Examining the Category Functioning of the ECERS-R

Across Eight Datasets

Ken A. Fujimoto

Loyola University Chicago

Rachel A. Gordon and Fang Peng

University of Illinois at Chicago

Kerry G. Hofer

Abt Associates

AERA Open, 4(1), 1-16, Published March 1 2018

Ken A. Fujimoto, Assistant Professor, Research Methodology Program, Loyola University Chicago, 820 N. Michigan Ave. Rm 1113, Water Tower Campus, Chicago, IL 60611. Phone (312) 915-6852. E-mail: kfujimoto@luc.edu.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130118 to the University of Illinois at Chicago (Rachel A. Gordon, PI) with a subcontract to Vanderbilt University (Kerry G. Hofer, PI). We gratefully acknowledge advice from Ariel Aloe and Everett Smith and research assistance from Rowena Crabbe, Danny Lambouths, Elisabeth Stewart, Kayla Polk, Jenny Kushto-Hoban, and Hillary Rowe in data extraction and literature review. The opinions expressed are those of the authors and do not necessarily represent views of the Institute, the U.S. Department of Education, or our consultants.

Abstract

Classroom quality measures, such as the Early Childhood Environment Rating Scale, Revised (ECERS-R), are widely used in research, practice, and policy. Increasingly, these uses have been for purposes not originally intended, such as contributing to consequential policy decisions. The current study adds to recent evidence of problems with the ECERS-R standard stop-scoring by analyzing eight studies offering 14 waves of data collection in approximately 4,000 classrooms. Our analyses, which featured the nominal response model, generalized partial credit model, partial credit model, within-category averages of total scores, and point-biserial correlations, revealed that all 36 items had categories that did not follow an ordinal progression with respect to quality. Additionally, our results showed that the category problems accumulated to the scale score. The results caution against the use of the standard raw scoring and encourage development of alternative scoring methods for the ECERS-R.

Examining the Category Functioning of the ECERS-R

Across Eight Datasets

Major policy efforts aim to make preschool universally available and to improve the quality of child care settings, with a goal of preparing all children for school (Child Trends, 2015; Pew Charitable Trusts, 2014; U.S. Department of Education, 2013a). Importantly for our study, policies often dictate that observational measures are incorporated in an attempt to ensure high classroom quality. Often, raw scores (e.g., averaging across all items) from these measures are compared to cut scores, contributing to consequential decisions for child care subsidy levels, Head Start funding, and public recognition with medals (gold, silver, bronze) or stars (5-star, 4-star, etc.). One widely used measure to assess the quality of child care centers is the Early Childhood Environment Rating Scale, Revised (ECERS-R; Harms, Clifford, & Cryer, 1998). The latest compendium of state Quality Rating and Improvement Systems found that 40% of states used only the ECERS-R and another 40% used ECERS-R along with another quality measure (Child Trends, 2015). A recent survey of state pre-kindergarten policies similarly found that 19 states relied on ECERS-R for program monitoring (Ackerman, 2014). With such consequences for funding and reputation, these measures can have an outsized influence on teacher practice, similar to high stakes student testing. Therefore, probing the psychometric properties of the measures is important.

Indeed, the validity of the ECERS-R scores for these uses has increasingly come into question because of the small associations between its scale scores and child developmental outcomes (e.g., Burchinal, Zaslow, & Tarullo, 2016; Burchinal, Kainz & Cai, 2011; Layzer & Goodson, 2006). Among many of the reasons for these low associations, recent studies pointed to limitations with the ECERS-R standard stop-scoring (e.g., Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013). At first glance, the ECERS-R seems to have a simple Likert-like scoring, with category scores increasing from 1 (*Inadequate*) to 7 (*Excellent*). A thorough examination of the items and scoring process, however, reveals the potential for the score categories not to follow an ordinal progression because assigning higher scores depends on scoring decisions for lower scores (referred to as "stop-scoring") and indicators that probe different aspects of quality are mixed together within some items (e.g., mixing of sanitation aspects of quality, like handwashing, with social, like conversations, as detailed below). Thus far, only a handful of studies have empirically tested the ordinal nature of the ECERS-R item categories, and a new version of the measure (i.e., the ECERS-3) has retained the same stop-scoring standard (Harms, Clifford, & Cryer, 2015).

Given the concerns that have arisen about the ECERS-R scores, the purpose of this study was to perform a comprehensive analysis of the category functioning of the ECERS-R items. We focused on whether the categories were: (a) ordered (i.e., followed an ordinal progression); (b) redundant (i.e., two categories represented similar quality levels); (c) disordered (i.e., a subsequent category represented lower quality); and (d) underutilized (i.e., categories had a low probability of being used). Our analytic approaches featured three item response theory models—the nominal response model (NRM; Bock, 1972), the generalized partial credit model (GPCM; Muraki, 1992), and the partial credit model (PCM; Masters, 1982). Although the PCM has been used more frequently in prior studies involving ECERS-R data, the NRM allows us to better diagnose the four types of problems the categories may have, and the GPCM allows us to examine how sensitive the results are to the PCM model assumptions that we detail below. Additionally, we calculated the within-category raw score averages and point-biserial correlations to examine how problems with the category functioning accumulated to the scale score level.

We used eight datasets, with 14 waves of data collections. Our data analysis procedures consisted of parallel and stacked analyses, which followed recent calls for integrative and coordinated data analysis and robustness checking (Curran et al., 2008; Duncan, Engel, Claessens, & Dowsett, 2014; Hofer & Piccinin, 2009). The advantages of these procedures were two-fold. The parallel analysis allowed us to determine whether the results replicated across the individual datasets (i.e., were robust across sample compositions and data collection; Hofer & Piccinin, 2009). Unfortunately, each dataset was not amenable for the NRM because of sample size limitations. The stacked analysis integrated the separate datasets into one, leading to a sufficient number of cases for the NRM (Marcoulides & Grimm, 2016). As a result, we provide the first detailed NRM results for the ECERS-R and more comprehensive information about category functioning than could have been obtained with the PCM alone. Additionally, the stacked analysis provided greater precision and power in detecting where the problems in the ordinal progression of the categories were occurring. By taking this multifaceted analytic approach, we gathered robust evidence on the category functioning of the ECERS-R items and provided detailed diagnostic information to guide future use and research involving the instrument.

The ECERS-R Scoring

Our examination of the ECERS-R scoring guidelines is why we expect problems with category usage. The instrument's unique scoring rules reflect its origins in the 1970s as a checklist created in response to early education centers' requests for guidance on self-improvement (Frank Porter Graham Child Development Institute, 2003). Reflecting these checklist origins, the ECERS-R includes over 400 indicators covering different aspects of quality (e.g., "sanitary conditions usually maintained," "pleasant social atmosphere," "books organized in a reading center;" Harms et al., 1998). To facilitate both observers and practitioners' ability to mentally digest these hundreds of indicators, the instrument developers organized them into a few dozen items. Within each item, the indicators were further grouped to represent different scores ranging from 1 to 7, with the indicators listed at the odd-numbered categories (labeled 1 = *Inadequate*, 3 = *Minimal*, 5 = *Good*, and 7 = *Excellent*).

To further reduce burden on observers, the developers created a stop-scoring rule calling for observers to stop checking the indicators for an item once they reach a category that does not meet the scoring rules. Figure 1 visually represents these rules. For category 1, all indicators are negatively oriented (e.g., "no interest centers defined"). If at least one of these indicators is endorsed, then the item receives a score of 1 and the observer moves on to the next item. If none of the category 1 indicators is endorsed, the observer considers the indicators of category 3. These indicators at category 3 (and categories 5 and 7) are positively oriented. If less than half of the indicators of category 3 are present, then the score remains in category 1, and the observer moves on to the next item. If at least half but not all of the indicators in category 3 are present, the score is a 2, and the observer moves on to the next item. If all of the indicators in category 3 are present, then the indicators of category 5 are considered. Category 5 is then scored in similar fashion as category 3. A score of 7 is only given if all indicators under that category are met.

This stop-scoring process reduces the burden on the observers because only a subset of indicators need to be considered for most items (especially when a classroom's scores fall in the lower categories). If the scale developers' placement of the indicators matched their actual locations on the quality continuum such that the indicators placed at higher categories truly reflected more quality than those listed at lower categories, then this scoring efficiency should not affect the categories' ordinal representation of quality. However, to the extent that the indicators do not reflect an ordered progression of true quality, the stop-scoring might produce problems with category underutilization, redundancy, and disorder. We feature three such issues revealed by scrutinizing indicator content: (a) complementary indicators, (b) basic-versus-advanced indicators; and (c) different-content indicators.

The first situation of complementary indicators is evident at categories 1 and 3 for some items, where the two categories have nearly equivalent indicators that are phrased in opposite directions. The ninth ECERS-R item (*Greeting/departing*) illustrates this issue. For example, “Greeting of children is often neglected” is an indicator under category 1, and “Most children greeted warmly” is an indicator under category 3. A classroom that meets the first condition (greeting is not neglected) would likely also meet the second condition (most children greeted warmly), potentially leading to categories 2 and 3 being underutilized. Redundancy of these categories might also result, due to slight variations between otherwise complementary indicators (e.g., words like “warmly”).

The second issue – presence of basic and advanced indicators at the same category level – may affect the chances of an observer perceiving evidence that meets the cutoff for odd scores (less than half of the indicators being observed) versus even scores (at least half but not all). Item 18 (*Informal Use of Language*) offers an example. All of this item’s indicators deal with the quality and quantity of conversations, but within categories 5 and 7, the indicators appear to tap into aspects that are: (a) basic (e.g., staff have individual conversations with most children) and (b) advanced (e.g., staff ask questions that encourage long and complex answers). The relative number of basic and advanced indicators at each of these categories will affect the chances of meeting the cutoff of “less than half “ versus the cutoffs of “half but not all” or “all.” To the extent that meeting the “less than half” cutoff is particularly uncommon, odd scores (3, 5 and 7) will be underused. Notice that this issue is complicated because it depends not only on the focal category’s indicators, but also those of the preceding and subsequent categories. Potentially this issue could also produce redundancy or disorder, to the extent that some basic indicators placed at higher categories overlap basic content placed at lower categories.

