Identifying High Quality Preschool Programs: New Evidence on the Validity of the ECERS-R in Relation to School Readiness Goals

Rachel A. Gordon
University of Illinois at Chicago

Kerry G. Hofer
Vanderbilt University

Ken A. Fujimoto
Loyola University Chicago

Nicole Risk and Robert Kaestner
University of Illinois at Chicago

Sanders Korenman
Baruch College/CUNY

Author Note

Correspondence concerning this paper should be addressed to Rachel A. Gordon, Institute of Government and Public Affairs, University of Illinois at Chicago, 815 West Van Buren St., Suite 525, Chicago, IL 60607. Phone: 312-413-0295. Fax: 312-996-1404. E-mail: ragordon@uic.edu.

The authors gratefully acknowledge funding from the Institute of Education Sciences (IES), U.S. Department of Education, through Grants R305A090065, R305A130118, and R305K05186, and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), through Grant R01HD060711. We also acknowledge data sharing from the Center for Family Policy and Research at the University of Missouri and the excellent assistance of Kristin Abner in conducting our surveys of experts. We thank Clancy Blair, Jeanne Brooks-Gunn, and Lauren Wakschlag for their advice on the definition of domain-specific child care quality dimensions and the anonymous experts who rated the ECERS-R indicators. The opinions expressed are those of the authors and do not necessarily represent views of IES, NICHD, or our consultants.

We also gratefully acknowledge the secondary data sources used for this study which had already been collected for the following projects: 1) "Scaling up the Implementation of a Pre-Kindergarten Mathematics Curriculum in Public Preschool Programs" (University of California Berkeley, PI: Prentice Starkey, IES R305K050186), 2) "Evaluating the Effectiveness of Tennessee Voluntary Pre-K Program" (Vanderbilt University, PI: Mark W. Lipsey, IES R305E090009), 3) "Focus in Early Childhood Curricula: Helping Children Transition to School" (Vanderbilt University, PI: Dale Farran, IES R305J020020) and 4) "Quality Rating Systems Evaluation" (University of Missouri, PI: Kathy R. Thornburg).

This is an electronic version of an article published in *Early Education and Development, 26,* 1086-1110. Early Education and Development is available online at: http://www.informaworld.com with the open URL of http://dx.doi.org/10.1080/10409289.2015.1036348

Abstract

Research Findings: The Early Childhood Environment Rating Scale, Revised (ECERS-R) is widely used, often to evaluate whether preschool programs are of sufficient quality to improve children's school readiness. We examined the validity of the measure for this purpose. Item response theory (IRT) analyses revealed that many items did not fit together to measure single dimensions, particularly when rated by consultants as indicating aspects of quality relevant for multiple domains of child development. IRT results also conflicted with scale developers' expectations in terms of whether markers that they attached to higher response categories represented higher quality empirically. When reanalyzed based on experts' ratings, IRT results also showed relatively few indicators captured the moderate to high range of quality. *Practice or Policy:* Our results suggest that policymakers should carefully consider whether measures designed for specific purposes are appropriate for other high-stakes uses. We encourage continued refinement of existing quality measures, development of new measures, and the accumulation of evidence for their various uses.

Keywords: child care quality, Early Childhood Environment Rating Scale, ECERS-R, domain-specificity, measurement.

Identifying High Quality Preschool Programs:
New Evidence on the Validity of the ECERS-R in Relation to School Readiness Goals

In his 2014 State of the Union address, President Obama re-iterated his call for expanding access to "high-quality" preschool (White House, 2014a). The "Pre-K Now" initiative sponsored by the Pew Charitable Trusts similarly focused on advancing "high-quality, voluntary pre-kindergarten for all three- and four-year-olds" and helped support a doubling of state funding and a near doubling of children served by state pre-kindergarten (pre-k) throughout the first decade of the 21st century (Pew Charitable Trusts, n.d., 2012). The political viability of these initiatives is predicated on high quality preschool's ability to set children on an upward trajectory of learning and productivity. The President's Early Childhood Learning initiative argues that expanding access will "shape key academic, social, and cognitive skills that determine a child's success in school and in life" (White House, 2014b). The Pre-K Now initiative pointed to evidence that "high-quality pre-k is an essential catalyst for raising school performance" (Pew Center on the States, 2011). Paralleling these initiatives is a recent focus on increasing the quality of all types of early child care, not just preschool. Quality Rating and Improvement Systems (QRIS) for child care emerged for this purpose in the late 1990s and had spread to three-quarters of the states by 2014 (Child Trends, 2014). An umbrella organization -- the QRIS National Learning Network (2013) -- states that their aim is to "elevate the quality of care in state early care and education systems and to support and improve children's development."

Although it is sensible for state and federal governments to invest in high rather than low or mediocre quality programs, there is currently limited evidence that strategies and measures for supporting quality are valid for this purpose. One common approach has been to write into policy particular observational measures of the quality of early childhood classrooms, and to penalize or reward programs with certain scores on these measures. The Improving Head Start for School

Readiness Act of 2007 required Head Start grantees to recompete for funding when they scored below a specific cutoff on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008). Many state QRIS provide child care programs with higher child care subsidy reimbursements based, in part, on their scores on the CLASS or the Early Childhood Environment System Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998; Tout et al., 2010 and most state prekindergarten programs use the measures to monitor programs (Ackerman, 2014). In fact, by 2014, thirty state QRIS used the ECERS-R (alone or in combination with other measures; Child Trends, 2014); and, in 2012-2013, nineteen states used the ECERS-R for monitoring their state pre-k programs (Ackerman, 2014). Practices like these share similar challenges with other educational initiatives that try to use test scores to assure a return on taxpayer investments (Jennings, 2012; Nichols & Berliner, 2007). These uses are predicated on the assumption that ample evidence exists that the measures accurately indicate quality that supports policy goals.

In this paper, we offer new evidence regarding this assumption. We focus on the extent to which the ECERS-R' items capture environmental inputs that support child development, given public and policymaker support typically rests on investments in early care and education leading to these outcomes. In doing so, we compare the ECERS-R scale developers' original organization of the items to a reorganization based on experts' ratings of the items' relevance for several domains of child development. We highlight the degree to which the original structure mixes items relevant for different developmental domains and how such mixing affects the psychometric properties of the original scale scoring. In short, we address the following unanswered questions: To what extent do the ECERS-R and its subscales measure aspects of quality specific to child developmental domains? How does the current scoring procedure of the ECERS-R contribute to (or detract from) its domain-specificity?

Background on the ECERS-R

The ECERS-R was not designed specifically for its current use in QRIS and other policy and evaluation efforts. For instance, it wasn't designed with the one and only purpose of identifying certain aspects of quality that support children's school readiness, nor to be precise enough to support high stakes decisions regarding whether programs fall above or below certain cutoffs.

Rather, 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). First published in 1980, and revised in 1998 (Harms & Clifford, 1980; Harms et al., 1998), the measure reflects the early childhood education field's concept of developmentally appropriate practice, including: 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 et al., 1998). In an interview reflecting on the scale, Harms further said: "in order to provide care and education that will permit children to experience a high quality of life while helping them develop their abilities, a program must provide for the three basic needs of children: a) protection of their health and safety, b) building positive relationships, and c) opportunities for stimulation and learning from experience…It takes all three to create quality care. No one component is more or less important than the others, nor can one substitute for another" (Frank Porter Graham Child Development Institute, 1999, p. 3-4). Based on this holistic perspective, we anticipate a mixture of various aspects of quality within scale items.

Many ECERS-R items are also organized around the way child care center directors and teachers structure the care setting, reflecting the practitioner-focused origins of the scale. The scale developers note that this organization makes it easy for observers to "collect information that is likely to be found under similar circumstances" (Cryer, Harms, & Riley, 2003, p. xi). Appendix 1 lists the ECERS-R items, including subscales of *Space and Furnishings*, *Personal Care Routines*, *Activities*, and *Program Structure* that organize various aspects of quality within different areas of the classroom (indoor space, gross motor space, space for privacy), events of the day (meals/snacks, greeting/departing, nap/rest), activities (blocks, music, art), and time use (schedule, free play, group time). We anticipate that this organization around events of the day also makes it likely that items mix aspects of quality relevant for multiple domains of child development. As we explain further below, the scale asks observers to look for numerous features -- referred to as *indicators* – that are attached to the scores of each item. The brief item labels do not always fully signal all of the developmentally relevant content captured by these indicators. For example, Item 10 "Meals/snacks" contains not only indicators of nutrition and sanitation but also indicators of the amount of conversation that takes place during meals and the tone of staff-child interaction, overlapping the content signaled by the labels of other subscales such as *Language-Reasoning* and *Interaction*. In order to provide a systematic assessment of this item content, the first goal of our study was to ask experts to rate the indicators within every item in terms of their relevance for particular domains of child development.

