Published in final edited form as:

Dev Psychol. 2013 January ; 49(1): 146–160. doi:10.1037/a0027899.

An Assessment of the Validity of the ECERS-R with Implications for Assessments of Child Care Quality and its Relation to Child Development

Rachel A. Gordon,

Department of Sociology and Institute of Government and Public Affairs, University of Illinois at Chicago

Ken Fujimoto,

Department of Educational Psychology, University of Illinois at Chicago

Robert Kaestner,

Department of Economics and Institute of Government and Public Affairs, University of Illinois at Chicago

Sanders Korenman, and

School of Public Affairs, Baruch College/CUNY

Kristin Abner

Department of Sociology and Institute of Government and Public Affairs, University of Illinois at Chicago

Abstract

The Early Childhood Environment Rating Scale-Revised (ECERS-R) is widely used to associate child care quality with child development but its validity for this purpose is not well established. We examined the validity of the ECERS-R using the multidimensional Rasch Partial Credit Model (PCM), factor analyses, and regression analyses with data from the Early Childhood Longitudinal Study-Birth Cohort. The PCM identified rating category disordering, indicating previously unrecognized problems with the scale's response process validity. Factor analyses identified neither a single factor nor the ECERS-R six subscales, replicating prior research regarding the scale's structural validity. Criterion validity results were mixed, with small effect sizes for regressions predicting child outcomes and moderate effect sizes for regressions predicting teacher-reported quality. Our results lend empirical support to recent critiques of the ECERS-R; and, we discuss implications for its future use and for the development of future measures.

Keywords

child care; child care quality; validity

Since the release of the Cost, Quality and Outcomes Study (Helburn et al., 1995), a consensus has emerged among child development experts that child care quality in the United States is “mediocre.” This lukewarm judgment of child care quality derives mainly from assessments based on the Early Childhood Environment Rating Scale (ECERS) and the Early Childhood Environment Rating Scale-Revised (ECERS-R; Cryer, Harms, & Riley, 2003; Harms, Clifford, & Cryer, 1998). The ECERS and ECERS-R have also played an

important role in documenting the positive, but notably small, associations between child care quality and child development (Early et al., 2006; Kontos, 1991; Mashburn et al., 2008; McCartney, 1984; Montes, Hightower, Brugger, & Moustafa, 2005; Peisner-Feinberg et al., 1999; Sylva et al., 2006; Zellman, Perlman, Le, & Setodji, 2008; Zill et al., 2003).

Scholars and practitioners have begun to raise concerns about the ECERS-R instrument, especially for high-stakes research and policy purposes (Emlen, 2010; Hofer, 2008, 2010; Layzer & Goodson, 2006; Perlman, Zellman, & Le, 2004). The ECERS-R instrument was “based on a checklist of items for improving the quality of environments in early childhood classrooms that Harms (one of the instrument creators) had compiled during nearly 20 years of teaching and observation” (Frank Porter Graham Child Development Institute, 2003, p. 9). The scale was first published in 1980, and a revised version was published in 1998 (Harms & Clifford, 1980; Harms, Clifford, & Cryer, 1998). The ECERS and ECERS-R reflect the early childhood education field’s concept of developmentally appropriate practice, which includes a predominance of child-initiated activities selected from a wide array of options; a “whole child” approach that integrates physical, emotional, social and cognitive development; and, highly trained teachers who facilitate development by being responsive to children’s age-related and individual needs (Bryant, Clifford, & Peisner, 1991; Copple & Bredekamp, 2009; Cryer, 1999; Harms, Clifford, & Cryer, 1998).

There is surprisingly little empirical evidence of the validity of the ECERS-R instrument to support its widespread use in research and policy contexts. Particularly lacking are studies that investigate the psychometric properties of the ECERS-R instrument using item response models. As we demonstrate in this article, item response theory models can provide more precise information about whether the response structure performs as intended than is true of classical test theory approaches. For example, factor analyses assume items are on ordinal or interval scales; item response theory approaches can be used to test this assumption. Factor analytic studies of the dimensionality of the ECERS-R, several of which failed to confirm the multi-dimensional structure specified by the ECERS-R developers (see Table 1), have also been limited to particular geographic areas and/or types of child care centers (Burchinal et al., 2008; Cassidy et al., 2005; Clifford et al., 2005; Early et al., 2006; Frede et al., 2007; Fuller, Kagan, Loeb, & Chang, 2004; Hofer, 2008; Holloway, Kagan, Fuller, Tsou, & Carroll, 2001; Howes et al. 2008; Perlman, Zellman, & Le, 2004; Sakai, Whitebook, Wishard, & Howes, 2003). The absence of a larger base of evidence establishing the validity of the ECERS-R is significant because of the important role that the ECERS-R has played in both research and policy. Judgments about the quality of child care and its influence on child development assume that the ECERS-R is valid for developmental research. If this is not the case, then the consensus about the mediocre quality of child care and the influence of quality of care on child development may need revision.

In this article, we assess three aspects of the validity of the ECERS-R: response process validity, structural validity, and criterion validity. In keeping with the latest views that validity is specific to the use of a measure, where possible we consider each aspect of validity in relation to three potential uses of the ECERS-R: for developmental research, for program improvement, and for subsidy policymaking (Joint Committee on Standards for Educational and Psychological Testing, 1999). As far as we are aware, ours is the first study to assess the ECERS-R using item response models for a sample of children in the U.S. We also conduct factor analyses of structural validity and regression analyses of criterion validity of the ECERS-R, as in some previous studies, but for a large and diverse (e.g., by type, funding, and geography) group of centers in which a nationally-representative sample of preschoolers were enrolled.

Method

Data

We used data from the 4-year-old follow-up of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). Children in the ECLS-B were selected from birth records in 46 states in 2001. The sample size was 10,700, resulting from a response rate of 74% at the initial 9-month interview. Fifty-one percent of these study children were boys; 54% were non-Hispanic White, 26% were Hispanic, 14% were non-Hispanic African American, 3% were Asian/Pacific-Islander, and 4% were of other race-ethnicities. At the 4-year-old follow-up interview, the sample size was 8,950, reflecting the exclusion of children who had died or had moved permanently out of the country, as well as those who could not be located, refused to participate, or lived more than 150 miles from the nearest interviewer (U.S. Department of Education, National Center for Education Statistics, 2001).

The ECLS-B investigators selected a sample of child care settings from 4-year-olds in care 10 or more hours per week based on type of care (home-based, Head Start, and non-Head Start centers) and child poverty (Snow et al., 2007, p. 127); all 4-year-olds whose care setting had been observed at two years and were still in care for at least 10 hours at four years were included. Approximately 65% of sampled center-based providers participated (Snow et al., 2007, p. 307). The ECLS-B statisticians created sampling weights that adjust for the initial sampling design (e.g., oversampling of American Indians and Asians, low birth weight, and twin births), family non-response initially and over time, and provider non-response (Snow et al., 2007).

Our analytic sample contains 1,350 centers and preschools. Seventeen percent are Head Start programs. An additional 58% are school- or church-based preschools (25% located in public schools, 21% in private schools, and 12% in religious schools or churches), and an additional 24% are classrooms serving preschool-aged children in other community centers (13% for-profit and 11% non-profit). The centers represent geographic areas across the U.S.; forty-two percent are located in the South, 20% in the Northeast, 20% in the Midwest and 18% in the West. Most are in urban areas (with 50,000 or more residents; 73%), although 12% are in urban clusters (with less than 50,000 residents) and 15% are located in rural areas. About two-fifths (41%) are located in ZIP Codes where less than 10% of young children (under age 5) are poor, about one-fifth (20%) are located in ZIP Codes where 10-19% of young children are poor, and the remaining two-fifths (39%) are located in ZIP Codes where 20% or more of children are poor.

In sum, the ECLS-B sample, after adjustment for sampling methodology, is nationally representative of 4-year-olds who were in child care 10 or more hours per week in 2005-2006. This sample is advantageous for our study because it allows us to replicate recent correlational and factor analytic studies of the ECERS-R for subgroups that have been examined in prior studies but with different protocols, allows us to extend these analyses to understudied subgroups (e.g., community-based centers), and provides the large sample size needed for item response theory models.

Measures

Descriptive statistics for all measures are in the Appendix.