Although these issues of complementary and basic-versus-advanced indicators of the same content have not been featured in prior IRT studies of the ECERS-R, the final issue of mixing different aspects of quality has been discussed. Scholars and users of the ECERS-R have raised concerns that preschool classrooms can be scored in a lower category due to lax health and safety practices despite possessing other aspects of quality such as warmth and responsiveness of caregivers (Gordon et al., 2013, 2015; Layzer & Goodson, 2006). For instance, on the tenth item (*Meals/snacks*), stringent criteria for sanitary conditions (e.g., most children and adults wash their hands before eating) must be met before observers can consider the social aspects of mealtime (e.g., rich conversation and supportive relationships). The scale developers' placement of indicators reflects a common belief held in the field that health and safety are more fundamental aspects of quality whereas the socio-emotional and academic nature of teacher-child interactions are more advanced aspects. However, if the placement differs from empirical ordering, it could lead otherwise higher quality classrooms to be scored in the lowest category. This mixing of different aspects of quality is particularly evident in the ECERS-R items that cover children's personal care routines. Therefore, we expect category redundancy and disorder to be especially likely for these items.

Prior Studies on the ECERS-R Scoring

Just a few empirical studies have examined the ECERS-R for these potential problems with category functioning. Although their results are suggestive, these studies have not yet leveraged all of the item response theory tools. Most importantly, their focus on the PCM over the NRM is limiting, because the PCM cannot separate category underutilization from category redundancy or disorder. This limitation is accentuated by the recent debates about the meaning of reversed thresholds under the PCM (Adams, Wu, & Wilson, 2012; Andrich, 2013), which can also be informed by the NRM's detection of the specific type of problems evident in the categories.

More specifically, Gordon et al. (2013) applied the PCM to ECERS-R item-level data from over 1,300 classrooms participating in the nationally-representative Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) study, gathered using the stop-scoring rule. They found that every ECERS-R item had at least one pair of adjacent category thresholds that was out of order. Mayer and Beckh (2016) also used the PCM with a nationally-representative German sample of 270 classrooms and replicated reversals of adjacent threshold estimates. These findings suggest some problems with category usage, but not what the problems are. One culprit could be category underutilization, potentially occurring more often in categories that are odd-valued and correspond to lower scores, and in the personal care routines items, as we previously noted above. The two published PCM studies only reported about the number of items with reversals in adjacent thresholds and did not detail which items and categories. Thus, an important contribution of our study is using the precision achieved by our large stacked data file to pinpoint where such problems occur.

In the broader literature regarding the PCM, some researchers have also argued that reversals of adjacent thresholds from the PCM likely reflect unusual samples rather than problems with an instrument itself (Adams et al., 2012). That is, the fact that a category appears underused may simply reflect a sample that happened to exclude people (or classrooms) reflecting those scores. Although the two existing PCM-based studies of the ECERS-R relied on nationally-representative samples – which should be less likely than convenience samples to have excluded classrooms representing certain scores on certain items – our replication of results across numerous datasets representing a range of care settings helps adjudicate whether the underutilization in the ECERS-R data is because of the instrument or the sample. Although thresholds are less precisely estimated in our parallel analysis than the stacked analysis, these dataset-specific results could also offer insight into the root of the underutilization. Replicated problems in the same categories of the same items across the datasets would suggest that the instrument is the issue, because the problems would not be specific to any one sample.

Other researchers have emphasized the possibility that reversals of the thresholds from the PCM model could arise because of disorder in the meaning of the categories (Andrich, 2013). In this case, interpretations of the overall scale scores (either total or averaged) are muddled because a lower score could represent greater amounts of quality than a higher score. We leverage the NRM to separate these problems of actual category disorder and category underutilization. Our approach is consistent with researchers' renewed attention to the effectiveness of the NRM for testing category functioning in rating scale data (Preston, Reise, Cai & Hays, 2011; Preston & Reise, 2015; Thissen, Cai, & Bock, 2010). The NRM is more flexible than the PCM (the latter being nested within the former) and can distinguish among possible disorder, redundancy, underutilization, and order of categories for individual items. However, the NRM has been infrequently used, possibly because of its data demands. Sample sizes under 2,000 may be underpowered in identifying certain types of category problems (Preston & Reise 2015). Our stacked dataset provided the needed sample size, allowing us to test whether the NRM fit better than the PCM and to illuminate the reasons for category threshold reversals under the PCM. We also analyzed the data with the GPCM. This model has not been used as frequently as the PCM for examining the category functioning of the ECERS-R. Including the GPCM, however, allowed us to determine whether the category problems reported based on the PCM reflected its constraint of all items to equally discriminate on the classroom quality level.

To the best of our knowledge, our manuscript is the first to offer detailed results from the NRM applied to ECERS-R data. Gordon and colleagues (2013) mentioned testing alternative measurement models, including the NRM, but merely offered a sentence summary that the NRM fit the data the best and that its category discriminations lacked order. Other ECERS-R studies suggest the NRM might reveal problems with category disorder as well as category underutilization. Two studies analyzed indicator-level data from the ECERS-R, where observers had evaluated all indicators (rather than following the stop-scoring rule). In the first study, Lambert and colleagues (2008) analyzed indicators for a subset of ECERS-R items scored for 300 classrooms in Jamaica and Grenada. Consistent with possible category disorder, their estimated indicator difficulty levels differed from the instrument developers' placement (e.g., an indicator that the instrument developers had placed at a score of 7 was estimated via Rasch modeling to reflect lower quality than an indicator placed at a score of 5 on the same item). Likewise, Gordon and colleagues (2015) analyzed indicator-level data for 36 of the 43 ECERS-R items, with the data coming from several hundred U.S. classrooms. They similarly found that two-thirds of the items had at least one pair of indicators that were empirically ordered in a different manner from the ECERS-R instrument developers' placement. Beyond these indicator-level analyses, two studies also looked at possible disorder at the total score level, finding that within-category raw score averages and point-biserial correlations did not always increase with the score categories (Gordon et al., 2015; Mayer & Beckh, 2016). These results suggest that problems with category usage in the ECERS-R may be extensive and sizable enough to matter at the scale-score levels, although replication is needed beyond these two studies.

Summary and Focus of Our Study

Examining the ECERS-R scoring procedures suggests possible problems with category usage. Yet, just a few studies have examined category functioning for the instrument. Our study advances the literature by using multiple analytic strategies (i.e., NRM, GPCM, PCM, within-category raw averages, and point-biserial correlations) and approaches (parallel and stacked analyses) replicated across eight datasets consisting of 14 waves. Our findings have important implications for the appropriateness of using ECERS-R scores to draw inferences from research studies and for consequential policy decisions; these uses amplify the advantage of our analyzing a wide range of samples (including centers serving low-income children and funded by state pre-kindergarten or federal Head Start programs, all the focus of policy efforts). Our more comprehensive analyses, especially because we included the NRM, let us differentiate among possible reasons for problems with category usage (e.g., category underutilization versus category redundancy or disorder). The relevance of our findings is that they serve as evidence for the response processes aspect of validity as outlined in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Additionally, our findings offer more specific implications for future use and revision of the instrument than have prior studies. Given our examination of the ECERS-R scoring rules discussed earlier, we anticipate underutilized or redundant categories to occur particularly often for the lowest categories (where indicators are sometimes complementary), for disorder to be especially common for items that mix indicators of different aspects of quality (like the personal care routines items), and to see each of these problems more often at odd- versus even-numbered categories (due to differences in their scoring rules).

Method

Datasets

Our study involved secondary analysis of data from eight large-scale research projects that were conducted during the 2000s, all of which included ECERS-R item scores. Table 1 contains the sample sizes and demographic information by dataset and wave.

Program types and family demographics. Four of the eight studies focused on the federal Head Start program, which targeted low-income children. These included the 2000 and 2003 cohorts of the Head Start Family and Child Experiences Survey (FACES), both nationally representative surveys of Head Start classrooms; and the Head Start Impact Study (HSIS) and the Early Head Start Research and Evaluation Project (EHSRE), both of which randomly assigned eligible children into Head Start or control groups. We did not use FACES 1997 because it relied on the original ECERS rather than the ECERS-R. We also did not use FACES 2006 and 2009 because their public releases did not contain all ECERS-R items. The Fragile Families and Child Wellbeing Study (Fragile Families) was also primarily comprised of low-income children (65%), although the children attended a wider range of classroom types than was the case in the Head Start samples (about one-sixth Head Start, and nearly one-tenth state pre-k, with almost three-quarters being other community based centers). For the ECLS-B, we focused on classrooms attended by low-income children because an earlier study (Gordon et al., 2013) had already examined the psychometric properties of the ECERS-R in the full ECLS-B sample. Among low-income ECLS-B preschoolers, 40% of the children attended Head-Start-funded centers and 25% attended school-based classrooms; the remaining 35% attended other community-based centers.

In contrast to the samples that primarily served lower-income children, the majority of the children in the Preschool Curriculum Evaluation Research Initiative (PCER) and Quality Interventions for Early Care and Education (QUINCE) samples had families with incomes at or above 200% of the federal poverty line. The PCER sample mostly attended state pre-kindergarten classrooms (58%) and the majority of the QUINCE sample attended private, community-based settings (80%).

Research teams and ECERS-R training. Most of the datasets come from large-scale studies and were implemented by survey firms, including Westat (FACES 2000, FACES 2003, and HSIS), MPR (Mathematica Policy Research; EHSRE, Fragile Families, portion of PCER), and RTI (Research Triangle Institute; ECLS-B, portion of PCER). These studies included a sizable number of classrooms on which ECERS-R observations took place (most had approximately 200-400 classrooms, and a few had close to 1,000 classrooms), with some classrooms observed twice (in both fall and spring of a single academic year) or three times (fall and spring of the first year; spring of the second year).

