Listening and talking	1250	4.96	1.33	1.00–7.00
Personal care routines	1250	3.87	1.50	1.00–7.00
Program structure	1250	5.00	1.50	1.00–7.00
Cognitive Outcomes				
Language	1250	8.59	1.90	4.65–13.63
Literacy	1250	13.53	7.11	5.46–34.68
Mathematics	1250	22.66	7.32	5.57–41.55
Socio-emotional Outcomes				
Teachers' ratings				
Approaches toward learning	1100	3.74	0.81	1.00–5.00
Social competence	1100	3.67	0.67	1.33–5.00
Restlessness ^a	1100	3.65	0.97	1.00–5.00
Internalizing ^a	1100	4.00	0.78	1.00–5.00
Externalizing ^a	1100	4.11	0.94	1.00–5.00
Two Bags Task				
Quality of play	1000	4.12	0.87	2.00–7.00
Negativity ^a	1000	4.45	0.91	2.00–7.00
Engagement	1000	6.65	0.77	1.00–7.00

^a denotes the scale was reverse scored.

Table 2

Regression Coefficients and Standard Errors (in Parenthesis) for the Linear and Piecewise Regression Models

Outcome	N	Linear regression slopes	Piecewise regression slopes	
			1 slope < 3.4	3.4 slope 7.0
Developmental Composite	1000	0.0343 (0.0339)	0.3252 (0.1220) **	-0.0364 (0.0378)
Cognitive Composite	1250	0.0104 (0.0225)	0.1142 (0.0788)	-0.0137 (0.0289)
Language	1250	0.0458 (0.0265)	0.0353 (0.1013)	0.0482 (0.0306)
Literacy	1250	-0.0081 (0.0258)	0.0866 (0.1105)	-0.0300 (0.0329)
Mathematics	1250	-0.0153 (0.0240)	0.1483 (0.0688) *	-0.0532 (0.0307)
Socio-emotional Composite	1000	0.0409 (0.0353)	0.3058 (0.1393) *	-0.0227 (0.0403)
Teachers' ratings				
Approaches towards learning	1100	0.0343 (0.0277)	0.2452 (0.1121) *	-0.0159 (0.0359)
Social competence	1100	0.0103 (0.0268)	0.0829 (0.1514)	-0.0070 (0.0369)
Restlessness ^a	1100	0.0409 (0.0323)	0.2997 (0.1312) *	-0.0207 (0.0363)
Internalizing ^a	1100	-0.0509 (0.0323)	0.0388 (0.1132)	-0.0722 (0.0422)
Externalizing ^a	1100	0.0837 (0.0299) **	0.3037 (0.1127) **	0.0314 (0.0348)
Two Bags Task				
Quality of play	1000	-0.0149 (0.0279)	0.0739 (0.1219)	-0.0362 (0.0408)
Negativity ^a	1000	0.0016 (0.0352)	0.3035 (0.1190) *	-0.0655 (0.0425)
Engagement	1000	0.0078 (0.0275)	0.1884 (0.0835) *	-0.0356 (0.0426)

Notes.

* denotes significance at 0.05 level.

** denotes significance at 0.01 level.

^a denotes the scale was reverse scored.