ECERS-R subscores and total scores—Observers completed the ECERS-R as part of the child care observation that the Research Triangle Institute conducted for the ECLS-B (RTI; Snow et al., 2007, p. 250-261). Observers met two of three criteria: (1) experienced working with young children or in child care, (2) held a bachelor's degree in early childhood

or a related field, or (3) experienced working on research studies that required observations or that involved child care or schools. The ECERS-R trainer had originally been trained at, and subsequently conducted trainings for, the Frank Porter Graham Institute, where the ECERS-R was developed. To be certified for the study on the ECERS-R each observer had to assign scores that were within one point of a consensus score for 30 of the 37 items (80%) during two practice observations in actual centers. Table 2 provides the percentage of the 1,350 centers that received each scale score (1 to 7) on each item. In the table, items are grouped by modal category, and the modal category is highlighted. Items in the same subscale tend to have similar scoring modes. The modal category is 1 or 2 for most items on the *Personal Care Routines* subscale, 4 for most items on the *Activities* subscale, and 7 for most items on the *Language and Reasoning* and *Interaction* subscales. The values of 3, 5, and 6 are never modal categories.

Alternative measures of child care quality—We coded a number of alternative measures of quality that indicate developmentally appropriate practice (some of which also capture what has been characterized as structural aspects of quality). Some of these were rated by the same observer who also rated the ECERS-R. Others were taken from a phone interview with the center director and the child's classroom teacher. With the exception of group size and caregiver:child ratio, we reverse coded the scores as needed so that higher scores on the alternative measures of quality would be expected to correlate with higher ECERS-R scores.

Observed measures: The *Arnett Caregiver Interaction Scale (Arnett CIS)* is an observational measure of caregiver-child interactions modeled after four well-established parenting styles: authoritarian, authoritative, permissive, and uninvolved (Arnett, 1989; Baumrind, 1967). Although Arnett organized the scale into four dimensions based on a principal components analysis, the factors have not been replicated by later studies; therefore, we followed the common practice of using the total sum of the items (author citation). We also use the *size of the child's group* and *ratio of children to caregivers*. Group size and child:caregiver ratio are based on the average of the first three counts of children, paid staff and volunteers made by the observer who also rated the ECERS-R.

Provider-reported measures: The child's classroom teacher reported her *educational level* and her *early childhood education credentials*, the latter of which is the sum of five indicators of: a) whether the teacher had four or more college classes in early childhood education, b) whether the teacher had ever taken special workshops or seminars in early childhood education, c) whether the teacher had specialized training in early childhood education in the past year, d) whether the teacher had or was working toward a Child Development Associate credential, and e) whether the teacher had any state-awarded certificates or credentials in early childhood education. *Interest areas* is the sum of the teacher's report of whether the classroom had "interest areas or centers for activities" organized around 10 different themes (e.g., reading area with books, water or sand table, private area for one or two children to be alone). *Child-centered activities* is the ratio of the reported hours spent in child-selected activities to the total time spent across: a) adult-directed whole class activities, b) adult-directed small group activities, c) adult-directed individual activities, and d) child-selected activities. *Language activities* and *math activities* is the sum of the frequency of eleven and ten items, respectively, coded from 0 = *never* to 5 = *everyday* such as "Work on learning names of letters," "Write own name," "Count out loud," and "Play math-related games."

Child outcome measures—To facilitate interpretation, we coded all child outcomes so that a higher score indicated better outcomes. All of the child outcomes were measured at the same study wave as when the child care setting was observed.

Cognitive outcome measures: We used two composite scores in the domains of reading and math that ECLS-B staff created using item response theory models (Najarian, Snow, Lennon, & Kinsey, 2010; Snow et al., 2007, 2009). These composite scores were created by ECLS-B staff because the investigators drew items to directly assess children's cognitive development from a variety of sources rather than from a single standardized measure, and single item scores are not included in the data.

The reading composite score included items from three subtests of the Preschool Language Assessment Scales (Simon Says, Art Show and Let's Tell Stories; Duncan & De Avila, 1998), from the Peabody Picture Vocabulary Test—Third Edition (Dunn & Dunn, 1997), and a measure of emergent early literacy, including letter sounds, early reading, phonological awareness, knowledge of print conventions, and matching words (Snow et al., 2007, 2009). The mathematics assessment composite score included items from the Test of Early Mathematics Ability, the ECLS-K cohort, and other sources in the following areas: number sense, geometry, counting, operations, and patterns (Najarian, Snow, Lennon, & Kinsey, 2010; Snow et al., 2007, 2009).

Socio-emotional outcome measures: We created socio-emotional measures based on parent-reported items that the ECLS-B study designers selected from the Preschool and Kindergarten Behavior Scales—Second Edition (PKBS-2; Merrell, 2003), Social Skills Rating System (SSRS; Gresham & Elliott, 1990), and Early Childhood Longitudinal Study-Kindergarten Cohort survey (Snow et al., 2007). Because the ECLS-B does not include composite scores for these items (Najarian et al., 2010), we created three subscales based on exploratory factor analyses and item response theory models (author citation): 1) *social competence* (e.g., how well the child plays with others, is liked by others, and is accepted by others), 2) *emotional and behavioral regulation* (e.g., lack of aggression, anger, and worry; expressions of happiness), and 3) *attention and concentration* (child pays attention well, does not disrupt the class, and is not overly active). We created Rasch measures, which are on a logit scale (Masters, 1982).

Health outcome measures: We re-coded the parent's overall rating of the child's health into a dichotomous indicator: *excellent* (coded 1), or *very good, good, fair, or poor* (coded 0). The parent also reported whether the child had a doctor-verified *respiratory illness, gastrointestinal illness, or ear infection*, as well as whether the child had experienced an injury that required a doctor's visit, since the last interview. We created dummy variables to indicate the absence of illness or injury.

Interviewers took two measurements for both the child's height and weight and these were averaged by the survey investigators. We computed the child's body mass index (BMI) and then defined the child as *not overweight* versus *overweight* based on CDC Growth Charts appropriate for the child's age.

Control variables—In the analysis of criterion validity, we adjusted for a number of covariates that may be associated with both the quality of child care and with child outcomes. Covariates included center, child, family, and community level characteristics.

Center characteristics: We created variables to measure the location and funding stream of all centers: Head Start, public school, private school, religious school/church, community non-profit or community for-profit. From information provided by the director, we

constructed variables measuring accreditation status, licensing status, and, if licensed, the number of children for which the center was licensed to provide care. We also indicated whether the center was willing to accept Child Care Development Fund subsidies.

Child characteristics: Child covariates include: child gender and racial and Hispanic identification (we coded these as Hispanic, non-Hispanic Black, or non-Hispanic of other race versus non-Hispanic White); whether the child was born low birth weight, whether the child was ever breast fed, the number of child's well-child doctor visits since the last interview, and whether the child had received WIC since the last interview. Outcomes measured when the child was two years included: the Bayley Short Form–Research Edition (BSF-R) mental, motor, and behavior scores (Bayley Short Form–Research Edition, 2001), whether the child had a doctor-confirmed respiratory illness, gastrointestinal flu, ear infection, or injury, a composite measure of the child's temperament (the Infant/Toddler Symptom Checklist; DeGagni, Poisson, Sickel, & Wiener, 1995), the child's overall health as reported by the mother and the child's BMI (based on measured height and weight).

Family characteristics: Mother demographic covariates include: home language was not English or the mother was born outside of the U.S.; whether there were any other children less than age 6 or any children ages 6 to 18 in the household; the mother's marital status; the mother's employment status; and maternal age. Family socioeconomic characteristics include: whether the family had used food stamps or TANF since the last interview, and a composite measure of family socioeconomic status created by ECLS-B staff by averaging five variables: (1) father/male guardian's education; (2) mother/female guardian's education; (3) father/male guardian's occupational prestige score; (4) mother/female guardian's occupational prestige score; and (5) household income. For households with only one parent, the ECLS-B staff averaged the available components (Snow et al., 2007, p. 405). Finally, we controlled for whether the mother was interviewed (versus another primary caregiver; 95% were mothers).

Community characteristics: We also merged data from the Decennial Census of Population and classified the child's ZIP Code of residence according to whether fewer than 10%, 10 to 19%, or 20% or more of young children (under age 5) in the ZIP Code had family incomes below 100% of the federal poverty level. To adjust for cross-region variation, we included dummy indicators for region (South, Midwest, West, or Northeast) of residence, and urbanicity of the ZIP Code (rural, urban area of fewer than 50,000 people, urban area of 50,000 people or more).

Analytic Approach

We examined three aspects of validity. Response process validity is fundamental to the meaning of item scores and thus relevant across uses of the scale (for research, practice and policy). Structural validity and criterion validity may differ depending on the uses of the scale, since researchers may be interested in measuring somewhat different aspects of quality than practitioners or policymakers (thus these different users may desire the scale to measure different dimensions and to correlate with different criteria).