The studies' documentation generally reported extensive training of observers in using the ECERS-R, which included a combination of lectures, discussions, in-home or take-home exercises, and practices (e.g., based on videotapes and visits to classrooms). Each study had a process for certifying observers, such as agreement with a "gold standard" trainer or with consensus codes. Several studies reported following the reliability benchmark recommended by the ECERS-R authors, with raters achieving an inter-rater reliability of at least 80% (within one point at the item level) to be qualified for performing study observations. The QUINCE study was coordinated by the institution where the ECERS-R was developed and included a daylong training by one of the instrument's authors. These features made QUINCE important to our replication objectives, despite it offering the smallest sample size.

ECERS-R Data

Consistent with the majority of the studies on the ECERS-R, we focused on the first 36 ECERS-R items (omitting the item that was only scored if a child with an identified disability attended the program, as well as items focused on parents and staff). Appendix A includes a summary of these ECERS-R items' scores in the stacked 14 datasets/waves. The IRT models we used in this study require a certain number of classrooms to be rated with each category within an item so that the category parameters can be estimated. The datasets with fewer classrooms had instances in which one or more categories within an item went unused. Thus, for our parallel analysis (i.e., each dataset analyzed individually), we followed prior research and collapsed unused categories with adjacent categories to ensure every category had at least one case (Linacre, 2004; Preston et al., 2011). Such collapsing was especially needed for the QUINCE dataset, which had the smallest overall sample size (we collapsed categories for six items in Wave 1 and 20 items in Wave 2; see Appendix B). Category collapsing was also needed for one to three items in each wave of the FACES datasets (e.g., for Item 2, *Furniture for routine care, play, and learning*, which was scored close to the maximum scale score of 7 in all waves; see Appendix B).

Our stacked dataset was formed using only the first wave of each dataset to avoid non-independence of observations, leading to item-level scores from 4,048 classrooms (after rounding the PCER and ECLS-B sample sizes, as per their reporting requirements; see Table 1). We recognize that by stacking the datasets, we implicitly assume equivalence of parameters across datasets. Although sample sizes were insufficient to formally test for invariance across datasets with the NRM model, in another study, we used a less data-demanding factor analytic model that did not require cases in every category of every item and found minimal noninvariance in factor loadings, which had negligible impact at the score level (citation omitted for blind review). The consistency in the pattern of adjacent thresholds across the datasets for the PCM also suggested that the lack of invariance was minimal. By assuming measurement invariance, our stacked dataset resulted in every category within an item to be used (95% of all possible categories had 100 or more ratings), and thus none of the categories required collapsing for our stacked analysis portion of the study. To account for the different datasets possibly representing separate subpopulations within the overall population, we allowed the latent means and standard deviations to vary across the datasets in our NRM, GPCM, and PCM models using a multiple group analysis.

Analytic Approaches to Detecting Problems in the Score Categories

We used IRT and raw score approaches in detecting problems in the category functioning of the ECERS-R Items. We provide details of the IRT models that we used for this study and their relevant parameters for diagnosing category problems, and then we describe our raw score approaches (i.e., within-category raw score averages and point-biserial correlations).

Item response theory approaches. In our presentation of the NRM, GPCM, and PCM models, we use the following indices. The datasets are indexed using g (where g = 1, 2,…, 8 and each value represents a dataset) in our multiple-group analyses. This subscript is not needed when each dataset is analyzed by itself because all classrooms belong to the same dataset (or group) in this case. The classrooms are indexed using i (where i = 1, 2,…, n_g, and n_g is the number of cases in dataset g). The items are indexed using j (where j = 1, 2,…, 36), and category scores are indexed using k (where k = 1, 2,…, m_j, with m_j being the highest score category for item j). In the stacked dataset, m_j equaled 7 for all items. In the parallel analysis, m_j equaled 7 for the items in each individual dataset where all categories were used and less than 7 for those items that had one or more unused categories (see Appendix B for details about which items within a dataset required rescoring).

The nominal response model. The NRM arrives at the probability of a rating of k on item j conditional on the quality level for classroom i in dataset g (θ_{ig}) through:

$$P(Y_j = k|\theta_{ig}) = \frac{\exp(a_{jk}\theta_{ig} + c_{jk})}{\sum_{t=1}^{m_j} \exp(a_{jt}\theta_{ig} + c_{jt})}, \quad (1)$$

where a_{jk} and c_{jk} represent the k^{th} category's discrimination and intercept, respectively, for item j, and the t is a substitute index for summing across all score categories.

When examining the ordering of the categories, the category boundary discriminations (CBDs) are of primary interest (Preston & Reise, 2015; Preston et al., 2011). The CBD (a^*_{jk}) for two adjacent categories ($k-1$ and k) within item j is obtained by taking the difference of the category discriminations for these categories, that is,

$$a_{jk}^{*} = a_{jk} - a_{j(k-1)}, \tag{2}$$

which represents the relative discrimination between these two categories (Thissen et al., 2010). When a CBD is large and positive the adjacent discriminations are considered ordered, which indicates that the corresponding categories are ordered in meaning (i.e., category k represents more quality than category $k - 1$). When a CBD is 0 or positive but small, the adjacent discriminations are considered roughly equivalent, and the corresponding categories are redundant in meaning (i.e., category k and $k - 1$ represent the same quality level). When a CBD is negative, the adjacent category discriminations are reversed and the corresponding categories are disordered (i.e., category k represents less quality than category $k - 1$; Preston et al., 2011; Preston & Reise, 2015; Thissen et al., 2010). For example, if $a_{9,5}^{*}$ is negative, then category 5 would reflect less quality than category 4 for Item 9.

In addition to the CBDs, the scoring function value (SFV_{jk}) for category k within item j is of interest when examining category functioning. Thissen and colleagues (2010)'s parameterization of the model decomposes the discrimination for category k into an overall discrimination (a_j) for item j and an SFV_{jk}, that is,

$$a_{jk} = a_j \times \mathrm{SFV}_{jk}, \tag{3}$$

so that the SFV of interest is obtained through

$$\mathrm{SFV}_{jk} = \frac{a_{jk}}{a_j}. \tag{4}$$

Because the SFVs are reexpressions of the category discriminations, they also indicate category order, redundancy, and disorder in the same locations as the CBDs. However, the benefits of examining the SFV estimates are that: (a) they can be compared to the integer scores assigned to the categories by the scale developers, offering evidence regarding the extent to which these assigned scores are empirically supported; and (b) they accumulate the effects of redundancy or disorder in earlier categories (Preston et al., 2011; Thissen et al., 2010).

The NRM can also indicate when categories are underutilized through the category thresholds, which are reexpressions of the intercepts. The threshold (b_{jk}) for category *k* within item *j* is obtained through

$$b_{jk} = \frac{c_{j(k-1)} - c_{jk}}{a_{jk} - a_{j(k-1)}} \tag{5}$$

(Thissen et al., 2010). Formally, this threshold represents the required amount of quality for a classroom to have a .5 probability of being rated in category *k* given that the focus is on categories *k* and *k* − 1. This definition translates into the point along the quality scale where a classroom has an equal probability of being rated in these two categories. When all categories are ordered and sufficiently used, the category thresholds will be ordered (i.e., $b_{j2} < b_{j3} < \ldots < b_{jk} < b_{jm_j}$). Reversal of a pair of adjacent thresholds, $b_{j(k-1)} > b_{jk}$, and their equivalence, $b_{j(k-1)} = b_{jk}$, are both evidence for concluding that the lower of the pair of categories (*k* − 1) is underutilized. Ordering of a pair of adjacent thresholds, however, does not ensure that category *k* – 1 will be well used. This is because the reversals in other pairs of thresholds affect each category's use, as our results demonstrate below.

To specify the multiple-group aspect of the model, the quality levels were assumed to be distributed as a univariate normal with a mean (*M*) and standard deviation (*SD*) in the population for each dataset, that is,

$$\theta_{ig} \sim \mathrm{n}\left(\mu_g, \sigma_g\right), \tag{6}$$

where μ_g and σ_g are the *M* and *SD*, respectively, for the sample in dataset *g*. The *M* and *SD* for the first dataset (i.e., $g = 1$) were fixed to 0 and 1, respectively, for model identification.

The generalized partial credit model and the partial credit model. The GPCM is nested within the NRM and is achieved when the set of SFVs within each item (see Equation 3) increase in increments of 1 (i.e., from 0 to 6) rather than being freely estimated. Thus, only the overall item discriminations (a_j) are freely estimated in Equation 3. This constraint on the SFVs is how the assumption that the categories within an item are ordered and equally discriminating is incorporated into the GPCM.

The PCM is nested within the GPCM. Hence, the PCM is also nested within the NRM. The PCM is achieved from the GPCM by constraining the overall item discriminations to be equal across all items. That is, $a_1 = a_2 = \ldots = a_j = a_{36} = a$. Because of the constraints on the SFVs in the GPCM and the PCM, the formula for the category threshold can be simplified to,

$$\tilde{b}_{jk} = \frac{c_{j(k-1)} - c_{jk}}{a_j}. \tag{7}$$

For the PCM, a_j is further replaced with a in the denominator, to reflect the assumed equality of the overall item discriminations. We use the tilde to differentiate the GPCM and PCM thresholds from the NRM threshold (i.e., $\tilde{b}_{jk}$ vs. b_{jk}). This distinction reflects the different assumptions that the GPCM and PCM make about the categories compared to the NRM. As Equation 7 shows, all categories within an item have the same value in the denominator because of the constraints the GPCM and PCM need to assume that the categories are ordered and equally related to the measured trait. The NRM does not make either of these assumptions, as reflected in Equation 5.