The standard scoring procedure for the ECERS-R reinforces its holistic approach, and makes it difficult for researchers and policymakers to pull out specific aspects of quality and examine whether these most strongly support particular domains of development. Each item is scored on a scale with odd-value labels from *1 = Inadequate* quality to *3 = Minimal* quality to *5 = Good* quality to *7 = Excellent* quality. In the standard scoring, indicators for lower scores must

be met before indicators of higher scores are evaluated. That is, observers "stop scoring" when they reach a response category at which an indicator is not observed. This standard scoring approach reduces response burden, in that observers do not have to consider indicators above the stop point. It may also reflect a philosophical perspective that centers should not get credit for higher-level aspects of quality that they are doing well (e.g., being warm and responsive in their interactions with children) if they are not doing lower-level aspects of quality well (e.g., assuring basic cleanliness and safety), consistent with a desire to measure global quality of the child care environment (Clifford, Reszka & Rossbach, 2010; Cryer, et al., 2003).

The stop scoring was not based on empirical evidence, however. The sorting of indicators into items and the placement of indicators at different scale levels was based on the scale developers' understanding of quality, based on their experiences in classrooms and understanding of the literature (Clifford, Reszka & Rossbach, 2010; Cryer et al., 2003; Harms et al., 1998), rather than psychometric evidence. Specifically, there was not empirical evidence that indicators placed by scale developers at lower category levels (e.g., *1* and *3*) in fact reflected lower levels of an underlying dimension of quality than indicators placed at higher category levels (e.g., *5* and *7*). A second goal of our study was to test these assumptions.

The stop scoring approach challenges policymakers, practitioners, and researchers who wish to isolate particular quality components in order to examine how they intercorrelate (e.g., "Are some centers high in some aspects of quality, such as those that promote health and safety, but lower in others, like those that support language development?") and how they relate to child outcomes (e.g., "Do aspects of quality that experts rate to be highly supportive of language development correlate more highly with language outcomes than do health-specific aspects of quality?"). For policymakers, evaluators, and researchers interested in school readiness and child development, it is thus important to check whether alternative scoring approaches would produce

more domain-specific measures of quality, than the standard stop scoring approach. Doing so is consistent with attempts beyond the ECERS-R to consider whether and how to develop measures of child care quality specific to domains of child development (Burchinal, Kainz, & Cai, 2011; Forry, Vick, & Halle, 2009; Vandell & Wolfe, 2000; Zaslow, Halle, et al., 2006; Zaslow, Martinez-Beck, Tout, & Halle, 2011). A focus on domain-specificity can also be advantageous from a measurement perspective, since psychometricians recommend that measure development begin by carefully defining each dimension, differentiating the dimensions from one another, and writing of items specific to each dimension (Wolfe and Smith, 2007a, 2007b). Our third goal was to examine the possibility of reorganizing the indicators into more domain-specific sets based on expert ratings.

In short, an important issue that is in need of investigation is the extent to which the ECERS-R and its subscales measure aspects of quality specific to child developmental domains and how the current scoring procedures of the ECERS-R contribute to (or detract from) its domain-specificity. Our paper fills this gap by reporting expert ratings of the relevance of the ECERS-R indicators for particular aspects of child development, item response theory (IRT) tests of the ECERS-R's current structure, and IRT tests of potential new domain-specific structures identified by the experts.

Prior Research on the Validity of the ECERS-R

Most studies examining evidence for the validity of the ECERS-R have focused on its associations with child outcomes. Indeed, a large (but somewhat inconsistent) literature associates higher child care quality with better child outcomes, including quality as measured with the ECERS-R (Gormley, 2007; Vandell, 2004). A consensus among researchers has emerged, however, that effect sizes of associations between child care quality and child outcomes are small, especially when studies account for covariates (Besharov & Morrow, 2006;

Duncan & Gibson-Davis, 2006). For example, Burchinal, Kainz, and Cai (2011) meta-analyzed peer-reviewed journal articles that reported estimates of the association between quality measures and child outcomes. The authors found that partial correlations between the ECERS-R and children's cognitive and socioemotional outcomes ranged from .02 to .09 in absolute magnitude. Gordon, Fujimoto, Kaestner, Korenman and Abner (2013) likewise reported standardized coefficients below .10 when associating ECERS-R scores with child cognitive, socioemotional, and health outcomes in the Early Childhood Longitudinal Study, Birth Cohort.

A number of published articles have also used factor analyses to examine the conceptual structure of the ECERS-R and its subscales. These studies consistently identified three dimensions that generally combined the six ECERS-R subscales: (1) *Language-Reasoning* and *Interaction*, (2) *Space and Furnishings*, *Activities*, and *Program Structure*, and (3) *Personal Care Routines* (e.g., Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; Clifford et al., 2005; Gordon et al. 2013; Perlman, Zellman, & Le, 2004; Sakai, Whitebook, Wishard, & Howes, 2003). Effect sizes have been found to be modestly higher, but still small, within domains (e.g., Abner and colleagues, 2013, Burchinal and colleagues, 2011, and Gordon and colleagues, 2013, all found that the dimension combining *Language-Reasoning* and *Interaction* had partial correlations with reading comprehension below .10). However, these analyses relied on the 36 items shown in Appendix 1, rather than trying to repackage indicators into more domain-specific sets.

There have been few published IRT analyses of the ECERS-R, although scholars are increasingly calling for such approaches (Bryant, Burchinal, & Zaslow, 2011; Gordon et al., 2013). The IRT studies that have been conducted fail to confirm the category order imposed by the standard stop scoring approach. Lambert and colleagues (2008) analyzed the indicators for the *Language-Reasoning* subscale using data from 300 classrooms in Jamaica and Grenada.

They found evidence of indicator disordering (e.g., within an item, an indicator for a score of *7* was estimated to represent lower quality than an indicator of a score of *5*). However, they did not look at the other ECERS-R subscales, and of course it is unclear whether their results would generalize to the U.S. Gordon and colleagues (2013) conducted an item-level IRT analysis using the ECLS-B. They found at least one pair of disordered categories for every ECERS-R item. The current paper's indicator-level analysis tests Gordon and colleagues' conclusion that the item-level disorder might reflect the mixing of different aspects of quality among an item's indicators. The ECERS-R scale developers have recently used multiple methodologies, including IRT, to begin to develop new scoring approaches (Clifford, Sideris, & Neitzel, 2012); however, the recently released third edition of the measure retains much of its existing structure, including mixture of different aspects of quality within items and stop scoring (Harms, Clifford & Cryer, 2015). The standard stop-scored ECERS-R also remains widely written in state policy and is already embedded in numerous research studies and evaluations (Ackerman, 2014; Child Trends, 2014). Our study provides additional IRT evidence to inform current and future uses of the scale.

Summary and Research Questions

Our research extends prior examinations of the validity of the ECERS-R by assessing the domain-specificity and the conceptual order of its indicators. To examine our first research question – To what extent does the ECERS-R capture aspects of quality specific to particular domains of child development? – we asked experts to rate the indicators of the ECERS-R for their relevance to specific developmental domains. We also conducted an indicator-level IRT analysis of the ECERS-R, using data collected without the stop-scoring approach, in order to examine our second research question: Is there evidence that indicators attached to higher rating categories reflect higher quality? We also used the IRT analyses to address our third research question: Do the indicators better fit together as organized by expert ratings or the standard

ECERS-R subscales? Finally, we used the IRT results to graph the estimated location of the indicators on the underlying dimensions of quality, allowing us to examine our final research question: Does the empirical ordering facilitate conceptual interpretation of the new sets of indicators based on expert ratings?

Method

Survey of Experts

Sample. To obtain an impartial review of the items and go beyond the idiosyncratic views of our team, we asked several dozen experts to rate the relevance of the ECERS-R indicators for measuring aspects of quality that promote child development within several domains. We recruited experts in early childhood development and early childhood education by email through the first and second authors' professional networks. We focused on advanced graduate students (41%), post-docs (11%), and new assistant professors and practitioners (48%) whom we expected had the relevant expertise as well as the time and interest to complete the survey. We required experts to have at least a master's degree in early childhood education (36%), developmental psychology (20%), or a related field such as human development or education (44%), so that they would have a strong understanding of domains of child development and quality practices that might promote them. The sample included 76 experts (2 males and 74 females). Participants were compensated $60 for each survey that they completed.