Response process validity: Ordering of categories—Observers assigned a score of 1 to 7 to each ECERS-R item by checking several indicators anchored to the odd-numbered categories. The ECERS-R standard scoring system assumes order in item scores because indicators of higher scores are rated only if indicators of lower scores are met (referred to as “stop-scoring”). Most studies that use the ECERS-R, including the ECLS-B and other large-scale studies (such as the Head Start Family and Child Experiences Survey, Welfare, Children, and Families: A Three City Study, and the Fragile Families and Child Wellbeing

Study), implement this stop-scoring.¹ Ordering has not been tested empirically in prior U.S. studies (see Ba türk & I iko lu, 2008 and Lambert et al., 2008 for studies from outside the U.S.).

The intended ordering of the ECERS-R rating scale may be violated in practice. We anticipated disorder because the stop-scoring described above is implemented with ECERS-R items that mix dimensions. For example, *Ecersr10: Meals/snacks* includes indicators of nutrition (*food served of unacceptable nutritional value*), caregiver-child interactions (*nonpunitive atmosphere during meals*), language (*meals and snacks are times for conversation*), and sanitation (*sanitary conditions not usually maintained*), among others. If a setting is high on some of these dimensions (e.g., caregiver-child interactions, language, and nutrition) but low on others (e.g., sanitation), two observers might give different final scores, especially since the indicators to justify their scores are usually not retained and because observers interpret subjective terms and can probe caregivers for information. As a result, when presented with conflicting information across an item's indicators, some observers may (either consciously or subconsciously) follow the recommended scoring structure and give a low score (in our example because the indicators of sanitation are not met), whereas other observers may mark a higher score to reflect the higher level of other aspects of quality.

In order to study the response process validity of the ECERS-R, we conducted a series of analyses using the Rasch Partial Credit Model (PCM) in which we examined the ordering of the rating scale categories. The PCM is appropriate for multi-category rating scales such as the ECERS-R (Embretson & Reise, 2000). Importantly, the PCM does not force order between adjacent category thresholds and thus can be used to test for ordering (Andrich, 1996; Andrich, de Jong, & Sheridan, 1997). It is recommended that in a PCM, there are at least 10 ratings per category for each item; otherwise, it is recommended to collapse adjacent categories in order to achieve a sufficient number of ratings before running any analyses (Linacre, 2004; Wright & Linacre, 1992). Given the large ECLS-B sample size, we only had to collapse categories in three instances: *Ecersr12*, *Ecersr10*, and *Ecersr2* which resulted in these items having six rather than seven categories (five rather than six thresholds).²

The PCM produces estimates of the thresholds on the (latent) center quality scale (or sub scale) that separate two adjacent categories within each item. We used 95% confidence intervals for adjacent thresholds to define: (a) *order* if the upper bound of the confidence interval for the lower threshold was below the lower bound of the confidence interval for next higher threshold, (b) *equivalence* if the confidence intervals of the lower threshold and higher adjacent threshold overlapped, and (c) *disorder* if the upper bound of the confidence interval for the next higher adjacent threshold fell below the lower bound of the confidence interval for the adjacent lower threshold.

We estimated three sets of PCMs using Conquest statistical software (Version 2.0; Wu, Adams, Wilson, & Haldane, 2007): (1) a PCM in which all 36 items were assumed to measure a single dimension, (2) six separate PCMs based on the ECERS-R six subscales (see Table 1), and (3) three separate PCMs based on combining dimensions whose measures in the PCMs for the six subscales correlated at about .80 or above. We assessed disordering in each of these models.

¹As we discuss below, investigators may score all indicators rather than implement stop-scoring (e.g., see Hofer 2008, 2010 for examples) and doing so may reduce the problems with disorder that we find at the item level.

²Similar results regarding disorder are evident when we do not collapse these three pairs of categories (results available from the authors).

To examine the sensitivity of our results to the assumptions of the PCM, we also tested for order using three alternative models: (1) the nominal response model (NRM; Bock, 1972) which is an adjacent category model that adds a discrimination parameter to each category of every item, (2) the generalized partial credit model (GPCM; Muraki, 1992) which extends the partial credit model by adding a discrimination parameter to each item, and (3) the graded response model (GRM; Samejima, 1969) which assumes the categories are ordered. These models led to four conclusions that complement and extend the results presented here: (1) the NRM model, which assumes no order in the discrimination parameters and no order in the category thresholds, fit best; the discrimination parameters lack order, meaning that (2) the items should not be simply summed (or averaged); and (3) the level of quality that an indicator represents depends on the quality of the center that is being observed; and, (4) higher categories generally do not indicate higher quality (which is inferred from the lack of a linear ordering of the discrimination parameters; results available from the authors).

Structural validity: Number of dimensions—We estimated several factor analyses to investigate the dimensionality of the ECERS-R. We present results of an exploratory factor analysis with oblique rotation treating the data as ordinal (using weighted least squares with a diagonal weight matrix; results of additional exploratory and confirmatory models available from the authors).³ We followed guidelines in Brown (2006) to evaluate models: (1) the Non-normed Fit Index (NNFI), also known as Tucker-Lewis index (TLI), close to or above 0.95, and (2) the root mean square error of approximation (RMSEA) around 0.06 and below. For interpretation, we focused on items that loaded at or above 0.40 on a factor. We interpreted the results in relation to the ECERS-R six subscales, prior studies, and broad developmental domains. For use in regression analyses, we produced factor scores based on the three-factor solution.

Criterion validity: Associations with child outcomes and alternative measures of quality—To examine criterion validity, we conducted regression analyses to obtain estimates of associations of the ECERS-R with child outcomes and with alternative measures of quality of care (including the types of developmentally appropriate practices the ECERS-R is designed to capture). We report results of separate regressions that used the ECERS-R total score and the three factor scores as predictors. For continuous outcomes, we report standardized coefficients from OLS regressions that adjust for center, child, family, and community controls (listed in the Appendix).⁴ For dichotomous outcomes, we report the change in the predicted probability of scoring one on the outcome for a one standard deviation increase in the predictor, holding covariates (listed in the Appendix) constant at their means (Long, 1997). We evaluate the size of the associations in two ways: (1) using the cutoffs recommended by Cohen (1992), with standardized coefficients of about 0.20 or less as small, about 0.50 as medium, and about 0.80 as large and (2) in comparison to standardized coefficients for control variables. We adjusted coefficients for oversampling using the sampling weights and we adjusted the standard errors for the clustering of children within ZIP Codes.

³When the data are treated as ordinal, EFA attempts to reproduce the polychoric correlations among the items (Brown, 2006) while assuming that categories are ordered (i.e., constraining the thresholds in the following manner: $\tau_2 < \tau_3 < \dots < \tau_K$, where τ_K is the point on the quality scale that separates receiving a score in category k , or above and below category k ; Muthén & Muthén, 1998-2007; Tabachnick & Fidell, 2006).

⁴We draw similar conclusions based on simple correlations, without controls (results available from the authors).

Results

Response Process Validity: Ordering of Categories

Table 3 summarizes the degree of order found for the threshold estimates based on the Rasch models.⁵ All 36 items demonstrate at least one instance of disorder, even when we allowed for multiple dimensions; about two-thirds of the items also demonstrate equivalence (25 to 26 items depending on the model). Most items (25) had two points of disorder that occurred around Category 3 and around Category 5 (data not shown).

Figure 1 visualizes the disordering of the Rasch thresholds for an example item: *Ecercs34: Schedule*. The charts plot the threshold estimates based on the 36-item model (Figure 1a) and based on the six subscale models (Figure 1b). In both cases, this item shows disordering in two places: (a) between threshold 2-3 and threshold 3-4 and (b) between threshold 4-5 and threshold 5-6. That is, the estimated level of overall child care quality needed to receive a score of a 4 rather than a 3 is less than the estimated level of quality needed to receive a score of a 3 rather than a 2. Likewise, the estimated level of quality needed to receive a score of a 6 rather than a 5 is less than the estimated level of quality needed to receive a score of a 5 rather than a 4. Figures 1c and 1d show that the simple average scores are disordered or equivalent in comparable places for this item (disordered between category 2 and 3; equivalent between category 4 and 5). In additional analyses (results not shown), we found disorder or equivalence for the average total score for all items and for the average subscale scores on 32 of the 36 items.