In Equation 7, all categories for an item have the same value in the denominator because of the constraints needed to assume the categories are ordered. As a consequence, any disordering in the categories will manifest in the thresholds under the GPCM and PCM (Andrich, 2013). Equivalence and reversals of thresholds under the GPCM and PCM, then, can reflect one or more of the problems with category disorder, redundancy, and underutilization. In contrast, the NRM freely estimates the overall discriminations and SFVs, limiting its thresholds to more purely reflect underutilization.

Analytic steps. Our stacked analysis proceeded in the following manner to determine whether problems in the category functioning were occurring. In the first step, we compared the relative fit of the NRM, GPCM, and PCM models to the stacked dataset, using the likelihood ratio test and Akaike Information Criterion [AIC]. Next, we tested each item for whether the SFVs increased in increments of one unit (i.e., fixed the first through seventh SFVs to 0 through 6, respectively) while all other items' SFVs were freely estimated. Conceptually, this tested whether an item fit the GPCM (i.e., constraining categories to be ordered and equally discriminating on the measured trait) while treating all other items as fitting the NRM. We then compared the fit of each reduced model (i.e., the model with one item constrained) to the full NRM model, using the LRT and AIC. For the 36 different LRTs, we performed the Benjamini-Hochberg (1995) correction to reduce the chance of false discoveries (see Appendix F for details). We also performed a similar investigation in that we tested whether each item conformed to the NRM while setting all of the other items to fit the GPCM (we thank an anonymous reviewer for suggesting these approaches).

Next, we performed category-level examinations of all items flagged as problematic during the previous step, which involved inspecting the CBDs (and their 95% confidence intervals [*CI*s]) from the initial NRM analysis (see Appendix F for details on how the standard errors were obtained to form the *CI*s). We considered two adjacent categories as disordered when the upper limit of the 95% *CI* for their corresponding CBD was less than 0, clearly redundant when the 95% *CI* included 0, redundant because the categories were not being distinguished enough when the *CI* included values greater than 0 and up to 0.5, and ordered when the lower bound of the 95% *CI* was greater than 0.5. The reason for the range of 0 to 0.5 for redundant categories is because a CBD can be viewed as the discrimination for two adjacent categories

(Preston & Reise, 2014; Thissen et al., 2010). When the CBD is 0, the corresponding categories are indistinguishable. A convention for a CBD cutoff to indicate that the two categories are distinguished enough does not exist. Thus, we adopted 0.5 because researchers found that data generated with CBDs greater than 1.5 led to ordered data conforming to the Guttman pattern with unrealistic precision while data generated with CBDs less than 0.5 had unrealistic poor properties (Preston & Reise, 2014), suggesting that the generated data with CBD values set to less than 0.5 did not resemble ordinal data. We also examined whether the adjacent category thresholds under all three models were ordered, equivalent, or reversed (also using 95% *CI*s, but for the category thresholds in this case; see Appendix D).

In our parallel analysis, we only used the PCM because none of the individual datasets for this portion of our study had the 2,000 plus cases that Preston and Reise (2015) found was typically needed for adequate power to test for category order with the NRM. We established whether adjacent pairs of PCM thresholds were disordered, equivalent, or ordered within each dataset as in the stacked analysis. We then compared across the datasets to see whether disordered and equivalent pairs of adjacent thresholds occurred in similar category locations.

Within-category raw score averages and category-to-total point-biserial correlations. We also examined the within-category raw score averages and the category-to-total point-biserial correlations to investigate the impact any problems in the categories detected with the IRT models might have at the scale score level (Adams et al., 2012; Wetzel & Carstensen, 2014). For these calculations, we first followed standard ECERS-R scoring by averaging item scores to form a *total raw score* for each classroom (although we used the first 36 rather than all 43 items as noted above). We next repeated the following calculations separately for each item. To obtain the *within-category raw score averages* for each item, we identified the classrooms rated with the same category score and then averaged those classrooms' total raw scores. To obtain the *category-to-total point-biserial correlations* for each item, we correlated the total raw scores with dummy indicators as to whether a classroom was rated in each category (0 = no, 1 = yes). When higher categories represent increasing levels of quality, these within-category averages and point-biserial correlations should increase with the score categories. This monotonic increase is expected even though the point-biserial correlations will be negative for the lower scores because of the multi-category response structure of the ECERS-R (Adams et al., 2012; Mayer & Beckh, 2016).

Visual Comparison of the Results

We created category probability curves (CPCs) to visually represent the results of the stacked analyses and to emphasize the similarities and differences between the IRT models. In general, CPCs indicate the probability of a classroom being rated with each category of an item as a function of quality level. When all of the categories for an item are used sufficiently and in an ordinal manner, the item's CPCs resemble Figure 2a. The visual hallmark of ordered categories is when the curves peak at successively higher quality levels (e.g., the mode for category 3 is higher along the quality scale than the mode for category 2). This result coincides with positive CBDs. The visual hallmark of a category sufficiently utilized is also evident in Figure 2a: Each category's curve peaks above all other curves (i.e., is most probable) at some point along the quality scale. For instance, category 1's curve is above all other curves until a quality level of –2.3, at which point category 2 becomes the most probable; category 2 maintains this status until a quality level of –1.5, where category 3 becomes the most probable. The adjacent categories' curves intersect at the quality points just noted, which are also the values for the thresholds. Consistent with the sufficient utilization of each category in this example, the thresholds are ordered (i.e., the threshold for categories 2 and 3 [–1.5], is greater than the threshold for categories 1 and 2 [–2.3]).

In contrast to this ideal situation, Figure 2b includes the CPCs for an item where the categories are not functioning as intended. That is, categories 2 and 3 are disordered, with the mode for category 3's curve being located lower along the quality scale than the mode for category 2's curve. This is the visual hallmark of category disorder and is consistent with the negative CBD for this item's categories 2 and 3 ($a_3^* = -1.0$). Category 5, on the other hand, is underutilized. Here, the visual hallmark is the category's probability curve never being above all other curves, consistent with thresholds being reversed around focal category 5 ($b_5 = -0.5$ and $b_6 = -0.8$). In other words, the curves for categories 4 and 5 intersect at a higher point along the quality scale (−0.5) than the curves for categories 5 and 6 (−0.8). Note that the example in Figure 2b illustrates disorder at a different category than underutilization in order to simplify the presentation. In reality, both problems can be present in the same category, as we will see in the ECERS-R results. We provide additional examples of CPCs in Appendix C.

Software

The parameters of the NRM, GPCM, and PCM were estimated using flexMIRT (Cai, 2017). The within-category raw score averages and point-biserial correlations were calculated with Stata. We note two differences between the flexMIRT parameterization and our presentation. First, flexMIRT fixes the first and last categories' SFVs to zero and six, respectively, for model identification. To facilitate interpretation and place the SFVs in the context of ECERS-R scoring (i.e., 1 to 7), we present our results after adding a constant of 1. Doing so does not alter any of the conclusions about the SFVs. flexMIRT also parameterizes the thresholds differently from Equation 7. We used the notation above to emphasize the similarities and differences between the NRM, GPCM, and PCM, although again our notation does not alter conclusions about threshold order, equivalence, and reversal (see Appendix D for details on flexMIRT's parameterization).

Results

We first present results of the model comparison and results of the item-level tests based on the analyses of the stacked dataset. Then we present category-level results regarding the CBDs and thresholds. We end with the within-category raw scores, point-biserial correlations, and CPCs from the stacked analyses.

Model Comparisons and Item-Level Tests

For the stacked dataset, the NRM (AIC = 412,753) was strongly favored over the GPCM (AIC = 414,879) and the PCM (AIC = 417,734) based on the information criteria. The likelihood ratio test indicated that the NRM's improvement in model fit over the GPCM (χ^2[180] = 2,486, $p < .01$) and PCM (χ^2[215] = 5,410 , $p < .01$) were statistically significant. Between the GPCM and PCM, the information criteria favored the GPCM, and its improvement in model fit over the PCM was statistically significant, $\chi^2(35) = 2{,}924.34$, $p < .01$. The NRM being favored over the other two models suggests that a subset of items have categories that do not follow an ordinal progression and/or the categories within each of those items do not equally contribute to the measured trait.

Although the GPCM displayed greater model fit to the data over the PCM, we present the results from the PCM and reserve the results from the GPCM for Appendix E for the following three reasons. First, the two models produced very similar findings in terms of category threshold conditions (as detailed in the appendix). Second, the PCM matches the assumptions of the ECERS-R scale developers (i.e., the standard scoring uses the simple average of items). Finally, doing so allows our PCM results to be compared to prior published studies on the ECERS-R, which primarily have been based on the PCM when those studies used an IRT model for ordinal data.

Regarding the item-level tests, all items were statistically significant even after correcting for false discovery. This finding indicates that none of the items had SFVs that increased in increments of 1 (i.e., none of the items conformed to the GPCM). Conceptually, this means that none of the items had categories contributing equally to overall classroom quality, creating the possibility that disordered and redundant categories were present. Based on the item-level results, we proceeded with our analysis at the category-level for all items.

Nonorder in Category Boundary Discriminations and Scoring Function Values

The NRM identified extensive nonorder in the categories through the CBDs (values reported in Table 2). Of the 216 adjacent pairs of categories, there were 7 instances where the 95% *CI*s for the CBDs were below 0 (3%; superscript *a* in the table), which indicates that the categories corresponding to these CBDs were disordered. That is, the higher of the corresponding categories represented lower levels of quality than their immediately prior categories. There were 31 instances where the *CI*s for the CBDs included 0 (14%; superscript *b* in the table), which indicates that each pair of categories associated with these CBDs were clearly redundant. There were 150 instances where the *CI*s were above 0 but below or included 0.5 (69%; superscript *c* in the table). This indicates that the categories within each pair associated with each of these CBDs were not being distinguished enough to represent different levels of quality based on our cutoff of 0.5. Lastly, there were 55 instances where the *CI*s were above 0.5 (26%; superscript *d* in the table), which indicates that the categories within each pair associated with these CBDs were ordered.