Measures. We first identified three meta-domains – cognitive, socioemotional, and health -- and eight sub-domains – e.g., promote math skills, promote social competence, reduce injuries -- of child development often studied in child care research (see Appendix 2; Zaslow, Halle, et al., 2006; Zaslow, Martinez-Beck, et al., 2011). Since consensus definitions of domain-specific aspects of child care quality have not previously been published, we asked three senior consultants to help us define the aspects of quality relevant to these domains. The three

consultants were developmental psychologists, all senior scholars with expertise in at least one of the domains, and were not included in our 76 experts. Expert review studies often use such definitions to guide ratings (Wolfe & Smith, 2007a, 2007b), and doing so allowed us to prime the experts' general content knowledge by making the domain definitions explicit.

As shown in Appendix 1, the first six ECERS-R subscales contain 383 indicators in total. Like other researchers, we omitted an additional seven ECERS-R items, primarily from the *Parents and Staff* subscale, from all analyses. In order to reduce response burden and cost, we further excluded 129 of the ECERS-R indicators from our expert survey. We followed Hofer (2008) who created a shortened version of the ECERS-R by first removing clearly redundant indicators (e.g., retaining "Blocks and accessories accessible for daily use" and removing "Few blocks are accessible for children's play") and then removing indicators that experts in her study identified as not at all relevant to child care quality. Our study differs from Hofer's by asking about domain-specific rather than general relevance to quality. We also reduced response burden by splitting the remaining 254 indicators into four separate surveys that followed the same structure, but contained just one of the four sets of indicators (62 to 70 indicators per survey). Our goal of at least 30 expert ratings was met for all surveys (53% of experts rated one set of indicators, 36% two sets, and 12% three sets).

Each expert was asked to rate each of the indicators as: *Not at all relevant*, *Only indirectly relevant*, *Somewhat relevant*, or *Highly relevant* for each domain. We dichotomized expert ratings into relevant or not relevant (combining the *somewhat* and *highly* relevant response categories and the *not at all* and *only indirectly* relevant response categories, respectively). Across the eight sub-domains the percentage of ratings with a missing expert rating ("I cannot adequately rate this item") ranged from 1-2% and the percentage with two missing ratings was less than 0.5%. No indicators had more than two missing ratings.

We defined expert agreement based on whether at least half of the experts rated an indicator as relevant for any of the domains, for more than one domain, and for a particular combination of domains (details and sensitivity analyses of other criteria for defining agreement are available from the authors). We generally repeated these analyses for the eight sub-domains and the three meta-domains; however, we focused our presentation of combinations on the three meta-domains (for which there were seven possible combinations of one, two or all three meta-domains whereas there were 254 possible combinations of the eight sub-domains).

Observations of Child Care Centers

Sample. For the IRT analyses, we combined two secondary datasets in which raters had scored all of the ECERS-R indicators rather than using the stop-scoring rule. Investigators from the University of California Berkeley and Vanderbilt University gathered data in 122 early childhood classrooms (52 public classrooms and 66 Head Start classrooms) as part of an early math curriculum evaluation project. Investigators from the University of Missouri gathered 160 observations in Head Start classrooms as part of a state quality rating systems pilot study. When we pooled these data sets, we had a total of 282 observations.

Measures. These studies had gathered all ECERS-R indicators without following the standard ECERS-R stop-scoring rules. Observers used the standard ECERS-R scoring sheets but checked every indicator within every item as either "Yes" observed or "No" not observed. We assigned values to the observers' check marks so that a higher score represented a more positive attribute. In other words, all positively oriented indicators (e.g., "Good ventilation, some natural lighting through windows or skylight") were assigned a value of *1* if checked "Yes" and all negatively-oriented indicators (e.g., "Insufficient space for children, adults, and furnishings") were assigned a value of *1* if checked "No." We focused our IRT analyses on 365 of the 383 indicators from the first six ECERS-R subscales (see again Appendix 1). The 15 excluded

indicators had substantial missing data because observers chose the "Not Applicable" option (e.g., indicators about enrolled children with disabilities). Additionally, we excluded two indicators because they had no variation in responses (and thus difficulty levels and standard errors could not be estimated) and one indicator that produced lack of convergence in initial models.

Analytic Approach

We calculated percentages based on the expert ratings and conducted IRT analyses of the observational data. For the IRT approach, we used the Rasch item bundle model (RIBM; Wilson & Adams, 1995) estimated in Stata 11.0 and Conquest 2.0 (Wu, Adams, Wilson, & Haldane, 2007). The traditional Rasch model connects the probability of a dichotomous item (the ECERS-R indicators in our case) being scored "Yes" with the level of the latent construct (the classroom's quality level in our case; see de Ayala, 2009; Gordon, 2015 for accessible introductions). The traditional Rasch model assumes these probabilities are independent across items, conditional on the item difficulties and latent trait levels. This assumption is violated when items are clustered, such as clustering of sets of indicators together on the same scoring page in the ECERS-R. The RIBM accounts for this type of non-independence. Our initial results confirmed that the RIBM model fit better than a traditional Rasch model (results available from the authors). Because of the very large number of indicators, we conducted a separate RIBM analysis within the six ECERS-R subscales and within each set of indicators identified by experts.

We calculated several statistics from the RIBM. One set of statistics tested whether there was empirical evidence for the ECERS-R authors' placement of indicators in rating categories. The RIBM estimated indicator *difficulty* levels, which is the estimated location on the latent quality construct where a classroom had a 50:50 chance of being rated "Yes" for that indicator. If

the standard scoring was consistent with empirical ordering then these locations should follow the order shown in the ECERS-R manual. That is, indicators placed at the rating category of a *1* should be positioned lower – be "easier" to observe -- than those of a *3*; those placed at a *3* should be easier than those placed at a *5*; and those at a *5* should be easier than those at a *7*. We used 99% confidence intervals of the indicators' difficulty estimates to check whether the empirical order followed the ECERS-R authors' order. We defined *ordered* indicators if the indicators placed by the ECERS-R authors at a lower rating category were empirically estimated to be easier than indicators that the ECERS-R authors had placed at a higher rating category, incorporating error of estimation in the confidence interval (i.e., the upper bound of the lower indicator was below the lower bound of the higher indicator). We defined *overlapping* indicators if the confidence intervals of the two indicators overlapped. We defined *disordered* indicators if the indicators placed by the ECERS-R authors at a lower rating category were empirically estimated to be harder than indicators that the ECERS-R authors had placed at a higher rating category (i.e., the lower bound of the lower indicator was above the upper bound of the higher indicator).

To examine whether indicators worked together better in the six ECERS-R subscales or the dimensions identified by the experts, we used indicator fit statistics from the RIBM analysis. High positive fit values corresponded to an indicator receiving ratings that were unexpected (e.g., the model predicted a high probability of a rating of "Yes" but the observed rating was "No"). Although there are different reasons for high fit statistics, one reason for such unexpected responses could be because a set of items does not measure a single dimension. Analysts sometimes attempt to improve item sets by iteratively removing the worst fitting indicators until fit criteria are met (e.g., targeting mean square values based on the sum of the squared residuals that are greater than two when standardized; Bond & Fox, 2009; Linacre, 2012). We examined

the relative performance of the subscales by comparing the number of indicators that misfit on the initial run and the number of indicators removed through this iterative process.

The RIBM model also estimates the location of each preschool classroom on each underlying dimension in the same units as the item locations. Because the Rasch model is based on the logistic distribution, the units are the log of the odds (the ratio of the probability of success to the probability of failure), also known as *logits*. We created item-classroom maps to illustrate how well the indicators were targeted at the preschool classrooms in our sample (i.e., looking to see whether the indicators covered well the full range of the underlying quality dimension reflected in the centers). The maps also showed the empirically-estimated order of the indicators, and we demonstrate below how these maps can be used to verify a priori expectations about how the indicators should be arrayed and to inform post-hoc re-interpretations of quality domains.

Results

Survey of Experts

Given that we focused on the 254 indicators that Hofer's experts (2008) had already identified as generally relevant to child care quality, it is not surprising that every indicator was rated as relevant to at least one of the eight sub-domains (and consequently at least one of the three meta-domains) by at least half of the experts. However, experts generally differed in terms of which domains the attribute supported; and, where experts agreed, the results were typically domain-general (i.e., indicators rated as relevant to more than one domain) rather than domain-specific (i.e., indicators rated as relevant to just one domain). When we looked across all eight domains, for only six indicators did the majority of experts rate the attribute as relevant to the same domain. In all six of these cases, the indicators dealt with sanitation (e.g., handwashing, exclusion of sick children) and the experts agreed the attribute was relevant only to reducing the

spread of illness. Agreement happened more often at the coarser meta-domain level, but it was still the case that for just 13% of indicators did experts agree about relevance to a single meta-domain. For 61% of indicators there was also agreement, but for relevance to two or three meta-domains. For the remaining quarter of the indicators, no combination of meta-domains was agreed upon by at least half of the experts.