In short, the category ordering assumed by the scale's developers is not consistently evident. We expect that the observed disorder reflects the fact that indicators within a single item reflect multiple dimensions. When observers follow the scoring instructions and assign a low score on an item due to the indicators of one of the dimensions, the higher quality of the center on other dimensions is missed. When observers violate the scoring instructions and assign a higher score on an item due to indicators of other dimensions, the lower quality of the center on one dimension is missed. Both situations may occur in practice.

Structural Validity: Number of Dimensions

Table 4 presents the factor loadings from the exploratory factor analysis for the single-factor, three-factor, and six-factor solutions. The model fit statistics indicated that a single-factor solution is not consistent with the data because the NNFI was below .95 and the RMSEA was above .06 (NNFI=.906 and RMSEA=.133). The six-factor solution is feasible (NNFI=.990 and RMSEA=.044), but the factor loadings shown in Table 4 do not reveal the six ECERS-R subscales. In most cases, items from at least two of the scale's original six subscales load on a factor (Factor1, Factor2, Factor3, Factor6). Factor5 has items from just one original subscale, although just two of the ten *Activities* items load on this factor. Finally, Factor4 is most consistent with the original subscale, with four of the six *Personal Care* items loading above .40 on the factor.

The three-factor solution was also viable based on the model fit statistics (NNFI=.980 and RMSEA=.062). The factor loadings shown in Table 4 reveal a structure similar to other recent factor analytic studies (Cassidy et al., 2005; Clifford et al., 2005; Early et al., 2006; Frede et al., 2007; Sakai et al., 2003), which generally combine several of the subscales. Almost all of the items that load above .40 on Factor1 come from *Space and Furnishings*, *Activities*, and *Program Structure*. Factor2 primarily draws items from *Personal Care*. And,

⁵We summarized the results this way for ease of interpretation. Disordering was also clearly evident in category probability curves, a traditional method of displaying response probabilities in the psychometric literature (results available from the authors).

(with one exception) all items that load above .40 on Factor3 come from *Language-Reasoning* and *Interaction*. We might interpret the items of *Personal Care* as somewhat more focused on the health domain and the items in *Language-Reasoning/Interactions* as somewhat more focused on caregiver-child interactions that may foster cognitive and socio-emotional development. However, the factors have relatively high inter-correlations (.75 to .79) which may reflect the fact that many items capture multiple aspects of quality.

In short, consistent with prior factor analyses of the ECERS and ECERS-R, we did not find evidence of a single global aspect of quality nor for six subscales of quality. Our three-factor solution is consistent with some prior factor analyses of the ECERS and more recent factor analyses of the ECERS-R, which often find two- or three-factor solutions with similar items loading on the factors (e.g., Cassidy et al., 2005; Clifford et al., 2005; Early et al., 2006; Frede et al., 2007; Sakai et al., 2003).

Criterion Validity: Associations with Child Outcomes and Alternative Measures of Quality

Table 5 shows the standardized coefficients for each child outcome measure and each alternative measure of quality. For each outcome, the results come from four separate regressions with one of our four ECERS-R measures as a predictor: the total score (the score typically used in research and for policy purposes) and the three factor scores (reflecting our best fitting factor structure). As noted above, the models control for center, child, family, and community characteristics (listed in the Appendix), and adjust for oversampling and for clustering of children within ZIP Codes.

Particularly relevant for uses of the scale in developmental research, we found few associations with child outcomes. The standardized coefficients from regressions of child development outcomes on the ECERS-R total score and factor scores are uniformly small in magnitude, with the largest being 0.08. This is small given conventional cutoffs (Cohen, 1992) and relative to standardized coefficients for key controls variables (e.g., family SES has a standardized coefficient of 0.23 and 0.26 in association with children's reading and math scores, respectively). In fact, there are no significant associations between the ECERS-R and children's reading and math scores. We found a few significant associations with children's socio-emotional and health outcomes although they are small in size and not in a pattern consistent with domain-specificity. For example, although the *Personal Care* factor (which includes the health-specific item *Ecrrsr13*) associates positively and significantly with excellent health, so does the *Language-Reasoning/Interactions* factor. In addition, the predicted probability of the absence of respiratory illness increases significantly with the *Space and Furnishings*, *Activities*, and *Program Structure* factor and not with *Personal Care* (although both estimates round to .02). Finally, *Personal Care* associates significantly with two measures of child socio-emotional development and at higher magnitudes than the potentially more domain-specific *Language-Reasoning/Interactions* factor.

Particularly relevant for uses of the ECERS-R by practitioners, there is more evidence of criterion validity with alternative measures of quality. Standardized coefficients are highest with the alternative observational measure, the Arnett CIS; and, as expected, the largest standardized coefficient is seen for the ECERS-R factor measuring the most similar construct as the Arnett CIS (*Language-Reasoning/Interactions*) with a moderate to high effect size of 0.73. The standardized coefficients for the remaining aspects of quality are small to moderate in size, again with some evidence of higher associations with the most similar constructs. For example, the teacher report of interest areas, child-centered activities, and math activities all have standardized coefficients of about 0.30 to 0.40 for the factor that captures *Space and Furnishings*, *Activities*, and *Program Structure*. These same teacher reports of activities and interest areas have standardized coefficients of 0.21-0.25 with the

Language-Reasoning/Interactions factor. The association with teacher-reported language activities is somewhat smaller, at 0.13.

The measures of teacher education and training, group size and child:caregiver ratios, which are also relevant to developmentally appropriate practice, show small but statistically significant associations with the ECERS-R. The associations are larger for the teacher's education specific to early childhood education than her overall educational level, with the highest association being with the factor capturing *Space and Furnishings/Activities/Program Structure* (a standardized coefficient of 0.17). A higher ratio of children to caregivers is associated with lower ECERS-R quality, with the largest association being with the factor capturing *Language-Reasoning/Interactions* (a standardized coefficient of -0.16). Most associations with group size are non-significant, except for a small but positive association with *Space and Furnishings/Activities/Program Structure* (a standardized coefficient of 0.11).

In additional analyses we also find that, like the ECERS-R, these alternative measures of quality are generally not significantly associated with child outcomes and, when associations are significant, effect sizes are small (results available from the authors).

Discussion

Our results provide new insights into the validity of the ECERS-R for developmental research, for practice, and for policy. Relevant for all of these uses is our finding about the measure's response process validity, where we see at least one instance of disorder in the categories of all 36 items. As far as we are aware, ours is the first study to use item response theory in a U.S. sample to test for the order assumed by the ECERS-R developers. Also relevant across uses of the scale, we fail to find that the ECERS-R measures a single global aspect of quality or six subscales of quality. Instead, our factor analyses reveal three factors that are similar to prior studies, but are based on a broad set of centers where a nationally representative sample of preschoolers receive care. When we used these three factors, and the ECERS-R total score, to predict various criteria, our conclusions depend on the type of criterion and thus use of the scale. The greatest evidence was for criterion validity with respect to the aspects of developmentally appropriate practice that the ECERS-R was designed to measure. That is, associations with alternative measures of quality were often significant, were moderate to large in size, and were highest for correlations between ECERS-R factors and alternative measures of similar constructs. There was less evidence of criterion validity for developmental research, however. The ECERS-R total score and its factor scores were rarely significantly associated with child outcomes; and, when they were, the associations were small in size. Even with our extension of prior research to consider multiple developmental domains, associations were not consistently higher with factors that measured aspects of quality that might be expected to be most relevant for the outcome domain.

Our finding of small associations with child outcomes replicates prior studies of the ECERS-R. Similarly small effect sizes have been found for other measures of child care quality, suggesting that this is a broader issue in the field, and not unique to the ECERS-R (Burchinal, Kainz, & Cai, 2011). Many reasons have been offered for these small associations such as the fact that exposure to any child care setting is often limited (in terms of hours per week and months of attendance) especially relative to other contexts (such as family and neighborhood) and the fact that scores on standardized tests may be less sensitive to child care contexts than other child outcomes like children's stress, moods, or engagement (Dunn & Kontos 1997). However, our study provides an additional explanation that small correlations may be attributable, in part, to low validity of the measure itself. Notably, IRT

approaches have been underutilized with measures of child care quality (not just the ECERS-R),⁶ and so new insights into the validity of other measures may also be provided by examining them with an IRT approach. We encourage the field to increase attention to these methods, as is increasingly done in other areas of research (e.g., DeRoos & Allen-Meares, 1998; Piquero, Macintosh, & Hickman, 2002; Rapport, LaFond, & Sivo, 2009).