Turning to the category locations and item types where these problems were most extensive, the CBDs that were consistently negative or overlapping zero were concentrated around where the odd-numbered categories of 3 and 5 were the higher of the adjacent pairs of categories. For instance, all seven negative CBDs (with 95% *CI*s below 0) were for the boundary discrimination between categories 2 and 3 (a_3^*), meaning these categories were disordered.

Fourteen of 31 CBDs with *CI*s that included zero fell at this same position, and an additional 10 of the 31 occurred with the CBD corresponding to categories 4 and 5 (a_5^*), reflecting that these categories represented similar levels of quality (i.e., redundant categories). Regarding item types, negative and small CBDs were particularly concentrated in the items pertaining to children's personal care routines (Items 9 to 14). Four of these items had negative CBDs. The other two items had two and three CBDs with *CI*s that included zero.

The estimated SFVs are also in Table 2. As previously noted, the SFVs are reexpressions of the CBDs, and thus identify the same problems with redundancy and disorder. However, by expressing the problems on the metric of the ECERS-R standard integer category scores, the SFVs highlight the extent to which the empirical estimates deviate from the instrument scoring. For example, the negative CBD of −0.41 for categories 2 and 3 within Item 10 has little substantive meaning in contrast to the intuitive problem evident because category 3's estimated SFV of 1.07 is substantially lower than its assigned raw score value of 3. Table 2 also illustrates the way in which SFVs accumulate problems of redundancy and disorder in preceding categories, adding to the information evident in the CBDs alone. For instance, looking at Item 22, categories 3 and 4 are ordered (CBD = 0.78, with *CI* = [0.54, 1.02]) but the latter category's estimated SFV of 3.09 is well short of the assigned value of 4. This discrepancy reflects the preceding categories' CBDs for this item being small or negative. In contrast, all of the CBDs for Item 18 fell in the positive range, with point estimates above 0.5 (although just two exceeded 0.50 based on the *CI*s). Consistent with these values, Item 18's SFVs are all close to the assigned values (within .30 points; e.g., SFV = 2.70 for category 3).

Nonorder in Category Thresholds

Analysis of the stacked dataset. The category threshold estimates produced during the analysis of the stacked dataset with the PCM are in Table 3. The superscripts indicate problems of reversal (#) or equivalence (^) between the marked value and the immediately preceding threshold value. Nonorder in adjacent thresholds consistently occurred in two locations: for $\tilde{b}_4$ versus $\tilde{b}_3$ (reflecting a problem with category 3) and for $\tilde{b}_6$ versus $\tilde{b}_5$ (reflecting a problem with category 5). Two-thirds of the items had threshold reversals at the former location, and over 90% in the latter location. Reversal – and equivalence – of thresholds also sometimes occurred with those that bounded category 2 ($\tilde{b}_3$ vs. $\tilde{b}_2$) and category 6 ($\tilde{b}_7$ vs. $\tilde{b}_6$), but never for thresholds bounding category 4 ($\tilde{b}_5$ vs. $\tilde{b}_4$).

As noted earlier, the PCM threshold estimates may reflect problems introduced by assuming the categories are ordered when they are truly disordered, whereas the NRM can separately detect category disorder through the CBDs and category underutilization through the thresholds. In the case of the ECERS-R, the NRM revealed extensive problems in the threshold locations as well as in the CBDs. That is, under the NRM, every item had nonordered thresholds. The pattern of problematic thresholds under the NRM was similar to those observed under the PCM, with most items having two or more nonordered thresholds that typically pointed to underutilization of categories 3 and 5. However, the NRM showed more evidence of categories 2 and 4 being underutilized than the PCM did, where these categories were also considered to be disordered based on the CBDs. In other words, the forced ordering of the SFVs under the PCM when the categories are disordered resulted in category underutilization appearing at the higher category.

Analysis of each dataset (parallel analysis). As previously noted, the sample size precluded analyzing each dataset with the NRM, but our parallel analysis with the PCM confirmed that nearly every item had at least one instance of reversed thresholds for most of the 14 replicate datasets/waves. More specifically, this was true for the majority of items: 90-100% in 10 of the datasets/waves, 81% of items in PCER, and 64% in FF. Category threshold equivalence was more frequently observed in the QUINCE dataset, most likely because of its small sample size leading to standard errors that were over twice the size of those seen in other datasets, and in turn resulting in more overlap between adjacent confidence intervals. Nevertheless, in QUINCE, half of the items still had at least one pair of reversed category thresholds in the first wave of data collection as did one-quarter of the items in the second wave.

The locations of the threshold reversals were also consistent across datasets/waves. Every sample replicated problems of reversals between thresholds $\tilde{b}_6$ and $\tilde{b}_5$ for two-thirds of the items and between thresholds $\tilde{b}_4$ and $\tilde{b}_3$ for two-fifths of the items. These locations matched the places where threshold reversals commonly occurred in the stacked analysis. In contrast, reversal was evident in just seven to twelve percent of the items for thresholds $\tilde{b}_7$ versus $\tilde{b}_6$ and thresholds $\tilde{b}_3$ versus $\tilde{b}_2$, respectively, across the replicates. No instances of reversal between thresholds $\tilde{b}_5$ and $\tilde{b}_4$ occurred in the 14 replicates, which was also consistent with the stacked analysis.

Within-Category Averages and Category-Total Point-Biserial Correlations

The within-category averages of the raw scores are presented in Table 4. In nearly three-quarters (26 of 36) of the items, these values did not consistently increase from the lower to the upper of two adjacent categories (i.e., nonorder). Nonorder most frequently occurred for category 3 (18 items) and category 5 (8 items), and was especially evident for the personal care routines (Items 9 to 14). These findings were consistent with the location of problems in the category functioning identified by the NRM, GPCM, and PCM.

Table 4 also includes the category-total point-biserial correlations. The correlations did not monotonically increase for two-thirds (24 of 36) of the items. These violations frequently occurred around category 3, which was consistent with our other findings. In contrast, the point-biserial correlations were generally ordered in the upper categories. These differences could be because each point-biserial correlation uses data from all classrooms (with every classroom not in a focal category being in the reference group), whereas the within-category averages and the IRT parameters focus just on classrooms in each pair of adjacent categories. Another possible reason for the different results could be the lack of a conventional cutoff for the needed increments in point-biserial correlations to signal redundant categories, such as those observed in the upper categories of the ECERS-R (e.g., Item 9 had nearly identical values for categories 5 and 6).

Graphs of Category Probability Curves

We created CPCs for the three items we used in the introduction to explain why problems in the categories may exist. The CPCs provide a visual representation of category order (Item 18, *Informal use of language*), redundancy (Item 9, *Greeting/departing*), and disorder (Item 10, *Meals/snacks*), with all CPCs also evidencing category underutilization. We present the CPCs produced with the NRM and the PCM side-by-side for each item to facilitate comparisons.

Earlier, we saw that Item 18's CBDs and within-category averages reflected the most consistent category order, although the PCM identified reversed thresholds and the point-biserial correlations decreased from category 2 to 3. Visually, the ordered categories are evident in the CPCs from the NRM because the modes increase with the categories (see Figure 2c). Underutilization is also evident as only three categories (i.e., 1, 4, and 7) are the most probable across a noticeable segment of the quality scale, with one additional category only being the most probable for a very minor segment of the scale (i.e., category 3). The CPCs from the PCM for the most part follow similar form to the CPCs under the NRM, although the probability curve for category 3 is more suppressed under the PCM (see Figure 2d). The differences reflect the fact that even though the NRM detected order, its estimated CBDs and SFVs were not identical to the values assumed by the PCM (and the ECERS-R scoring). Only categories 1, 4, and 7 being the most probable in Figure 2d is because of the nonorder in four pairs of adjacent PCM thresholds, which reflects underutilization of categories 2, 3, 5, and 6. These results are also consistent with our examination of the indicator content, where we previously noted that this item mixed basic and advanced content in the higher categories (this item also has the issue of complementary indicators at the lower categories, which we exemplified with a different item in the introduction).

Turning to Item 9, we noted earlier that we found considerable evidence of redundancy, with the confidence intervals for three CBDs overlapping zero as well as four instances of equivalent within-category averages. In the CPCs from the NRM (Figure 2e), the redundancy is reflected where two curves peak at similar points along the quality scale, such as categories 2 and 3 (both peaking near −4). The curve for category 3 also closely tracks the curves for categories 1 and 2. These results are consistent with the highly complementary content between categories 1 and 3 for this item. The PCM constraints are evident in Figure 2f, where categories 2 and 3 are forced to be ordered and the mode for category 3 is now located higher along the quality scale than the mode for category 2. Although the reversed thresholds under the PCM may be affected by this forced category order, category underutilization is also evident under the NRM where only categories 1, 2, 4, and 7 are most probable.

Regarding Item 10, we noted earlier that the NRM detected disorder between categories 2 and 3 as well as considerable category redundancy (the confidence interval for one CBD including zero and three CBDs including 0.5). Although many of the curves for this item are also suppressed, the disorder of categories 2 and 3 is evident in that the mode for category 3’s curve is lower along the quality scale than the mode for category 2 (the former being off screen, at about –4.00, versus the later peaking just below –1.00; Figure 2g). The PCM forces categories 2 and 3 to be ordered, with the mode for the latter category’s probability curve now being higher along the quality scale than the mode for category 2 (just above versus just below –1.00; Figure

2h). Yet, in the NRM and the PCM, all but two of the categories are underutilized, with only categories 1 and 7 being most probable at some point. This item is essentially dichotomous, consistent with the possibility discussed in the introduction that the stringent sanitation requirements in category 1 may restrict many classrooms to that category. Once classrooms move out of this lowest category, they appear to jump to the highest category, in line with our item content review and prior studies that suggest these classrooms may possess higher levels of other aspects of quality.