Looking more specifically at the indicators where experts agreed, we identified four specific combinations of meta-domains. The majority (57%) of agreed-upon indicators were rated as relevant to both cognitive and socioemotional development. Just over one quarter (26%) were rated as relevant to all three meta-domains. These indicators generally came from across the ECERS-R subscales. Examples of indicators rated as relevant to both cognitive and socioemotional development included "Staff read books to children informally," "Staff usually respond to children in a warm supportive manner," and "Some opportunity for children to be a part of self-selected small groups." Examples of indicators rated as relevant to all three meta-domains were "Most staff sit with children during meals and group snacks," "Children taught to manage health practices independently," and "Staff explain reasons for safety rules to children."

In contrast, fewer indicators were rated as relevant to just one meta-domain, and these indicators were concentrated within subscales. Specifically, at least half the experts agreed that 17 indicators were relevant only to the Cognitive meta-domain, with nearly all of these indicators coming from the *Activities* subscale and the remainder coming from the *Language-Reasoning* subscale. Examples were "Books organized in a reading center," "Blocks and accessories accessible for daily use," and "Daily activities used to promote math/number learning." At least half of the experts agreed that an additional 16 indicators were relevant just for the Health meta-domain, and nearly all of these indicators came from the *Personal Care Routines* subscale with the remaining indicator coming from *Space and Furnishings*. Examples included "All furniture

is sturdy and in good repair," "Sanitary conditions usually maintained," and "Staff and children wash hands most of the time after toileting." No indicators were rated as relevant only to the Socioemotional meta-domain.

Observations of Child Care Centers

Indicator Ordering. We now turn to our RIBM analyses, where we first tested whether the empirical estimates of indicators' positions on the latent construct were consistent with the scale developers' placement of the indicators in the standard scoring sheets. Table 1 presents the number of indicators, number of comparisons between adjacent higher and lower indicators, the percentage of comparisons that were ordered, and the percentage of comparisons that were not ordered (overlapping or disordered). Recall that we relied upon 99% confidence intervals to make these determinations, and that ordered comparisons would be consistent with the scale developers' arrangement of the indicators within the items' rating categories whereas overlapping or disordered categories would be inconsistent with those arrangements. We reported the total across all indicators in the first row and results within the ECERS-R scale developers' subscales in the remaining rows.

Beginning with the total results (first row), of the 1,386 adjacent comparisons among 365 indicators, 56% were ordered and 44% were not ordered. The largest fraction of those that were not ordered had overlapping confidence intervals. Seven percent of all comparisons were out of order. Across the six ECERS-R subscales, order was most evident for *Activities* (70% ordered) and *Language-Reasoning* (68% ordered) followed by *Space and Furnishings*, *Interaction*, and *Program Structure* (55%-57% ordered). Order was least evident for *Personal Care Routines* (37% ordered). Fully one-fifth of the comparisons for *Personal Care Routines* were disordered. At the item level (not shown), we similarly found that within *Personal Care Routines*, the majority of comparisons were not ordered for almost every item. Although the majority of the

comparisons were ordered for items in other subscales, we importantly saw that every ECERS-R item had at least one adjacent comparison that was not ordered (overlapping or disordered); fully two-thirds had at least one adjacent comparison that was disordered.

We used a series of charts in Figure 1 to present visually examples of indicators that had mostly ordered categories, mostly overlapping categories, and mostly disordered categories. Along the horizontal axis of each chart, we grouped the indicators within the item score category that they were listed under on the ECERS-R score sheet (1, 3, 5, or 7). On the vertical axis, we charted the item difficulty based on our RIBM analyses. Indicators found higher on the vertical axis corresponded to "harder" indicators, i.e., those that raters were less likely to endorse. We used vertical lines to represent the 99% confidence interval for the estimate of an indicator's item difficulty, making it easy to see which indicators were ordered, overlapping or disordered. To facilitate these comparisons, we ordered the indicators within each category (1, 3, 5 and 7) by the point estimates of their item difficulties. A short label for each indicator was provided in the legend.

If the indicators were ordered, then they should rise in groups from left to right. That is, the indicators under a 1 should all have the lowest item difficulties, the indicators under a 3 the next highest, the indicators under a 5 somewhat higher, and the indicators under a 7 should have the highest item difficulties on the vertical scale. We presented in Figure 1a the results for the indicators of ECERS-R Item 23 "Sand/Water", which was one of the items with a high percentage of ordered indicators (77%). Although some overlap existed among indicators under neighboring scores (e.g., indicators "Sand/water play available" and "Different activities"), the general trend of the indicators as a group for this item increased in terms of difficulty as the score category increased.

We presented an example of an item with numerous overlapping indicators in the middle

chart, Figure 1b. This chart shows the indicators for ECERS-R Item 14 "Safety Practices" which had the highest percentage of overlapping indicators (68%). In this case, many indicators fell near the middle of the vertical axis across the rating categories of 1, 3, 5 and 7, with overlapping confidence intervals. The hardest item in this set was also disordered, coming from the "Score 3" ("No safety hazards").

Finally, we showed the indicators for the ECERS-R Item 10 "Meals/Snacks" in the bottom chart, Figure 1c. For this item, a near majority of comparisons showed lack of order (49%, or 36 out of 73 comparisons) including 30% overlapping and 19% disordered. In this case, a number of the indicators for the score of 1, 3, and 5 were positioned near the low end (about –3 or below on the logit scale) with confidence intervals again showing substantial overlap. Of note, the larger confidence intervals for these indicators reflect the fact that these indicators were very easy (the vast majority of preschool classrooms in our sample met the conditions for these indicators, thus creating very large uncertainty in their estimates; Hedeker & Gibbons, 2006, p. 358). Some indicators for the scores of 3, 5, and 7 appeared in the mid-range of the graph (about –1 to –3 on the logit scale). The most difficult indicators fell at the scores of 1 and 3 (near zero on the logit scale) and dealt with maintaining sanitary conditions (e.g., whether most adults and children washed their hands before meals and snacks). These indicators were harder than the indicators for a 7 (which captured whether the children helped during meals and used child-size utensils and whether meals were a time for conversation).

Fit of indicators. We now turn to our analyses of how well the indicators fit together to define each dimension. In Table 2, we summarized the number of misfitting indicators from our initial analysis as well as the number of indicators that we iteratively removed before the final RIBM analysis. In the top panel, we show results from our analyses of the original ECERS-R subscales. In the bottom panel, we show the results from our analyses of the four combinations

of meta-domains based on the expert surveys. In both cases, the analyses did not force the indicators to be ordered as in the ECERS-R standard scoring, but rather tested whether each set of items fit together to define a single dimension, given their empirical ordering.

The empirical results confirmed the experts' ratings: IRT results showed that fit to a single dimension was best for indicators that raters agreed measured a single meta-domain. Beginning by looking at the ECERS-R scale developers' original organization of the items into six subscales, we found that fit was worst for the *Activities* and *Program Structure* subscales where nearly half or more of the indicators were removed during the refitting process. In contrast, about one-quarter of the *Program Structure* indicators and close to 10% of the remaining three subscales had indicators removed. When we reorganized the items into sets based on the expert ratings, we found that a similar fraction – 14% - were removed when we focused on the items that experts agreed were relevant for two (Cognitive-Socioemotional) or all three (Cognitive-Socioemotional-Health) meta-domains. In contrast, all but one of the items that experts rated as relevant to only one meta-domain fit together.

Targeting of Indicators. In our final analysis, we considered how well the fitting items covered the dimensions and were targeted at the sampled classrooms. We showed in Figures 2 and 3 item-classroom maps for the expert-identified Health and the Cognitive meta-domains, respectively. We focused on these two dimensions because experts agreed the indicators were domain-specific and the indicators were of a manageable number and showed good fit.

In each figure, the distribution of classrooms appeared on the left (the Xs, each of which represent between 1 and 2 classrooms). The items appeared on the right (with labels that connect back to the items and indicators (e.g., 13.3.1 means the 1st indicator of Category 3 on Item 13) and provides the indicator label, abbreviated for display. The display was arrayed so that the bottom represented classrooms with relatively less of the latent quality construct and represented

items that were relatively easier for observers to rate (were relatively frequent). The top reflected relatively more of the construct and relatively harder (or rarer) items. The scale on the far left was the common logit scale used by the RIBM model to estimate the position of each item and classroom. The absolute location of this scale is arbitrary, and each display centers the classrooms at a mean of zero. Therefore, all item and person locations were relative to the classroom mean on a particular aspect of quality (health or cognitive).