The widespread adoption of the ECERS-R for a variety of programmatic, policy, and research purposes necessitates comprehensive validity studies that look at multiple aspects of validity using techniques drawn from both classical test theory and item response theory (Joint Committee on Standards for Educational and Psychological Testing, 1999; Kane, 2006). Three decades ago the ECERS and ECERS-R advanced the scientific measurement of child care quality, based primarily on accumulated professional experience and expert review, but without the benefit of the detailed psychometric approaches to measurement available today. The ECERS-R rightly took its place as the field standard, finding widespread applications in research and policy. However, given the psychometric tools now available, the results of our study suggest that an iterative measurement development process, attending to multiple aspects of validity from start to finish, may be needed in order to produce better measures of child care quality for future developmental research (Wolfe & Smith, 2007a, 2007b).

For researchers, practitioners, and policymakers who continue to use the ECERS-R, especially to compare to prior studies and/or in longitudinal research, we recommend that all indicators be scored. For researchers, scoring all of the indicators provides the needed “bridge” to earlier studies that imposed the stop scoring (since the score can be calculated both with and without stop scoring) and also allows them to put the indicators together in other ways (ideally based on IRT methods). For practitioners, scoring all indicators is in line with the ECERS-R origins as a checklist and provides more information than does stop-scoring about what centers are currently doing well and where they need to improve. For policymakers, scoring all indicators credits centers for all that they are doing well in the high stakes context of attaching funding to scores; doing so is potentially important given evidence that about one-quarter additional centers moved above one state’s cutoff for higher funding when all indicators were taken into account versus when the standard stop-scoring was used (Hofer 2008, 2010).

We also recommend that future scale development efforts be directed at the ECERS-R and alternative measures. Regarding the existing ECERS-R, as noted above, we anticipated disorder due to (a) the mixing of different aspects of quality within the indicators of a single item combined with the “stop scoring” rules, (b) the numerous subjective assessments observers must make during coding, (c) the need to rely on teacher reports for scoring some indicators, and (d) failure to retain all indicators, such that there is no record to justify the item score. Studies that use cognitive interviewing approaches might probe observers to verify whether these are in fact the root problems producing disorder (and to discover other challenges that make it difficult for observers to follow the standard scoring scheme and/or to assign a score; Joint Committee on Standards for Educational and Psychological Testing, 1999). These studies could compare observers’ experiences with the stop-scoring and all-indicator scoring approaches to provide, for example, estimates of the time it takes to implement each method and the challenges observers experience with each approach. To the extent that collecting all indicators is time-consuming, future IRT analysis of the ECERS-R indicators could also help identify indicators that are redundant and those that might be removed. Such IRT analyses could also identify sets of indicators that assess similar

⁶For example, Child Trends’ comprehensive compendium of measures cites only two uses of IRT across numerous reliability and validity analyses of 34 measures (Halle & Vick, 2007).

developmental constructs (e.g., health/safety) and/or sets of items that assess similar categories of interest to practitioners (e.g., activities).

Beyond the existing ECERS-R, we recommend particular attention to the following major issues in new scale development for measuring child care quality:

1. Dimensions of quality should be carefully defined and items written specifically for these dimensions.
2. In this process, dimensions of quality should be carefully aligned with the intended use of the measure (e.g., for developmental research, dimensions of quality specific to domains of child development; for policies aimed at improving school readiness, aspects of quality that will prepare children for school; for practitioners, dimensions of quality that address, for example, regulations and/or accreditation standards and/or dimensions of quality that are emphasized by the profession).
3. Item pools should be developed and evaluated with item response theory approaches to identify a set of items that measure different levels along each relevant dimension of quality; samples used in this development process should reflect the range of quality in the target population for the intended use of the scale.
4. Expected associations with outcomes should be explicitly stated as the measure is developed (e.g., where and why within- and cross-domain associations with specific measures of child development are expected) and studies should be designed with attention to increasing predictive validity (hours and months exposed to a setting, random assignment to settings of different quality or collection of extensive controls for confounds).

In general, our results support the increasing attention scholars and policymakers are giving to the measurement of child care quality (Forry, Vick, & Halle, 2009; Layzer & Goodson, 2006; Zaslow et al., 2006; Zaslow, Martinez-Beck, Tout, & Halle, 2011), as well as recent attempts to develop new measures of quality of care (e.g., Sylva et al., 2006; Pianta, LaParo, & Harms, 2009). We offer additional evidence that the low associations between child care quality and child development outcomes so commonly found in the literature may, in part, reflect the weak psychometric properties of the scales themselves for developmental research (including the types of disordering and lack of dimensionality identified here).

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090065, and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, through Grant R01HD060711. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Everett Smith for his expert consultation on our psychometric analyses, Carol Myford for detailed comments on an earlier version of this paper, and Claudia Melgar for assistance with the literature review.

Appendix: Descriptive Statistics for Variables Used in Regression Analyses

Appendix

Descriptive Statistics for Variables Used in Regression Analyses

	<i>M</i> or %	<i>SD</i>	<i>Min</i>	<i>Max</i>
ECERS-R				
Total score	4.55	1.07	1.25	6.97
Factor scores				

Factor score 1: Space and furnishings, activities, and program structure	0.00	0.46	-1.55	1.42
Factor score 2: Personal care	0.02	0.58	-1.82	1.75
Factor score 3: Language-reasoning and interaction	0.01	0.71	-2.35	2.02
Child Outcomes				
Cognitive				
Reading composite	-0.38	0.75	-2.38	2.60
Math composite	-0.35	0.78	-2.83	2.38
Socio-emotional				
Social competence	1.49	0.98	-1.61	3.90
Emotional and behavioral regulation	1.01	0.61	-1.39	2.75
Attention and concentration	0.78	0.86	-1.60	3.28
Health				
Child excellent health	54.29			
Child is not overweight	80.63			
No doctor verified respiratory illness	86.52			
No doctor verified gastrointestinal illness	96.13			
No doctor verified ear infection	57.18			
No injury that required doctor's visit	76.20			
Alternative Measures of Quality				
Observed measures				
Arnett Caregiver Interaction Scale	64.65	11.68	4	78
Observed group size	14.04	4.52	1.67	36.67
Observed child:teacher ratio	7.13	3.03	0.42	22.33
Provider-reported measures				
Teacher's educational level	17.30	1.98	3	22
Teacher's early childhood education credentials	3.21	0.98	0	5
Interest areas	8.87	1.27	2	10
Child-centered activities	0.33	0.17	0	0.80
Math activities	35.99	8.01	0	50
Language activities	41.69	7.61	5	55
Center-Level Controls				
Type of center				
Head Start	17.39			
Public school	25.12			
Private school	21.09			
Religious or church	12.30			
Community-based center, non-profit	10.65			
Community-based center, for-profit	13.46			
Size and license status				
Not licensed	17.88			
Licensed for 50 or fewer children	26.11			
Licensed for 51 to 100 children	20.59			
Licensed for 100 or more children	35.43			

Accredited	47.52			
Accept children with subsidies	46.11			
Child-Level Controls	<i>M</i> or %	<i>SD</i>	<i>Min</i>	<i>Max</i>
Child race/ethnicity				
Hispanic	23.81			
Non-Hispanic, White	53.33			
Non-Hispanic, Black	16.05			
Non-Hispanic, Other	6.79			
Child is female	48.95			
Child was born low birth weight	7.60			
Mother ever breast fed child	70.18			
Zero or one well child check-ups since the last interview	19.77			
Since last interview, child received WIC	32.70			
Bayley scores at two years				
Mental	119.98	32.08	0	157.99
Motor	75.65	21.59	0	99.56
Behavior	38.26	11.17	0	55
Child temperament composite at two years	8.87	4.23	0	21
Child in excellent health at two years	58.68			
Child BMI at two years	15.77	5.55	0	36.3
Child has no doctor-verified illness or injury at two years	39.23			
Family-Level Controls				
Mother's age				
Age 35+	38.19			
Age 25 – 34	49.52			
Age < 25	12.29			
Mother married	63.88			
Mother's employment status				
Not employed	36.09			
Full time	43.59			
Part time	20.32			
No English at home or mother not born in US	21.15			
Other children in the household				
Any other children <6 in the household	46.22			
Any children 6-18 in the household	53.35			
Family socioeconomic status	-0.03	0.82	-1.98	2.01
Since last interview, family used Food Stamps	25.60			
Since last interview, family used TANF	9.94			
Community-Level Controls				
Young child poverty rate in ZIP Code				
Less than 10%	40.51			
10-19%	20.10			
20% or more	39.39			

Urbanicity	
Rural	14.77
Urban (in a cluster with less than 50,000 residents)	11.93
Urban (in a cluster with 50,000 or more residents)	73.30
Region	
West	18.08
South	41.94
Northeast	20.11
Midwest	19.88

Note. $n = 1,150$ (rounded to nearest 50, as per ECLS-B data sharing agreement) for all variables except the child's math and reading composite and the child's weight status which are $n = 1,100$. All values weighted by the ECLS-B child care observation ("CCO") sampling weight. Dummies for child's age in months and whether the mother (versus another parent figure) was interviewed are controlled in regression models but not shown in the Appendix.