Discussion

Our study provides empirical evidence of problems with the category functioning that we anticipated based on our examination of the ECERS-R manual. We also advance on the handful of prior studies on this topic by using multiple analytic strategies (i.e., NRM, GPCM, PCM, within-category raw averages, and point-biserial correlations) and approaches (parallel and stacked analyses) involving eight datasets with 14 waves. Problems in category functioning were consistently evident across items, datasets, analyses, and approaches, and our comprehensive analysis helped pinpoint the locations and types of problems. For instance, problems were consistently evident with categories 3 and 5, likely reflecting the instrument's complex stop-scoring rules as we described in the introduction. For many items, the problems detected in category functioning reflected category underutilization and category redundancy. For other items – especially those capturing children's personal care routine items (Items 9 to 14) – the problems included category disordering. Regardless of the category functioning problems, the fact that the SFVs deviated from the scale developers' assigned scores for all items indicate that all categories within an item do not contribute equally to the measured trait (Preston & Reise, 2014). This finding, along with our other rigorous psychometric results, has important implications for using averages of ECERS-R developer-assigned scores for research and policy purposes.

As Preston and Reise (2014) cautioned in situations of small CBDs (which are based on the SFVs) like we found for the ECERS-R, "when category distinctions fail to discriminate, a researcher would not want to use a scoring strategy that aggregates raw integer item scores" (p. 392). Our findings raise concern with the current use of averaged scores for consequential decisions, echoing findings from earlier descriptive studies of the instrument (e.g., Hofer, 2010). In terms of research, the raw scores include error from the categories within an item not following an ordinal progression and equally discriminating. These could be contributing factors for the very small effect sizes between ECERS-R raw averages and child outcomes that are frequently reported.

Our study contributes to the literature on the category functioning of the ECERS-R items in several important ways. First, our study used parallel analysis to replicate findings across different datasets, indicating that the problems observed in the category functioning occurred in data from a range of different samples and data collection teams. This replication shows that the small set of published research demonstrating problems with the ECERS-R categories was not due to their unique samples. This replication drives our second major contribution because our samples included settings that are of focus in current policy efforts. As a result, our findings have direct implications for current use of the ECERS-R. A third contribution is that we were able to use the NRM and PCM in our stacked analyses. Doing so allowed us to differentiate the extent to which problems identified in prior PCM-based studies reflect only underutilization of a category versus also reflecting disorder and redundancy of categories.

Our findings are consistent with, but importantly extend, the small set of prior psychometric studies of the ECERS-R. For instance, our finding that category disorder occurred most often for the children's personal care routines items is consistent with Gordon and colleagues' (2015) indicator-level Rasch analyses. Their study revealed that nearly two-thirds of the indicators for these items were empirically ordered in a manner that differed from the category score where the instrument developers had placed them. Our current findings are also consistent with prior studies (Gordon et al., 2013; Mayer & Beckh, 2016), where reversed thresholds under the PCM were interpreted as problems in the categories stemming from the stop-scoring rule combined with the greatest mixing of indicators that tap into different aspects of quality for these personal care routines items. In addition to this potential problem of mixing indicators, in the current study, we also highlighted ways in which the broader mixing of basic and advanced indicators – along with the presence of complementary indicators – might limit, if not preclude, the use of certain categories. We also found that problems accumulated to the scale score level, with all but four items having disordered within-category averages or point-biserial correlations, replicating the single-study evidence of each problem in prior studies (Gordon et al., 2015 for averages; Mayer & Beckh, 2016 for correlations).

Although our study used multiple analytic strategies to establish problems in the category functioning of the ECERS-R items across many different datasets, we note some limitations. One limitation is that we could not use the NRM on the individual datasets because of their small sample sizes. In the parallel analysis with the PCM, we had to collapse categories with low frequencies in some datasets/waves, primarily the smallest QUINCE dataset where we saw more equivalent thresholds than reversed thresholds. Collapsing categories did not appear to have an impact on threshold disordering because, in the items that did not require any collapsing, the nonordering in the thresholds appeared in the same category locations as observed in the stacked analysis, which did not require any category collapsing. We encourage future replication studies with sufficiently large samples to confirm that collapsing does not affect threshold conditions during an IRT analysis. It is also the case that many of our datasets included primarily lower-income children. Although these datasets were an advantage because these children are often the target of policy and we found that all categories were used in our stacked dataset, additional replication with diverse samples is warranted. Such studies may wish to proceed in a two-step approach similar to what we used, especially when each dataset lacks the sufficient sample size for the NRM. The first step could include a parallel analysis of the datasets, using the PCM. If the category problems replicated across datasets, then the NRM could be fitted to the stacked dataset to differentiate issues of category redundancy, disordering, and underutilization. The stacked analyses could also include calculating within-category means and point-biserial correlations to inform how item-level problems accumulate to the scale score level.

Another limitation of this study is that we did not have access to indicator-level data, particularly data with all indicators scored rather than stop-scored. Analyzing complete indicator-level data could further illuminate the reasons for the problems in the category functioning that our study detected. Such indicator-level analysis could also inform alternative scoring systems for the ECERS-R (and the new ECERS-3) as well as further refinement of item content (e.g., Clifford, Sideris, & Neitzel, 2012). Finally, limited simulation and empirical studies exist for using the NRM to examine the category functioning of rating scale items. Particularly challenging for applied scholars is how to determine when a positive CBD is too close to zero to reflect a meaningful distinction between categories (i.e., redundancy). Regardless of whether a clear upper cutoff currently exists for CBDs to indicate order, CBDs of 0 or less than 0 are clearly problematic, of which there were many in our study. We encourage further methodological work to establish guidance regarding whether CBDs are large enough to indicate that their corresponding categories are sufficiently distinguished.

Although the new ECERS-3 manual advises users to consider scoring all indicators, it still retains the stop-scoring approach in its standard scoring guidelines and training materials and does not offer a specific scoring strategy based on all of the indicators. We recommend that practitioners, researchers, and policymakers move to alternative scoring methods (for both the ECERS-R and the ECERS-3) that yield quality estimates that are reliable and valid for research and policy use. By integrating models such as the NRM, GPCM, and PCM into iterative scale development, improved measures may yield larger correlations with children's school readiness. If the stop-scoring approach is retained in future scale revisions, empirical evidence demonstrating that the indicators are ordered as organized within item categories should be produced, along with other reliability and validity evidence. Until then, our results combined with those currently documented in the literature caution against using the ECERS-R with the stop-scoring rule for research, policy, and practice.

References

Ackerman, D. J. (2014). *State-funded prek policies on external classroom observations: Issues and status.* Princeton, NJ: Educational Testing Service.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*, 547−573.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "Threshold Disorder Controversy." *Educational and Psychological Measurement, 73*, 78−124.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29−51.

Burchinal, M., Zaslow, M., Tarullo, L. (2016). Quality thresholds, features, and dosage in early care and education: Secondary data analyses of child outcomes. *Monographs of the Society for Research in Child Development.*

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. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle. (Eds.), *Quality measurement in early childhood settings* (pp. 11−31). Baltimore, MD: Brookes Publishing.

Cai, L. (2017). flexMIRT® version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Child Trends. (2015). *The QRIS Compendium.* Retrieved from http://qriscompendium.org/about/

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, 2012).

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44,* 365−380.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50,* 2417−2425.

Frank Porter Graham Child Development Institute. (2003). A whole new yardstick. *Early Developments, Vol. 7 [Rating early childhood environments],* 8–11.

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.

Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N. C., Kaestner, R., & Korenman, S. (In Press). Identifying high-quality preschool programs: New evidence on the validity of the Early Childhood Environment Rating Scale–Revised (ECERS-R) in relation to school readiness goals. *Early Education and Development*, *26*, 1086−1110.

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.

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

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14,* 150−164.

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.

Layzer, J. I., & Goodson, B. D. (2006). The quality of early care and education settings: Definitional and measurement issues. *Evaluation Review, 30,* 556−576.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM.

Marcoulides, K. M., & Grimm, K. J. (2016). Data integration approaches to longitudinal growth modeling. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164416664117

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149−174.

Mayer, D. & Beckh, K. (2016). Examining the validity of the ECERS-R: Results from the German National Study of Child Care in Early Childhood. *Early Childhood Research Quarterly, 36*, 415−426.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Pew Charitable Trusts. (2014). *Pre-K Now.* Retrieved from http://www.pewtrusts.org/en/archived-projects/pre-k-now

Preston, K. S. J. & Reise, S. P. (2014). Detecting faulty within-item category functioning with the nominal response model. In S. P. Reise & D. A. Revicki. (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386−405). New York: Routledge.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*, 523−550.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini. (Eds.), *Handbook of polytomous item response theory models* (pp. 43-75). New York, NY: Routledge.

U.S. Department of Education. (2013). *Education department announces next rounds of race to the top, including another key investment to expand access to high-quality early learning opportunities.* Retrieved from http://www.ed.gov/news/press-releases/education-department-announces-next-rounds-race-top-including-another-key-invest

Wetzel, E. & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment, 21*, 765−774.