Ideally, the indicators would be distributed from the highest to the lowest level of preschool quality in order to maximize variability in our measures of classroom quality. Instead, Figure 2 showed that the Health indicators were concentrated in the lower half of our preschool classrooms' measures. The hardest indicator ("Adequate handwashing") was located at about a logit value of zero, with about half of sampled classrooms being located above this value. Additional health-specific items that were harder would be needed in order to better distinguish between preschool classrooms in this moderate to high health-specific quality range. We also saw at the low end of the distribution that some indicators were so easy that they offered little information: the items regarding smoking and some procedures to minimize spread of disease fell below the classroom with lowest quality in the sample. Sturdy furniture of good repair and essentials for emergencies were likewise estimated to be at the lowest extreme of the distribution. These items are typically covered in contemporary licensing and regulation standards, and to the extent that variation in the full range is important for policy, the measurement of health-specific quality could be improved by replacing these indicators with indicators somewhat higher (e.g., in the -2.5 to -3.5 logit range) or in the middle to upper end of the distribution.

The item-classroom map for the Cognitive meta-domain was somewhat different. The indicators were again predominately in the low to middle end of the latent dimension, although

two items fell at the higher end (math and number materials are many and varied and are accessible for a substantial portion of the day). Reviewing the labels in the figure reveals that the indicators that experts agreed were specific to supporting cognitive development primarily captured the number, variety and organization of materials and activities, rather than teacher-child interactions, coming from four items on these topics (Item 15 Books and pictures, Item 22 Blocks, Item 25 Nature/science and Item 26 Math/number). Like Figure 1, the item-classroom map in Figure 3 also revealed the difference between the empirical ordering of the indicators and the scale developers' original placement. For instance, an indicator that was placed at rating category 5 by the scale developers – "15.5.3 Books organized in a reading center" – was one of the easiest items relative to this sample of classrooms. On the other hand, other indicators of the 5th rating category were the hardest, harder than indicators from the 7th rating category (e.g., the hardest indicator in Figure 3 was Indicator 5.1 of the "Math/number" item, "26.5.1 Many developmentally appropriate materials of various types accessible," harder than Indicator 7.2 of this same item, "26.7.2 Materials are rotated to maintain interest").

Discussion

The goal of this paper was to extend prior research on the validity of the ECERS-R by assessing the domain-specificity and the conceptual order of its indicators. We found that experts agreed that the ECERS-R indicators were relevant to at least one domain of child development, consistent with Hofer's (2008) earlier results. We found little domain-specificity for these indicators, however, as most were rated as relevant to multiple domains. A subset of about three-dozen indicators was rated as relevant only to the Cognitive or only to the Health meta-domains, being drawn primarily from the *Activities* subscale in the former case and the *Personal Care Routines* subscale in the latter case. No indicators were rated as relevant only to the Socioemotional meta-domain, although most indicators were rated as relevant to that meta-

domain in combination with the Cognitive, and sometimes also the Health, meta-domain.

Our results also revealed that indicators best fit together to measure a single dimension when they were repackaged into sets that experts agreed were relevant to single domains of child development. Indicators that experts felt were relevant to two or three meta-domains of child development fit together less well, as did indicators of the original ECERS-R subscales. Indeed, fully half of the indicators of the ECERS-R *Activities* subscale did not fit together to define a single underlying dimension, nor did nearly half of the *Program Structure* indicators and nearly one-quarter of the *Interaction* indicators. We also found that the indicators that experts agreed measured aspects of quality relevant for cognitive development better covered the underlying dimension than did the indicators that experts agreed measured aspects of quality specific to health, although harder indicators of both dimensions would help increase variation in quality scores.

Given the number of early education programs relying on ECERS-R data for a variety of consequential decisions, the general lack of domain-specificity is discouraging. This is especially so from the perspective of scholars and policymakers who are attempting to define more narrowly aspects of quality relevant for particular domains of child development (Burchinal, Kainz, & Cai, 2011; Forry, Vick, & Halle, 2009; Vandell & Wolfe, 2000; Zaslow, Halle, et al., 2006; Zaslow, Martinez-Beck, Tout, & Halle, 2011). For example, some scholars have been pursuing the possibility that the predominant use of global rather than domain-specific measures of quality may be one reason that correlations between measures of child care quality and child outcomes are small in magnitude. Policymakers also frequently bank their investments in early childhood programs on closing achievement gaps between more and less advantaged children (Fuller 2007; Laosa & Ainsworth, 2007). One way to assure that programs are meeting these goals is to monitor their quality specific to promoting particular aspects of school

readiness, like language skills, math skills, and a positive approach to learning. Our results suggest that one measure that has been widely used in the literature and in state QRIS, the ECERS-R, generally does not provide this domain-specificity.

Lack of domain-specificity is less problematic from other perspectives, however, including developmentally appropriate practice, the philosophy on which the ECERS-R was based; the broader holistic approach to early childhood intervention that has been a hallmark of many programs, including Head Start; and, "domain-general" (rather than domain-specific) perspectives on development, especially the ways in which emotional maturity supports cognitive development (and cognitive maturity supports emotional development; Arnold & Doctoroff, 2003; Denham, 2006). From these perspectives any attempts to define aspects of child care quality that specifically support one domain of development may be seen as unproductive. Our finding that experts saw most indicators as relevant for both cognitive and socioemotional development (and sometimes also for health) is consistent with these perspectives. On the other hand, our experts did agree on a small subset of indicators that tapped aspects of quality relevant just for cognition or for health. The latter findings suggest the potential for additional efforts to define and measure domain-general and domain-specific aspects of quality. Such efforts might help the field distinguish the extent to which low effects sizes in associations of quality with child outcomes reflects this aspect of measurement versus other issues (e.g., children's differential exposure to quality settings, families' non-random selection into child care settings, varying validity of outcome measures; Burchinal, Kainz, & Cai, 2011; Hofer, 2008, Zaslow et al., 2006).

Our results generally exemplify the ways in which researchers can use psychometric tools to inform scale development. IRT approaches, for example, have been successfully applied in other areas of developmental research (e.g., DeRoos & Allen-Meares, 1998; Dunn & Dunn,

2007; Piquero, Macintosh, & Hickman, 2002; Rapport, LaFond, & Sivo, 2009), but have been utilized less frequently in the development of child care quality measures. Here we showed that over one-quarter of the ECERS-R indicators did not fit together with other indicators to measure their respective subscales. Fit was best for a subset of about three-dozen items that experts agreed measured only one domain of quality. The approach we used here to analyze the ECERS-R post-hoc can be used even more productively during scale development. That is, scholars can define and differentiate dimensions of quality, develop item pools, ask experts to rate items on their relevance for these dimensions, and then collect data and analyze relevant items to winnow out those that do not work together empirically (Wolfe & Smith 2007a, 2007b). The items that fit together can be examined in maps, such as those we show in Figures 2 and 3, to identify gaps where adding additional items would increase variation in the resulting measure. Maximizing variation in child care quality in this way would increase precision in regression models predicting child outcomes, another factor that may be attenuating associations.

We also used IRT approaches to test a unique aspect of the ECERS-R scale: the order implicit in its standard stop-scoring approach. This test is important, especially in high stakes contexts where the ECERS-R has been used, such as in QRIS. Providers and advocates have expressed concern that their scores sometimes do not reflect their true quality because failure to meet indicators at the low end of the scale, especially health and safety items, prevents the possibility of being rated on indicators at the higher end of the scale, especially on caregiver-child interactions (Zellman & Perlman, 2008). Indeed, Hofer (2008, 2010) found 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 method was used. In this paper we used a different technique – IRT – to demonstrate that an item's indicators do not always follow the order assumed by stop scoring. Overall, more than two-fifths of indicator comparisons

lacked order. Importantly, every ECERS-R item had at least one pair of adjacent indicators whose positions either overlapped or were out of order; fully two-thirds had at least one pair of adjacent indicators that were out of order.

Lack of order was greatest in the *Personal Care Routines* subscale which contains items that follow the structure that practitioners have complained about, where failing health and safety indicators at Score 1 or 3 prevents their receiving credit for other indicators at higher categories. The examples we showed in Figures 1b and 1c revealed that, although the scale developers placed indicators of handwashing and safety hazards in the lower score positions of stop scoring, these were the hardest indicators on the item, harder than indicators of other aspects of quality which scale developers placed in higher score positions (such as positive caregiver-child interactions and children participating in the activity). These results provide evidence in support of Gordon and colleagues' (2013) conclusion that the mixing of different aspects of quality combined with the stop scoring approach likely produced the category disorder evident in their item-level analyses. Tests of order could also be informative to future scale development, even when stop scoring is not used, to the extent that theories and concepts are used to develop items and categories thought *a priori* to be positioned at low, medium, and high levels on a dimension.