References

- Andrich, D. Measurement criteria for choosing among models with graded responses. In: Eye, A. v.; Clogg, CC., editors. *Categorical variables in developmental research: Methods of analysis*. Academic Press, Inc.; San Diego, CA: 1996. p. 3-35.
- Andrich, D.; de Jong, JH.; Sheridan, BE. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost, J.; Langeheine, R., editors. *Applications of latent trait and latent class models in the social sciences*. Waxmann Verlag GMBH; New York, NY: 1997. p. 58-68.
- Arnett J. Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*. 1989; 10(4):541-552.
- Ba türk R, I iko lu N. Analyzing process quality of early childhood education with many facet Rasch measurement model. *Educational Sciences: Theory and Practice*. 2008; 8:25-32.
- Baumrind D. Child-care practices anteceding three patterns of preschool behavior. *Genetic Psychology Monographs*. 1967; 75:43-88. [PubMed: 6032134]
- Bayley Short Form-Research Edition. *Bayley Scales of Infant Development: Second Edition (BSID-II)*. The Psychological Corporation, a Harcourt Assessment Company; San Antonio, TX: 2001. Adapted from N. Bayley.
- Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*. 1972; 37(1):29-51.
- Brown, T. *Confirmatory factor analysis for applied research*. Guilford Press; New York, NY: 2006.
- Bryant DM, Clifford RM, Peisner ES. Best practices for beginners: Developmental appropriateness in kindergarten. *American Educational Research Journal*. 1991; 28:783-803.
- Burchinal MR, Howes C, Pianta R, Bryant D, Early D, Clifford R, Barbarin O. Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science*. 2008; 12:140-153.
- Burchinal, M.; Kainz, K.; Cai, Y. How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In: Zaslow, M.; Martinez-Beck, I.; Tout, K.; Halle, T., editors. *Quality Measurement in Early Childhood Settings*. Brookes Publishing; Baltimore, MD: 2011.
- 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; 20:345-360.
- Clifford RM, Barbarin O, Chang F, Early D, Bryant D, Howes C, Pianta R. What is pre-kindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science*. 2005; 9(3): 126-143.
- Cohen J. A power primer. *Psychological Bulletin*. 1992; 112(1):155-159. [PubMed: 19565683]

- Copple, C.; Bredekamp, S., editors. Developmentally appropriate practice in early childhood programs serving children from birth through age 8. National Association for the Education of Young Children; Washington, DC: 2009.
- Cryer D. Defining and assessing early childhood program quality. *Annals of the American Academy of Political and Social Science*. 1999; 563:39–55.
- Cryer, D.; Harms, T.; Riley, C. All about the ECERS-R. Kaplan; Lewisville, NC: 2003.
- DeGangi, GA.; Poisson, S.; Sickel, RZ.; Wiener, AS. Infant/Toddler Symptom Checklist. The Psychological Corporation; San Antonio, TX: 1995.
- DeRoos Y, Allen-Meares P. Application of Rasch analysis: Exploring differences in depression between African-American and white children. *Journal of Social Service Research*. 1998; 23:93–107.
- Duncan, SE.; De Avila, EA. PreLAS 2000. CTB/McGraw-Hill; Monterey, CA: 1998.
- Dunn L, Kontos S. What have we learned about developmentally appropriate practice? *Young Children*. 1997; 52:4–13.
- Dunn, LM.; Dunn, LM. Picture Vocabulary Test–Third edition (PPVTIII). Pearson Publishing; Upper Saddle River, NJ: 1997.
- Early DM, Bryant DM, Pianta RC, Clifford RM, Burchinal MR, 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.
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Lawrence Erlbaum Associates; Mahwah, NJ: 2000.
- Emlen, AC. Solving the childcare flexibility puzzle. Universal Publishers; Boca Raton, FL: 2010.
- Forry, N.; Vick, J.; Halle, T. Evaluating, developing, and enhancing domain-specific measures of child care quality. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services; Washington, DC: 2009. (OPRE Research-to-Policy Brief #2)
- Frank Porter Graham Child Development Institute. Early developments. The University of North Carolina at Chapel Hill; Chapel Hill, NC: 2003.
- Frede, E.; Jung, K.; Barnett, WS.; Lamy, CE.; Figueras, A. The Abbott Preschool Program Longitudinal Effects Study (APPLES): Interim report. National Institute for Early Education Research; New Brunswick, NJ: 2007.
- Fuller B, Kagan SL, Loeb S, Chang Y. Child care quality: Centers and home settings that serve poor families. *Early Childhood Research Quarterly*. 2004; 19:505–527.
- Gresham, FM.; Elliott, SN. Social Skills Rating System manual. American Guidance Service; Circle Pines, MN: 1990.
- Halle, T.; Vick, JE. Quality in early childhood care and education settings: A compendium of measures. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services; Washington, DC: 2007.
- Harms, T.; Clifford, RM. Early Childhood Environment Rating Scale. Teachers College Press; New York: 1980.
- Harms, T.; Clifford, RM.; Cryer, D. Early Childhood Environment Rating Scale, Revised Edition. Teachers College Press; New York: 1998.
- Helburn, SW.; Culkin, ML.; Morris, JR.; Mocan, HN.; Howes, SC.; Phillipsen, LC.; Rustici, J. Cost, Quality, and Child Outcomes in child care centers. University of Colorado; Denver, CO: 1995. (Technical Report)
- Hofer, KG. Unpublished doctoral dissertation. Vanderbilt University; Nashville, TN: 2008. Measuring quality in pre-kindergarten classrooms: Assessing the Early Childhood Environment Rating Scale.
- Hofer KG. How measurement characteristics can affect ECERS-R scores and program funding. *Contemporary Issues in Early Childhood*. 2010; 11(2):175–191.
- Holloway SD, Kagan SL, Fuller B, Tsou L, Carroll J. Assessing child-care quality with a telephone interview. *Early Childhood Research Quarterly*. 2001; 16:165–189.

- Howes C, Burchinal M, Pianta R, Bryant D, Early D, Clifford R, Barbarin O. Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*. 2008; 23:27–50.
- Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing. American Educational Research Association; Washington, DC: 1999.
- Kane, MT. Validation. In: Brennan, RL., editor. Educational measurement (4th Edition; Sponsored jointly by the National Council on Measurement in Education and the American Council on Education). Praeger; Westport, CT: 2006. p. 17-64.
- Kontos S. Child care quality, family background, and children's development. *Early Childhood Research Quarterly*. 1991; 6:249–262.
- Lambert MC, Williams SG, Morrison JW, Samms-Vaughan ME, Mayfield WA, Thornberg KR. Are the indicators for the Language and Reasoning Subscales of the Early Childhood Environment Rating Scales-Revised psychometrically appropriate for Caribbean classrooms? *International Journal for Early Years Education*. 2008; 16:41–60.
- Layzer J, Goodson BD. The “quality” of early care and education settings: Definitional and measurement issues. *Evaluation Review*. 2006; 30:556–576. [PubMed: 16966676]
- Linacre, JM. Optimizing rating scale category effectiveness. In: Smith, EV., Jr; Smith, RM., editors. Introduction to Rasch measurement. Journal of Applied Measurement Press; Maple Grove, MN: 2004. p. 258-278.
- Long, JS. Regression models for categorical and limited dependent variables. Sage; Thousand Oaks, CA: 1997.
- Mashburn AJ, Pianta RC, Hamre B, Downer J, Barbarin O, Bryant D, Howes C. Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*. 2008; 79:732–749. [PubMed: 18489424]
- Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982; 47(2):149–174.
- McCartney K. Effects of quality of day care environment on children's language development. *Developmental Psychology*. 1984; 20:244–260.
- Merrell, KM. Preschool and Kindergarten Behavior Scales (PKBS-2). Pro-Ed, Inc; Austin, TX: 2003.
- Montes G, Hightower AD, Brugger L, Moustafa E. Quality child care and socioemotional risk factors: No evidence of diminishing returns for urban children. *Early Childhood Research Quarterly*. 2005; 20:361–372.
- Muraki E. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*. 1992; 16(2):159–176.
- Muthén, LK.; Muthén, BO. Mplus user's guide. Muthén & Muthén; Los Angeles: 1998-2007.
- Najarian, M.; Snow, K.; Lennon, J.; Kinsey, S. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool-Kindergarten 2007 Psychometric Report. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education; Washington, DC: 2010. (NCES 2010-009)
- Peisner-Feinberg, ES.; Burchinal, MR.; Clifford, RM.; Culkin, M.; Howes, C.; Kagan, SL.; Zelazo, J. The children of the cost, quality, and outcomes study go to school: Executive summary. University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Center; Chapel Hill, NC: 1999.
- 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.
- Pianta, RC.; LaParo, KM.; Harms, B. Classroom Assessment Scoring System Manual Pre-K. Brookes; Baltimore, MD: 2009.
- Piquero AR, Macintosh R, Hickman M. The validity of a self-reported delinquency scale: Comparisons across gender, age, race, and place of residence. *Sociological Methods and Research*. 2002; 30:492–529.
- Rapport MD, LaFond SV, Sivo SA. One-dimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A Rasch analysis. *Journal of Psychopathology and Behavioral Assessment*. 2009; 31:309–319.