Table 1

Demographics and Sampling Design for Each Dataset/Wave

	FACES 2000 N/%	FACES 2003 N/%	HSIS N/%	EHSRE N/%	QUINCE N/%	PCER N/%	FF N/%	ECLS-B N/%
Number of classrooms with ECERS-R								
Wave 1	267	326	915	984	81	310	365	800
Wave 2	261	305	747	-	56	310	-	-
Wave 3	195	-	-	-	-	-	-	-
Center characteristics (%)								
Head Start	100	100	71	45	14	31	18	40
State Pre-K or Public School	0	0	16	n/a	6	58	11	25
Other	0	0	21	n/a	80	11	72	35
Characteristics of children served (%)								
Low income	93	94	94	98	41	5	65	100
Female	50	52	50	49	52	49	45	50
Race/Ethnicity								
Hispanic	29	28	35	24	11	16	11	24
Non-Hispanic White	39	30	26	36	44	34	14	28
Non-Hispanic Black	27	35	33	36	31	42	59	34
Non-Hispanic Other	5	7	6	3	14	8	16	14
Years of ECERS-R Observations	2000-2001	2003-2004	2002-2003	2001-2003	2004-2005	2004	2001-2004	2004-2005
Target Population	Nationally representative samples of Head Start classrooms.		Nationally representative sample of Head Start classrooms plus classrooms where comparison group children enrolled.	Classrooms attended by children originally eligible for 17 Early Head Start programs.	Classrooms served by 24 CCR&R agencies in 5 states.	Twelve research teams in about one dozen states recruited centers/ classrooms.	Classrooms attended by children originally sampled from hospitals in 20 large U.S. cities, with an oversample of non-marital births.	Classrooms attended by children originally sampled from birth records in most states. We focused on low-income children.

Note. Sample sizes for PCER are rounded to the nearest 10, and the ECLS-B to the nearest 50, per NCES reporting requirements. Low income is defined as below 200% federal poverty guideline. n/a = Only Head Start funding is available in EHSRE; school location and state pre-k funding source are not known. FACES=Head Start Family and Child Experiences Survey; HSIS=Head Start Impact Study; EHSRE=Early Head Start Research and Evaluation Project; QUINCE= The Quality Interventions for Early Care and Education; PCER=Preschool Curriculum Evaluation Research Initiative; FF=Fragile Families and Child Wellbeing Study; ECLB-B=Early Childhood Longitudinal Study-Birth Cohort. CCR&R = child care resource and referral.

Table 2

Category Boundary Discriminations and Scoring Function Values from the Stacked Analysis with the Nominal Response Model

	Category Boundary Discriminations						Scoring Function Values (SFV)						
Item Labels (Abbreviated)	a^*_{j2}	a^*_{j3}	a^*_{j4}	a^*_{j5}	a^*_{j6}	a^*_{j7}	1	2	3	4	5	6	7
Space and Furnishings													
Item 1: Indoor space	0.39[c]	0.33[c]	0.39[c]	−0.11[b]	0.33[c]	0.38[c]	1	2.36	3.53	4.89	4.50	5.67	7
Item 2: Routine care furniture	0.54[c]	0.07[b]	0.42[c]	0.47[c]	0.05[b]	0.64[d]	1	2.48	2.67	3.82	5.10	5.23	7
Item 3: Comfortable furnishings	0.42[c]	−0.01[b]	0.35[c]	0.63[c]	0.35[c]	0.71[d]	1	2.02	2.01	2.87	4.42	5.27	7
Item 4: Room play friendly	0.69[c]	0.17[b]	0.54[c]	0.03[b]	0.61[c]	0.71[d]	1	2.50	2.88	4.06	4.13	5.45	7
Item 5: Privacy space	0.44[c]	0.26[c]	0.37[c]	0.45[c]	0.36[c]	0.67[d]	1	2.03	2.65	3.52	4.57	5.43	7
Item 6: Child−related display	0.23[b]	0.30[c]	0.56[c]	0.13[b]	0.40[c]	0.50[c]	1	1.65	2.50	4.10	4.46	5.58	7
Item 7: Gross motor space	0.10[b]	0.24[c]	0.19[c]	0.20[c]	0.40[c]	0.37[c]	1	1.39	2.35	3.12	3.90	5.50	7
Item 8: Gross motor equipment	0.19[c]	0.18[c]	0.23[c]	0.25[c]	0.19[c]	0.53[c]	1	1.73	2.42	3.30	4.25	4.99	7
Personal Care Routines													
Item 9: Greeting/departing	0.13[b]	0.12[b]	0.39[c]	−0.04[b]	0.24[c]	0.84[d]	1	1.46	1.90	3.30	3.17	4.01	7
Item 10: Meals/snacks	0.43[c]	−0.41[a]	0.72[c]	0.17[b]	0.47[c]	0.65[d]	1	2.27	1.07	3.19	3.69	5.07	7
Item 11: Nap/rest	0.47[c]	−0.16[b]	0.85[d]	0.00[b]	0.43[c]	0.68[c]	1	2.25	1.82	4.06	4.05[a]	5.20	7
Item 12: Toileting	0.55[c]	−0.35[a]	0.61[c]	−0.08[b]	0.70[c]	0.35[c]	1	2.86	1.66	3.73	3.47	5.83	7
Item 13: Health	0.68[c]	−0.32[a]	0.65[c]	0.02[b]	0.57[c]	0.67[d]	1	2.80	1.95	3.67	3.71	5.21	7
Item 14: Safety	0.46[c]	−0.63[a]	0.53[c]	0.36[c]	0.32[c]	0.70[d]	1	2.59	1.56	2.27	3.49	4.60	7
Language−Reasoning													
Item 15: Books	0.54[c]	−0.09[b]	0.98[d]	0.62[c]	0.42[c]	0.77[d]	1	2.00	1.82	3.64	4.79	5.57	7
Item 16: Child communication	0.86[c]	0.44[c]	1.07[d]	0.35[c]	0.97[d]	0.77[d]	1	2.15	2.75	4.18	4.66	5.96	7
Item 17: Language reasoning	0.75[d]	0.47[c]	0.47[c]	0.41[c]	0.39[c]	1.00[d]	1	2.29	3.09	3.90	4.59	5.27	7
Item 18: Informal use of language	0.51[c]	0.59[c]	1.03[d]	0.55[c]	0.50[c]	0.69[d]	1	1.79	2.70	4.30	5.15	5.93	7
Activities													
Item 19: Fine motor	0.82[d]	0.55[c]	0.76[d]	0.15[b]	0.97[d]	0.86[d]	1	2.19	3.00	4.11	4.33	5.74	7
Item 20: Art	0.97[d]	0.42[c]	1.22[d]	0.83[d]	0.38[c]	0.89[d]	1	2.24	2.77	4.33	5.39	5.87	7
Item 21: Music	0.55[c]	0.39[c]	0.49[c]	0.64[c]	0.59[c]	0.85[d]	1	1.95	2.61	3.44	4.54	5.54	7
Item 22: Blocks	0.49[c]	−0.07[b]	0.78[d]	0.73[d]	0.83[d]	0.69[d]	1	1.85	1.72	3.09	4.36	5.80	7
Item 23: Sand/water	0.38[c]	0.11[b]	0.49[c]	0.47[c]	0.32[c]	0.52[c]	1	2.00	2.28	3.58	4.80	5.64	7
Item 24: Dramatic play	0.39[c]	0.33[c]	0.86[d]	0.78[d]	0.87[d]	0.95[d]	1	1.56	2.04	3.28	4.40	5.64	7
Item 25: Nature/science	0.82[d]	0.27[c]	0.53[c]	1.06[d]	0.09[b]	0.73[c]	1	2.41	2.87	3.78	5.59	5.75	7
Item 26: Math	0.78[c]	0.32[c]	0.97[d]	1.03[d]	0.06[b]	1.17[d]	1	2.08	2.53	3.88	5.30	5.38	7
Item 27: Multimedia use	0.65[c]	−0.23[a]	0.79[d]	0.34[c]	0.37[c]	0.79[d]	1	2.43	1.92	3.67	4.43	5.25	7
Item 28: Diversity acceptance	0.20[c]	0.32[c]	0.34[c]	0.51[c]	0.33[c]	0.75[d]	1	1.49	2.27	3.10	4.35	5.17	7
Interaction													
Item 29: Gross motor supervision	0.10[b]	0.08[b]	0.74[d]	0.45[c]	0.63[c]	0.80[d]	1	1.21	1.38	2.97	3.93	5.28	7
Item 30: General supervision	0.41[c]	0.21[b]	0.63[c]	0.34[c]	0.59[c]	1.01[d]	1	1.77	2.17	3.35	3.98	5.10	7
Item 31: Discipline	0.59[c]	0.18[b]	0.62[c]	0.85[d]	0.60[c]	1.13[d]	1	1.89	2.17	3.11	4.39	5.29	7
Item 32: Staff−child interactions	0.51[c]	0.08[b]	0.42[c]	0.11[b]	0.48[c]	0.91[d]	1	2.22	2.42	3.43	3.70	4.83	7
Item 33: Child−child interactions	0.83[d]	0.21[b]	0.75[c]	0.62[c]	0.44[c]	0.90[d]	1	2.32	2.66	3.87	4.86	5.56	7
Program Structure													
Item 34: Schedule	0.94[d]	−0.73[a]	0.94[d]	0.28[c]	0.83[d]	0.99[d]	1	2.74	1.39	3.13	3.64	5.17	7
Item 35: Free play	1.34[d]	−0.68[a]	1.30[d]	0.65[c]	0.66[c]	1.23[d]	1	2.79	1.89	3.62	4.48	5.36	7
Item 36: Group time	0.32[c]	0.26[b]	0.65[c]	0.57[c]	0.31[c]	0.98[d]	1	1.62	2.13	3.39	4.49	5.10	7

Note. Values are category boundary discriminations (a^*_{jk}) and scoring function values (SFV) from the nominal response model. All item-level tests were statistically significant (Benjamini-Hochberg adjusted p-values are in Appendix F). That is, the model fit worsened for each item when the item was treated to fit the GPCM while all other models were specified to fit the NRM. Superscript letters reflect: (a) *CI* for $a^*_{jk} < 0$; (b) *CI* overlaps 0; (c) *CI* greater than 0 and including values up to 0.5; and (d) *CI* greater than 0.5. The CBDs reflect a linear transformation of the difference in SFVs and thus parallel conclusions apply for the difference in adjacent SFVs divided by a_j (the overall item discrimination). We added a constant of one to the SFVs to align with the standard ECERS−R scoring. Results are from the analysis of the first or only waves of the eight datasets stacked together, n = 4,048 classrooms.