Our study has limitations. Three consultants helped us define domain-specific aspects of quality and 76 experts rated the relevance of the ECERS-R indicators for these domains. Using different consultants and experts or using different domain definitions might have led to different results. To reduce response burden, we also asked experts to rate a subset of 254 of the 383 indicators of the first six ECERS-R subscales. Although we focused on indicators that another set of experts rated as relevant to child care quality in general (Hofer 2008), our findings may have differed if we had asked experts to rate all indicators. It is also the case that our RIBM results showed that over one-third of confidence intervals overlapped, and we might have been

able to identify more ordered -- and more disordered -- adjacent indicators with a larger sample size (since standard errors of item difficulty estimates would be smaller in larger samples; Hedeker & Gibbons, 2006). At the same time, the numerous comparisons that we conducted were not independent. We used 99% rather than 95% confidence intervals, but some differences may reflect sampling error. Our results also do not necessarily generalize beyond the nearly 300 preschool classrooms in our datasets.

We also focused on the Rasch model, which makes important assumptions, including that the strength of associations between a center's underlying latent quality level and its probability of receiving a yes score is the same across indicators (i.e., that the "discrimination" parameter is constant across indicators). An advantage of this assumption is that it simplifies interpretation by assuring that the ordering of items on the latent dimension does not vary depending on center quality and therefore allowing for the kinds of maps that we showed in Figures 2 and 3. We hope our study will encourage more IRT analyses of the ECERS-R, including studies that rely on other models that make different assumptions (e.g., see Gordon, 2015 for an introduction and review).

Another important limitation of the study is that we relied on expert ratings of the relevance of indicators for aspects of child development, but could not directly look at the indicators' validity in relation to school readiness. In other words, to fully test domain-specificity we would ideally have been able to correlate child outcomes with the dimensions we created by repackaging the indicators based on our experts' ratings. Unfortunately, child outcomes were not available in the datasets that we used, and examining such correlations would be an important direction for future research.

With these limitations in mind, our findings support current efforts by the ECERS-R scale developers to develop new scoring approaches for the ECERS-R and the ECERS-3

(Clifford et al., 2012; Harms et al., 2015). In the meantime, we recommend that users collect data on all indicators to allow for additional tests of the order of the instrument's indicators and examination of whether subsets of indicators selected to measure particular aspects of quality work together and correlate specifically with child outcomes. Relevant to both future revisions of the ECERS-R, and to other measure development, our results demonstrate the utility of measurement models, including IRT models, to evaluate measures of child care quality and provide some support both for domain-specificity and domain-generality of child care quality. These findings reinforce the importance of current efforts in the field to refine existing quality measures and to examine them with a broad array of psychometric tools (Forry, Vick, & Halle, 2009; Gordon et al., 2013; Pianta, LaParo, & Harms, 2009; Sylva et al., 2006; Zaslow, Martinez-Beck, Tout, & Halle, 2011;).

References

Abner, K.S., Gordon, R.A., Kaestner, R., & Korenman, S. (2013). Does child-care quality mediate associations between type of care and development? *Journal of Marriage and Family, 75*, 1203-1217.

Ackerman, D.J. (2014). *State-Funded PreK Policies on External Classroom Observations: Issues and Status*. Princeton, NJ: Educational Testing Service. Available at http://www.ets.org/Media/Research/pdf/PIC-STATE-PRE-K.pdf

Arnold, D.H. & Doctoroff, G.L. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology, 54,* 517-545.

Besharov, D.J., & Morrow, J.S. (2006). Introduction - Rethinking child care research. *Evaluation Review, 30*(5), 539-555.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd Ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bryant, D.M., Burchinal, M., & Zaslow, M. (2011). Empirical approaches to strengthening the measurement of quality: Issues in the development and use of quality measures in research and applied settings. Pages 33-47 in Zaslow, M., Martinez-Beck, I., Tout, K. & Halle, T. (Eds). *Quality Measurement in Early Childhood Settings*. Baltimore, MD: Brookes Publishing.

Bryant, D.M., Clifford, R.M., & Peisner, E.S. (1991). Best practices for beginners: Developmental appropriateness in kindergarten. *American Educational Research Journal, 28*, 783-803.

Burchinal, M., Kainz, K. & Cai, Y. (2011). 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. Pages 11-31 in Zaslow, M., Martinez-Beck, I., Tout, K. & Halle, T.

(2011). *Quality Measurement in Early Childhood Settings*. Baltimore, MD: Brookes Publishing.

Cassidy, D.J., Hestenes L.L., Hegde A., Hestenes S., & Mims S. (2005). 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, 20*, 345–360.

Child Trends. (2014). QRIS Compendium. Available at http://qriscompendium.org/

Clifford, R.M., Barbarin, O., Chang, F., Early D., Bryant D., Howes, C., … Pianta, R. (2005). What is pre-kindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science, 9(*3*),* 126-143.

Clifford, R.M., Sideris, J., & Neitzel, J. (2012). *New Scoring Mechanisms for the ECERS-R*. Paper presented at NAEYC's 21st National Institute for Early Childhood Professional Development in Indianapolis, IN (June 10-13, 2012).

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

Cryer, D. (1999). Defining and assessing early childhood program quality. *Annals of the American Academy of Political and Social Science, 563*, 39-55.

Cryer, D., Harms, T., & Riley, C. (2003). *All about the ECERS-R*. Lewisville, NC: Kaplan.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Denham, S.A. (2006). Social-emotional competence as support for school readiness: What is it and how to we assess it? *Early Education and Development, 17*, 57-89.

DeRoos, Y., & Allen-Meares, P. (1998). Application of Rasch analysis: Exploring differences in

depression between African-American and white children. *Journal of Social Service Research, 23,* 93-107.

Duncan, G.J., & Gibson-Davis, C.M. (2006). Connecting child care quality to child outcomes: drawing policy lessons from nonexperimental data. *Evaluation Review, 30*(5), 611-630.

Dunn, L.M. & Dunn, D.M. (2007). *Peabody Picture Vocabulary Test-Fourth Edition Manual*. Minneapolis, MN: Wascana.

Forry, N., Vick, J., & Halle, T. (2009). Evaluating, developing, and enhancing domain-specific measures of child care quality (OPRE Research-to-Policy Brief #2). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Frank Porter Graham Child Development Institute. (2003). *Early developments*. Chapel Hill, NC: The University of North Carolina at Chapel Hill.

Gordon, R.A. (2015). Measuring constructs in family science: How can IRT improve precision and validity? *Journal of Marriage and Family*, *77*, 147-176.

Gordon, R.A., Fujimoto, K., Kaestner, R., Korenman, S. & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for assessments of child care quality and its relation to child development. *Developmental Psychology, 49*, 146-160.

Gormley, W.T. (2007). Early childhood care and education: Lessons and puzzles. *Journal of Policy Analysis and Management, 26*(3), 633-671.

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

Harms, T., Clifford, R.M., & Cryer, D. (2015). *Early Childhood Environment Rating Scale, Third Edition*. New York: Teachers College Press.

Hedeker, D. & Gibbons, R.D. (2006). *Longitudinal Data Analysis*. New York: Wiley.

Hofer, K.G. (2008). Measuring quality in prekindergarten classrooms: Assessing the Early Childhood Environment Rating Scale (Doctoral dissertation, Vanderbilt University, 2008). *Dissertation Abstracts International*, 70A (11).

Hofer, K.G. (2010). How measurement characteristics can affect ECERS-R scores and program funding. *Contemporary Issues in Early Childhood*, 11 (2), 175-191.

Jennings, J. (2012). The effects of accountability system design on teachers' use of test score data. *Teachers College Record, 114*, 1-23.

Lambert, M.C., Williams, S.G., Morrison, J.W., Samms-Vaughan, M.E., Mayfield, W.A., Thornberg, K.R. (2008). 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, 16,* 41–60.

Laosa, L.M. & Ainsworth, P. (2007). Is public pre-k preparing Hispanic children to success in school? *NIEER Preschool Policy Brief*, Issue 13 (March).

Linacre, J.M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Nichols, S.L. & Berliner, D.C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge: Harvard Education Press.

Perlman, M., Zellman, G.L., & Le, V. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly, 19,* 398-412.

Pew Center on the States. (2011). Transforming public education: Pathway to a pre-k-12 future. Retrieved from http://www.pewstates.org/uploadedFiles/PCS_Assets/2011/Pew_PreK_Transforming_Public_Education.pdf

Pew Charitable Trusts. (n.d.). Pre-k now. http://www.pewstates.org/projects/pre-k-now-328067

Pew Charitable Trusts. (2012). Proof into policy: Pre-k milestones. Retrieved from http://www.pewstates.org/research/data-visualizations/proof-into-policy-pre-k-milestones-85899376577

Pianta, R.C., LaParo, K.M., & Harms, B. (2009). *Classroom Assessment Scoring System Manual Pre-K*. Baltimore, MD: Brookes.

Piquero, A.R., Macintosh, R., & Hickman, M. (2002). The validity of a self-reported delinquency scale: Comparisons across gender, age, race, and place of residence. *Sociological Methods and Research, 30,* 492-529.