- 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.
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 34(4, Pt. 2):100.
- Snow, K.; Derecho, A.; Wheelless, S.; Lennon, J.; Kinsey, S.; Morgan, K.; Einaudi, P.; Mulligan, GM. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Kindergarten 2006 and 2007 data file user's manual. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics; Washington, DC: 2009. (National Center for Education Statistics No. 2010-010)
- Snow, K.; Thalji, L.; Derecho, A.; Wheelless, S.; Lennon, J.; Kinsey, S.; Park, J. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Preschool year data file user's manual (2005–06) (National Center for Education Statistics No. 2008-024). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics; Washington, DC: 2007.
- Sylva K, Siraj-Blatchford I, Taggart B, Sammons P, Melhuish E, Elliot K, Totsika V. Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*. 2006; 21(1):76–92.
- Tabachnick, BG.; Fidell, LS. Using multivariate statistics. 5 ed.. Allyn and Bacon; Boston, MA: 2006.
- U.S. Department of Education, National Center for Education Statistics. ECLS-B 9-month sample sizes, population sizes, and response rates. 2001. (NCES 2008-024)Retrieved from http://nces.ed.gov/ECLS/pdf/9mo_samplesize.pdf
- Wolfe, EW.; Smith, EV, Jr. Instrument development tools and activities for measure validation using Rasch models: Part I--instrument development tools. In: Smith, EV., Jr; Smith, RM., editors. Rasch measurement: Advanced and specialized applications. Journal of Applied Measurement Press; Maple Grove, MN: 2007a. p. 202-242.
- Wolfe, EW.; Smith, EV, Jr. Instrument development tools and activities for measure validation using Rasch models: Part II--validation activities. In: Smith, EV., Jr; Smith, RM., editors. Rasch measurement: Advanced and specialized applications. Journal of Applied Measurement Press; Maple Grove, MN: 2007b. p. 243-290.
- Wright BD, Linacre JM. Combining and splitting categories. *Rasch Measurement Transactions*. 1992; 6(3):233–235.
- Wu, ML.; Adams, RJ.; Wilson, MR.; Haldane, SA. ACER ConQuest version 2.0: Generalised item response modeling software. ACER Press; Camberwell, Australia: 2007. [Computer software and manual]
- Zaslow M, Halle T, Martin L, Cabrera N, Calkins J, Pitzer L, Margie NG. Child outcome measures in the study of child care quality. *Evaluation Review*. 2006; 30:577–610. [PubMed: 16966677]
- Zaslow, M.; Martinez-Beck, I.; Tout, K.; Halle, T., editors. Quality measurement in early childhood settings. Brookes Publishing; Baltimore, MD: 2011.
- 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. (MG-650-QEL)
- Zill, N.; Resnick, G.; Kim, K.; O'Donnell, K.; Sorongon, A.; McKey, RH.; D'Elio, MA. Head Start FACES 2000: A whole-child perspective on program performance. U.S. Department of Health and Human Services, Administration on Children, Youth, and Families; Washington, DC: 2003.

Figure 1.

Bar Charts Demonstrating Disorder for an Example ECERS-R Item (Ecersr34: Schedule) from the ECLS-B

Note. A constant was added to the Rasch threshold estimates shown in charts ‘a’ and ‘b’ so that all values are positive. Charts ‘c’ and ‘d’ show the average of the ECERS-R total score or subscore within each category of the item. The total score and subscore are calculated following the standard procedure of averaging the scores of all 36 items (total score) or the subscale items (for *Ecersr34* the subscale is Program Structure as shown in Table 1). Standard error bars are shown around the threshold estimates (charts a and b) and category averages (charts c and d), except for threshold 6-7 because it is automatically the value that offsets the sum of the other threshold estimates to zero.

Table 1**Subscales and Items of the ECERS-R**

Space and Furnishings	Activities
1. Indoor space	19. Fine motor
2. Furniture for routine care, play and learning	20. Art
3. Furnishings for relaxation and comfort	21. Music/movement
4. Room arrangement for play	22. Blocks
5. Space for privacy	23. Sand/water
6. Child-related display	24. Dramatic play
7. Space for gross motor play	25. Nature/science
8. Gross motor equipment	26. Math/number
	27. Use of TV, video, and/or computers
Personal Care Routines	28. Promoting acceptance of diversity
9. Greeting/departing	
10. Meals/snacks	Interaction
11. Nap/rest	29. Supervision of gross motor activities
12. Toileting/diapering	30. General supervision of children (other than gross motor)
13. Health practices	31. Discipline
14. Safety practices	32. Staff-child interactions
Language-Reasoning	33. Interactions among children
15. Books and pictures	
16. Encouraging children to communicate	Program Structure
17. Using language to develop reasoning skills	34. Schedule
18. Informal use of language	35. Free play
	36. Group time
	37. Provisions for children with disabilities
	Parents and Staff
	38. Provisions for parents
	39. Provisions for personal needs of staff
	40. Provisions for professional needs of staff
	41. Staff interaction and cooperation
	42. Supervision and evaluation of staff
	43. Opportunities for professional growth

Source: (Harms, Clifford, & Cryer, 1998).

Note. We excluded the shaded items because the Early Childhood Longitudinal Study, Birth Cohort did not gather ECERS-R Items 38 to 43 (Snow et al., 2007, p. 81) and 61% of centers were missing on Item 37; this focus on the first 36 ECERS-R items is consistent with prior research (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; Clifford et al., 2005; Frede et al., 2007).

Table 2

Percentage of Centers Receiving Each Scale Value for Each ECERS-R Item, Sorted by Modal Category