QRIS National Learning Network. (2013). Retrieved from http://qrisnetwork.org/

Rapport, M.D., LaFond, S.V., & Sivo, S.A. (2009). One-dimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A Rasch analysis. *Journal of Psychopathology and Behavioral Assessment, 31,* 309-319.

Sakai, L.M., Whitebook, M., Wishard, A., & Howes, C. (2003). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing differences between the first and revised editions. *Early Childhood Research Quality, 18(*4*)*, 427-445.

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly, 21(*1), 76-92.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *ACF-OPRE Report. Compendium of Quality Rating Systems and Evaluations*. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Vandell, D.L. (2004). Early child care: The known and the unknown. *Merrill-Palmer Quarterly-*

Journal of Developmental Psychology, 50(3), 387-414

Vandell, D.L., & Wolfe, B. (2000). *Child care quality: Does it matter and does it need to be improved?* Washington, DC: U.S. Department of Health and Human Services.

White House. (2014a). State of the Union. Retrieved from http://www.whitehouse.gov/sotu

White House. (2014b). Education: Knowledge and skills for the jobs of the future. Retrieved from http://www.whitehouse.gov/issues/education/early-childhood

Wilson, M., & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika, 60(2),* 181-198.

Wolfe, E.W., & Smith, E.V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - Instrument development tools. In E. W. Wolfe & E. V. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications* (pp. 202-242). Maple Grove, MN: JAM Press.

Wolfe, E.W., & Smith, E.V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II - validation activites. In E. W. Wolfe & E. V. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *ACER ConQuest version 2.0: Generalized item response modeling software*. Statistical software.

Zaslow, M., Halle, T., Martin, L., Cabrera, N., Calkins, J., Pitzer, L., et al. (2006). Child outcome measures in the study of child care quality. 577-610.

Zaslow, M., Martinez-Beck, I., Tout, K. & Halle, T. (2011). *Quality Measurement in Early Childhood Settings*. Baltimore, MD: Brookes Publishing.

Zellman, G.L. & Perlman, M. (2008). *Child-Care Quality Rating and Improvement Systems in Five Pioneer States Implementation Issues and Lessons Learned*. Santa Monica, CA: RAND.

Table 1
Number and Percentage of Ordered, Overlapping, and Disordered Indicators, Overall and by ECERS-R Scale Developers' Subscales

	Number of		Number (Percentage) of Comparisons		
	Indicators	Comparisons	Ordered	Overlapping	Disordered
All Items	365	1386	781 (56)	502 (36)	103 (7)
Within ECERS-R Subscales					
Space and Furnishings	75	270	154 (57)	104 (39)	12 (4)
Personal Care Routines	70	306	112 (37)	133 (43)	61 (20)
Language-Reasoning	39	140	95 (68)	41 (29)	4 (3)
Activities	98	352	245 (70)	86 (24)	21 (6)
Interaction	53	209	115 (55)	92 (44)	2 (1)
Program Structure	30	109	60 (55)	46 (42)	3 (3)

Note. $n = 282$ centers. Values in Column 1 are the number of indicators in the analysis (total, and within subscales). Values in Column 2 are the number of comparisons between adjacent indicators (e.g., if there were 2 indicators of category 1 and 3 categories of category 3, then there would be 6 adjacent comparisons). Columns 3 to 5 provide the number and percentage of comparisons that are ordered, overlapping, and disordered. Based on 99% confidence intervals of the indicators' difficulty estimates, we defined: *ordered* indicators if the indicators placed by the ECERS-R authors at a lower rating category were empirically estimated to be easier than indicators that the ECERS-R authors had placed at a higher rating category (i.e., the upper bound of the lower indicator was below the lower bound of the higher indicator), *overlapping* indicators if the confidence intervals of the two indicators overlapped, and *disordered* indicators if the indicators placed by the ECERS-R authors at a lower rating category were empirically estimated to be harder than indicators that the ECERS-R authors had placed at a higher rating category (i.e., the lower bound of the lower indicator was above the upper bound of the higher indicator).

Table 2

Number and Percentage of Indicators Initially Misfitting and Removed During Refitting for the ECERS-R Scale Developers' Subscales and Expert-identified Meta-Domains

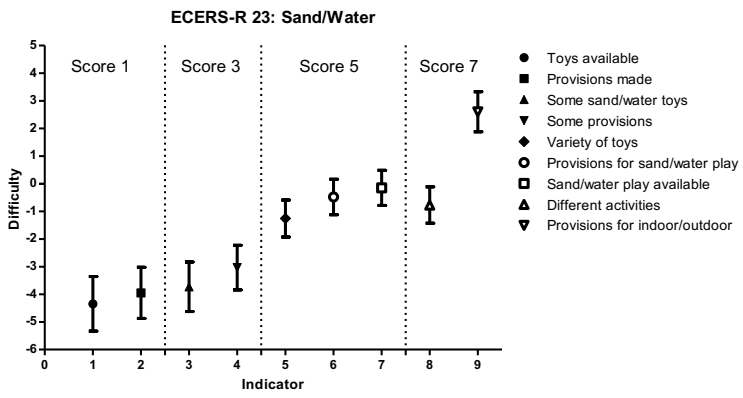
	Total Number of Indicators	Indicators Initially Misfitting		Indicators Removed During Refitting	
		Number	Percent	Number	Percent
ECERS-R Subscale					
Space and Furnishings	75	5	7	6	8
Personal Care Routines	70	3	4	8	11
Language-Reasoning	39	2	5	3	8
Activities	98	9	9	58	59
Interaction	53	5	9	12	23
Program Structure	30	7	23	14	47
Expert-Identified Meta-Domains					
Health	14	0	0	0	0
Cognitive	17	1	6	1	6
Cognitive-Socioemotional	99	10	10	14	14
Cognitive-Socioemotional-Health	44	7	16	6	14

Note. $n = 282$ centers.

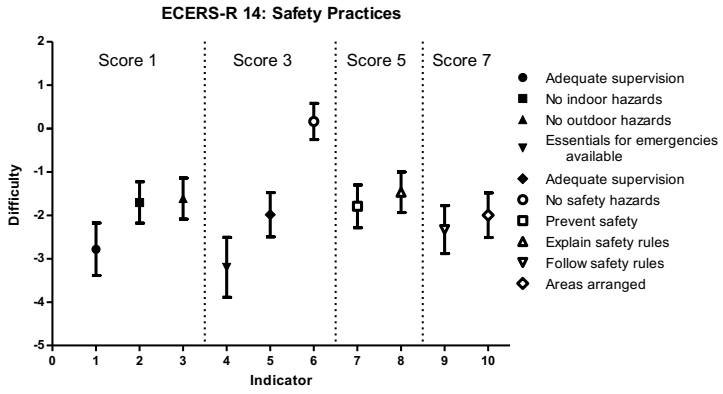
Figure 1

Examples of Items with Many Ordered (Panel a), Overlapping (Panel b), and Disordered (Panel c) Adjacent Comparisons

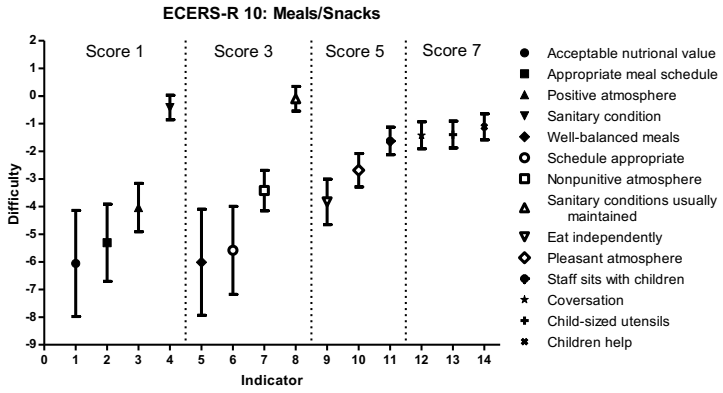
a)



b)



c)



Note. Plotted values are indicator difficulty levels estimated using the RIBM (with 99% confidence intervals). Indicators were first grouped within rating category levels and then ordered by their difficulty point estimate. Shortened labels are provided for each indicator. The abbreviated labels of indicators of "Score 1" are positively oriented, reflecting the fact that we reverse-scored them for analysis.