Item	Subscale	Scale Value						
		1	2	3	4	5	6	7
Eciersr12: Toileting	Personal Care Routines	31%	26%	2%	10%	1%	9%	23%
Eciersr10: Meals/Snacks	Personal Care Routines	41%	19%	1%	9%	6%	9%	16%
Eciersr27: Multimedia Use	Activities	17%	27%	3%	22%	6%	13%	11%
Eciersr7: Gross Motor Space	Space and Furnishings	9%	29%	6%	20%	10%	15%	12%
Eciersr34: Schedule	Program Structure	4%	30%	2%	27%	2%	8%	28%
Eciersr11: Nap/Rest	Personal Care Routines	18%	31%	4%	30%	2%	4%	12%
Eciersr13: Health	Personal Care Routines	7%	53%	1%	6%	3%	10%	20%
Eciersr17: Language Reasoning	Language-Reasoning	6%	6%	20%	27%	6%	8%	27%
Eciersr6: Child-Related Display	Space and Furnishings	2%	10%	26%	28%	10%	15%	10%
Eciersr23: Sand/Water	Activities	15%	3%	18%	29%	8%	16%	12%
Eciersr5: Privacy Space	Space and Furnishings	7%	7%	22%	30%	5%	10%	18%
Eciersr21: Music	Activities	4%	30%	14%	31%	8%	7%	6%
Eciersr28: Diversity Acceptance	Activities	6%	7%	23%	33%	10%	8%	12%
Eciersr3: Comfortable Furnishings	Space and Furnishings	8%	6%	19%	33%	5%	11%	18%
Eciersr20: Art	Activities	4%	10%	22%	37%	4%	10%	13%
Eciersr25: Nature/Science	Activities	16%	22%	11%	38%	1%	3%	9%
Eciersr19: Fine Motor	Activities	2%	6%	10%	41%	4%	12%	25%
Eciersr22: Blocks	Activities	8%	8%	5%	49%	8%	17%	5%
Eciersr24: Dramatic Play	Activities	6%	9%	11%	50%	10%	11%	3%
Eciersr26: Math	Activities	8%	4%	13%	53%	4%	6%	11%
Eciersr15: Books	Language-Reasoning	4%	6%	7%	60%	2%	3%	20%
Eciersr29: Gross Motor Supervision	Interaction	8%	7%	5%	20%	20%	18%	23%
Eciersr8: Gross Motor Equipment	Space and Furnishings	9%	20%	5%	17%	5%	17%	27%
Eciersr14: Safety	Personal Care Routines	28%	23%	1%	10%	2%	7%	29%
Eciersr35: Free Play	Program Structure	5%	8%	5%	28%	5%	15%	34%
Eciersr31: Discipline	Interaction	5%	5%	4%	11%	16%	23%	35%
Eciersr1: Indoor Space	Space and Furnishings	4%	6%	6%	28%	4%	14%	37%
Eciersr16: Child Communication	Language-Reasoning	2%	3%	5%	20%	5%	23%	42%
Eciersr4: Room Play Friendly	Space and Furnishings	4%	7%	9%	17%	6%	14%	43%
Eciersr18: Informal Language	Language-Reasoning	2%	3%	8%	28%	4%	12%	44%
Eciersr30: General Supervision	Interaction	7%	8%	2%	13%	11%	16%	45%
Eciersr33: Children Interactions	Interaction	3%	5%	3%	14%	2%	27%	46%
Eciersr36: Group Time	Program Structure	7%	2%	6%	15%	5%	16%	49%
Eciersr2: Routine Care Furniture	Space and Furnishings	2%	2%	0%	8%	3%	23%	62%
Eciersr9: Greeting/Departing	Personal Care Routines	3%	6%	4%	12%	3%	9%	63%
Eciersr32: Staff-Child Interactions	Interaction	5%	5%	2%	10%	1%	7%	71%

Note. The modal category for each item is highlighted. The items are ordered first by their modal category and second by the percentage in that modal category.

Table 3

Summary of Number of ECERS-R Items Demonstrating Disorder and/or Equivalence based on Rasch Thresholds

	Number of Items With			
	Neither Disorder Nor Equivalence	Disorder Only	Equivalence Only	Both Disorder And Equivalence
<u>Rasch Model</u>				
36 items	0	11	0	25
6 subscales	0	10	0	26
3 combined subscales	0	10	0	26

Note. Disorder and equivalence defined based on 95% confidence intervals for thresholds. For disordered thresholds, the confidence interval of the lower adjacent threshold is higher than the next adjacent higher threshold. For equivalent thresholds, the confidence interval of the adjacent thresholds overlap.

Table 4
Factor Loadings from Exploratory Factor Analysis: One-, Three-, and Six-Factor Solutions

Single Factor		Three Factor Solution			Six Factor Solution					
Item	Solution	Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Space and Furnishings										
Eeersr1: Indoor Space	0.425									
Eeersr2: Routine Care Furniture	0.505	0.402			0.573					
Eeersr3: Comfortable	0.605	0.733			0.671					
Eeersr4: Room Play Friendly	0.623	0.595			0.669					
Eeersr5: Privacy Space	0.622	0.565								
Eeersr6: Child-Related Display	0.487	0.514			0.567					
Eeersr7: Gross Motor Space			0.700			0.777				
Eeersr8: Gross Motor Equipment	0.414					0.500				0.461
Activities										
Eeersr19: Fine Motor	0.728	0.756							0.553	
Eeersr20: Art	0.714	0.729								
Eeersr21: Music	0.580	0.474								
Eeersr22: Blocks	0.690	0.834								
Eeersr23: Sand/Water	0.572	0.624			0.416					
Eeersr24: Dramatic Play	0.651	0.834			0.484					
Eeersr25: Nature/Science	0.704	0.722								
Eeersr26: Math	0.692	0.645							0.779	
Eeersr27: Multimedia Use	0.593									
Eeersr28: Diversity Acceptance	0.492	0.524			0.533					
Program Structure										
Eeersr34: Schedule	0.595	0.415								
Eeersr35: Free Play	0.724	0.613								0.509
Eeersr36: Group Time	0.716	0.534					0.490			0.422
Personal Care										
Eeersr9: Greeting/Departing	0.505			0.414						0.495

Item	Single Factor Solution	Three Factor Solution			Six Factor Solution					
		Factor1	Factor2	Factor3	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Ecersr10: Meals/Snacks	0.553		0.669					0.559		
Ecersr11: Nap/Rest	0.491		0.516					0.411		
Ecersr12: Toileting	0.522		0.617					0.588		
Ecersr13: Health	0.538		0.620					0.672		
Ecersr14: Safety	0.505		0.705			0.584				
Language-Reasoning										
Ecersr15: Books	0.677	0.585								
Ecersr16: Child Communication	0.727	0.467		0.514			0.612			
Ecersr17: Language Reasoning	0.658			0.639			0.696			
Ecersr18: Informal Language	0.736			0.758			0.871			
Interaction										
Ecersr29: Gross Motor	0.608		0.402	0.467			0.561			
Ecersr30: General Supervision	0.744			0.651			0.754			
Ecersr31: Discipline	0.761			0.733			0.855			
Ecersr32: Staff-Child	0.750			0.842			0.983			
Ecersr33: Children Interactions	0.733			0.708			0.839			

Note. Values are the factor loadings from the EFA treating the items as ordinal. To reduce the volume of results, only factor loadings greater than .40 are presented. The remaining factor loadings are available from the authors.

Table 5

Standardized Coefficients from Regressions of Child Outcome and Alternative Quality Measures on ECERS-R Total Score and Three Quality Factors

	ECERS-R	Three Factor Scores		
	Total Score	SpcAct-Struct	Personal Care	Lang-Int
Child Outcomes				
Cognitive				
Reading composite	0.03	0.05	-0.01	0.06
Math composite	-0.04	-0.04	-0.01	-0.02
Socio-Emotional				
Social competence	0.06	0.06	0.06	0.05
Emotional and behavioral regulation	0.04	0.02	0.07 *	0.06 *
Attention and concentration	0.04	0.03	0.08 *	0.05
Health				
Child excellent health ^a	0.04 *	0.03	0.05 *	0.06 *
Child is not overweight ^a	0.01	0.01	0.02	0.01
No doctor verified respiratory illness ^a	0.02	0.02 *	0.02	0.01
No doctor verified gastrointestinal illness ^a	0.00	0.00	0.00	0.00
No doctor verified ear infection ^a	0.01	0.01	0.04	0.02
No injury that required doctor's visit ^a	-0.01	-0.01	-0.01	0.00
Alternative Measures of Quality				
Observer Measures				
Arnett Caregiver Interaction Scale	0.60 *	0.49 *	0.46 *	0.73 *
Observed group size	0.06	0.11 *	0.01	0.06
Observed child:teacher ratio	-0.16 *	-0.12 *	-0.14 *	-0.16 *
Provider-Reported Measures				
Teacher's educational level	0.11 *	0.10 *	0.03	0.15 *
Teacher's early childhood education credentials	0.14 *	0.17 *	0.09 *	0.13 *
Interest areas	0.35 *	0.40 *	0.21 *	0.25 *
Child-centered activities	0.30 *	0.37 *	0.19 *	0.21 *
Math activities	0.26 *	0.29 *	0.17 *	0.21 *
Language activities	0.12 *	0.13 *	0.03	0.13 *

Note. SpcActStructure=Space and Furnishings, Activities, and Program Structure. LangInt =Language-Reasoning and Interaction. The table summarizes results from 80 separate regression models, each with one of the ECERS-R quality measures as a predictor (columns) and one of the child outcomes or alternative measures of quality as an outcome (rows). For OLS models, values are standardized regression coefficients, adjusting for all of the control variables listed in the Appendix. For logit models, values are the change in predicted probability of a one on the outcome for a standard deviation increase in the ECERS-R score, holding all covariates (listed in the Appendix) at their means. All regressions are weighted by the ECLS-B child care observation ("CCO") sampling weight and standard errors are adjusted for clustering within ZIP Codes. $n = 1,150$ (rounded to nearest 50, as per ECLS-B data sharing agreement) for all variables except the child's math and reading composite and the child's weight status which are $n = 1,100$.

^a Logit regression.

*
p<.05 (one-sided p-values; all values except associations with group size, child:teacher ratio and school-readiness beliefs expected to be positive).