Figure 2

Map of Fitting Indicators and Preschool Classrooms for Expert-Identified Health Dimension

Measures in Logits	Expert-Identified Health Dimension	
	Classrooms	Indicators
	Higher	More Difficult
4	X	
	X	
	X	
	X	
	X	
3	XX	
	XX	
	XXX	
	XXXXX	
2	XXXXX	
	XXXXX	
	XXXXXXXX	
	XXXXXX	
1	XXXXXXX	
	XXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	
0	XXXXXXXXX	"13.3.1 Adequate handwashing by staff and children takes place after wiping noses…"
	XXXXXXX	"10.3.3 Sanitary conditions usually maintained [meals/snacks]"
	XXXXXXXX	
	XXXXXXX	"12.3.1 Sanitary conditions are maintained [toileting/diapering]"
		"12.3.3 Staff and children was hands most of the time after toileting"
	XXXXXXX	
	XXXXXXXX	"12.1.1 [R] Sanitary conditions of area are not maintained [toileting/diapering]"
-1		"12.1.3 [R] Handwashing often neglected by staff or children after toileting/diapering"
	XXXXXX	
	XXXXXX	
	XXXXXX	"12.5.1 Sanitary conditions easy to maintain [toileting/diapering]"
		"13.3.2 Staff usually take action to cut down on the spread of germs"
-2	XXXXXX	
	XXXXX	"12.3.2 Basic provisions made for care of children [toileting/diapering]"
	XXX	
	XXX	
-3	XX	
	X	
	X	
	XX	"14.3.3 Essentials needed to handle emergencies available"
	X	"2.5.2 All furniture is sturdy and in good repair"
-4		
		"13.3.4 Procedures used to minimize spread of contagious disease"
-5		"13.3.3 Smoking does not take place in child care areas"
		"13.1.2 [R] Smoking is allowed in child care areas, either indoors or outdoors"
	Lower	Easier

Note. Each "X" represents 1.7 centers. The labels are provided for each indicator, with numbers connecting back to items and indicators (e.g., 13.3.1 means the 1st indicator of Category 3 on Item #13). We added [R] to the front of indicators of response category 1 to indicate that we reversed these negatively-oriented indicators prior to analysis. Where needed to distinguish meaning, the words in square brackets reflect the overall focus of an item. Fourteen items are graphed, because two indicators allowing not applicable were excluded from the analysis ("10.3.5 Allergies posted and food/beverage substitutions made" and "13.7.2 Individual toothbrushes properly labeled and stored; used at least once during the day in full-day programs").

Figure *3*

Map of Fitting Indicators and Preschool Classrooms for Expert-Identified Cognitive Dimension

Measures	Expert-Identified Cognitive Dimension	
in Logits	Classrooms	Indicators
	Higher	More Difficult
	X	
3	X	
	X	
	X	
	XXX	
	XXX	
2	XXX	
	XXXX	
	XXXXXXX	
	XXXXXXX	
	XXXXXXXXXX	"26.5.1 Many developmentally appropriate materials of various types [math/number]"
1	XXXXXXXX	
	XXXXXXXX	
	XXXXXXXXX	"26.5.2 Materials are accessible for a substantial portion of the day [math/number]"
	XXXXXXXXXXX	
	XXXXXXXXXXXXX	
0	XXXXXXXXXXXX	
	XXXXXXXXX	"25.7.1 Nature/science activities requiring more input from staff [at least biweekly]"
	XXXXXXXXXXXXX	"25.7.2 [Materials] extend children's hands-on experiences [nature/science]"
	XXXXXXXXXXXX	"26.7.1 Math/number requiring more input from staff [at least biweekly]"
	XXXXXXXXXXX	"26.5.4 Daily activities use to promote math/number learning"
	XXXXXXXXX	"26.7.2 Materials are rotated to maintain interest [math/number]"
-1	XXXXXXXX	
	XXXXXXX	"15.7.2 Some books relate to current classroom activities or themes"
	XXXX	
	XXXXXX	
	XXX	"25.1.1 [R] No games, materials, or activities for nature/science accessible"
-2	XX	"26.5.3 Materials well organized and in good condition [math/number]"
	XXX	"26.3.1 Some developmentally appropriate math/number materials accessible"
	XX	"22.3.3 Blocks and accessories accessible for daily use"
	X	
	XX	
	X	
-3	X	"22.1.1 [R] Few blocks are accessible for children's play"
	XX	
	X	"15.5.3 Books organized in a reading center"
		"26.1.2 [R] Math/Number taught primarily through rote counting or worksheets"
-4		
		"26.1.1 [R] No math/number materials accessible"
	X	
	Lower	Easier

Note. Each "X" represents 1.4 centers. Shortened labels are provided for each indicator, with numbers connecting back to items and indicators (e.g., 26.5.1 means the 1st indicator of Category 5 on Item #26). We added [R] to the front of indicators of response category 1 to indicate that we reversed these negatively-oriented indicators prior to analysis. Where needed to distinguish meaning, the words in square brackets reflect the overall focus of an item. Sixteen items are graphed, because the one misfitting indicator was excluded ("26.3.2 Materials accessible daily").

Appendix 1

Number of Indicators as Organized by Scale Developers into Items and Subscales

	Number of Indicators		
		Used in Current Study	
	Total	Observed Centers[a]	Expert Ratings[b]
Total	**383**	**365**	**254**
I. Space and Furnishings	**82**	**75**	**24**
1. Indoor space	14	13	4
2. Furniture for routine care, play and learning	10	7	2
3. Furnishings for relaxation and comfort	9	9	2
4. Room arrangement for play	12	11	5
5. Space for privacy	7	7	1
6. Child-related display	9	8	2
7. Space for gross motor play	10	10	3
8. Gross motor equipment	11	10	5
II. Personal Care Routines	**77**	**70**	**55**
9. Greeting/departing	12	10	6
10. Meals/snacks	18	14	8
11. Nap/rest	12	12	8
12. Toileting/diapering	14	14	13
13. Health practices	11	10	10
14. Safety practices	10	10	10
III. Language-Reasoning	**39**	**39**	**39**
15. Books and pictures	11	11	11
16. Encouraging children to communicate	9	9	9
17. Using language to develop reasoning skills	8	8	8
18. Informal use of language	11	11	11
IV. Activities	**101**	**98**	**55**
19. Fine motor	9	9	5
20. Art	9	8	6
21. Music/movement	10	10	2
22. Blocks	11	11	3
23. Sand/water	9	9	2
24. Dramatic play	12	12	6
25. Nature/science	10	10	7
26. Math/number	10	10	10
27. Use of TV, video, and/or computers	11	9	5
28. Promoting acceptance of diversity	10	10	9
V. Interaction	**53**	**53**	**51**
29. Supervision of gross motor activities	10	10	9
30. General supervision of children	11	11	10
31. Discipline	12	12	12
32. Staff-child interactions	10	10	10
33. Interactions among children	10	10	10
VI. Program Structure	**31**	**30**	**30**
34. Schedule	11	11	10
35. Free play	10	9	10
36. Group time	10	10	10

Note. Values are the number of indicators. Wording of 36 items and 6 subscales from Harms, Clifford, and Cryer, 1998. [a] As described in the text, we excluded a small number of indicators from our IRT analyses of the observed centers, primarily indicators that were often missing when not applicable. [b] As discussed in the text, we excluded a larger number of indicators from the expert surveys, to reduce response burden and redundancy.

Appendix 2
Study-Defined Domains and Meta-Domains

Three Meta-Domains	Eight Domains	Domain Definition
Cognitive	Promote Language Skills	Materials, activities, and child-caregiver interactions that expose children to spoken and written language, such as looking at and reading books or talking among caregivers and children.
Cognitive	Promote Math Skills	Materials, activities, and child-caregiver interactions that expose children to numbers, spatial relations, measurement, classification and patterning.
Cognitive	Promote a Positive Approach to Learning	Materials, activities, and child-caregiver interactions that embed learning throughout the day, connect skill development to daily experiences, and promote fun and enthusiasm across activities that promote skill development.
Socioemotional	Reduce Behavior Problems	Child-caregiver interactions, particularly around rules and discipline, that are warm and consistent as opposed to: (1) harsh, (2) irregular, and/or (3) lax.
Socioemotional	Promote Social Competence	Materials, activities, and child-caregiver interactions that promote concern and respect for others, understanding and respect for rules, and skills in joining groups, in leadership and teamwork, and in conflict-resolution.
Socioemotional	Promote Emotional Regulation	Helping children recognize and label their emotions and assisting children in using strategies to constructively respond to their emotions, including the provision of space, materials, activities, and child-caregiver interactions that allow children to remove themselves from stressful situations, distract themselves from distressing interactions, and soothe themselves with comforting objects and/or interactions.
Health	Reduce the Spread of Illness	Sanitary conditions or infection control practices that reduce the spread of infectious diseases. Practices could include those that the caregiver implements herself and that she trains the child(ren) to implement.
Health	Reduce Injuries	Environmental conditions that reduce exposure to situations where accidents might occur (e.g., barriers) and caregiver practices that reduce injury risk, especially when contextual risk cannot be avoided (e.g., supervision).