Table 3

Thresholds from the Stacked Analysis with the Partial Credit Model

	PCM Thresholds					
Item Labels (Abbreviated)	$\tilde{b}_{j2}$	$\tilde{b}_{j3}$	$\tilde{b}_{j4}$	$\tilde{b}_{j5}$	$\tilde{b}_{j6}$	$\tilde{b}_{j7}$
Space and Furnishings						
Item 1: Indoor space	−1.65	−1.22^	−4.61#	4.09	−3.13#	−2.20
Item 2: Routine care furniture	−1.76	−0.69^	−5.95#	−1.65	−2.06^	−1.92^
Item 3: Comfortable furnishings	−0.77	−3.78#	−0.88	2.71	−1.54#	0.30
Item 4: Room play friendly	−3.00	−2.37^	−1.96^	0.87	−2.30#	−1.30
Item 5: Privacy space	0.24	−4.14#	−0.18	2.15	−0.68#	−0.30^
Item 6: Child−related display	−7.09	−2.88	−0.68	2.90	−0.38#	2.39
Item 7: Gross motor space	−3.27	1.10	−2.82#	1.50	−0.16#	0.52
Item 8: Gross motor equipment	−2.67	2.12	−1.58#	1.50	−1.61#	−1.45^
Personal Care Routines						
Item 9: Greeting/departing	−2.35	−1.56^	−2.89#	2.48	−2.24#	−4.15#
Item 10: Meals/snacks	0.78	3.98	−4.56#	0.55	−0.97#	−1.03^
Item 11: Nap/rest	−2.09	1.89	−4.06#	4.03	0.82#	−1.41#
Item 12: Toileting	0.47	3.92	−4.72#	3.57	−3.98#	−1.31
Item 13: Health	−4.87	5.68	−3.92#	1.63	−2.46#	−2.05^
Item 14: Safety	−0.16	3.31	−3.68#	2.85	−1.67#	−3.05#
Language−Reasoning						
Item 15: Books	−1.60	−1.73^	−5.43#	6.00	−1.03#	−2.09#
Item 16: Child communication	−1.72	−3.41#	−3.96^	2.03	−3.57#	0.32
Item 17: Language reasoning	−1.75	−2.82#	−0.85	2.20	0.34#	−1.41#
Item 18: Informal use of language	−0.71	−4.12#	−3.74^	3.18	−1.83#	−2.04^
Activities						
Item 19: Fine motor	−2.09	−2.47^	−3.72#	3.63	−2.24#	−1.33
Item 20: Art	−3.22	−1.78	−1.53^	3.47	−1.01#	0.13
Item 21: Music	−5.87	−0.01	−1.32#	2.84	0.63#	1.43
Item 22: Blocks	−0.63	−0.38^	−5.70#	2.57	−1.99#	3.50
Item 23: Sand/water	2.24	−3.81#	−1.57	2.48	−0.74#	1.55
Item 24: Dramatic play	−3.52	0.01	−3.93#	2.49	0.22#	3.97
Item 25: Nature/science	−2.01	0.97	−2.32#	5.73	−0.31#	−0.48^
Item 26: Math	0.66	−4.36#	−3.33	4.02	0.42#	0.06^
Item 27: Multimedia use	−2.27	2.61	−4.54#	3.07	−0.50#	1.31
Item 28: Diversity acceptance	−1.59	−2.99#	−1.32	3.06	1.41#	0.01#
Interaction						
Item 29: Gross motor supervision	−0.41	−0.90^	−3.84#	0.06	0.75	0.48^
Item 30: General supervision	−1.40	0.27	−3.76#	−0.16	−0.83#	−1.42#
Item 31: Discipline	−1.36	−1.24^	−3.04#	−1.15	−0.40	−0.33^
Item 32: Staff−child interactions	−1.22	0.07	−3.80#	2.97	−3.86#	−4.37^
Item 33: Child−child interactions	−2.25	−0.54	−4.13#	2.39	−4.64#	−0.38
Program Structure						
Item 34: Schedule	−5.59	4.15	−5.57#	3.88	−2.60#	−1.83
Item 35: Free play	−2.53	−0.22	−4.09#	2.26	−1.88#	−0.95
Item 36: Group time	1.90	−4.22#	−2.58	1.10	−2.11#	−1.93^

Note. Values are based on the slope−threshold specification of the partial credit model, where monotonically increasing values indicate ideal category functioning. Thresholds begin with category 2, because they are defined relative to the immediately prior category (see Equation 7). # Value is reversed in relation to the threshold just below it (lower *CI* bound of lower threshold above upper *CI* bound of higher threshold). ^ Value is statistically equivalent to the threshold just below it (i.e., overlapping confidence intervals). Results are from the analysis of the first or only waves of the eight datasets stacked together, *n* = 4,048 classrooms.

Table 4

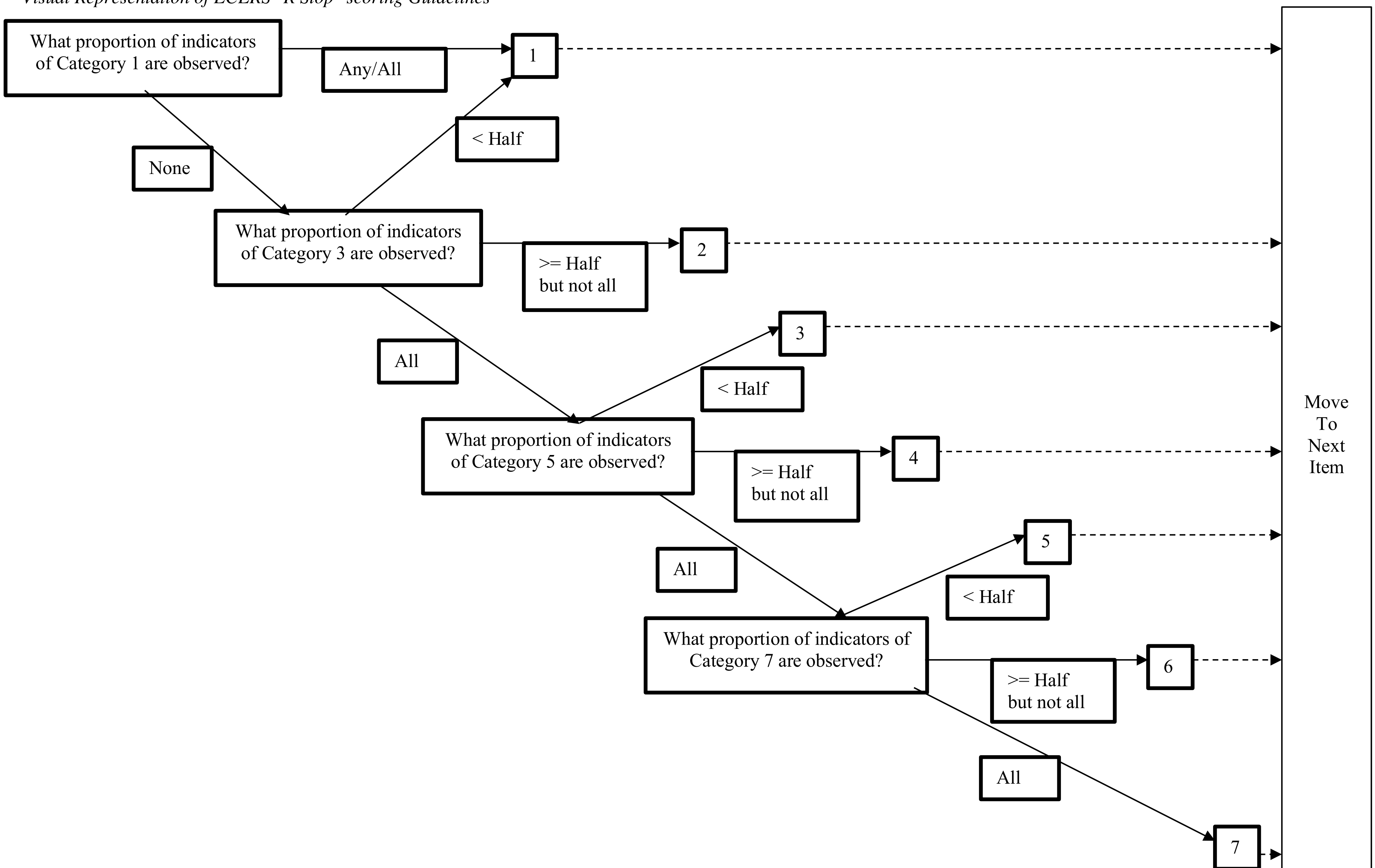
Within−Category Raw Score Averages and Category−Total Point Biserial Correlations from the Stacked Analysis

	Within−Category Average (*M*)							Category−Total Point−Biserial Correlation (*r*)						
Item Labels (Abbreviated)	M_1	M_2	M_3	M_4	M_5	M_6	M_7	r_1	r_2	r_3	r_4	r_5	r_6	r_7
Space and Furnishings														
Item 1: Indoor space	3.16	3.69	4.14	4.59	4.57^	4.92	5.30	−0.31	−0.22	−0.13	−0.12	−0.04	0.03	0.37
Item 2: Routine care furniture	2.64	3.19	3.22^	3.78^	4.05^	4.48	5.19	−0.25	−0.16	−0.11	**−0.23**	−0.14	**−0.16**	0.43
Item 3: Comfortable furnishings	3.66	4.13	4.14^	4.48	5.06	5.34	5.80	−0.27	−0.15	**−0.29**	−0.16	0.06	0.20	0.45
Item 4: Room play friendly	2.79	3.55	3.67^	4.31	4.39^	4.87	5.44	−0.29	−0.25	**−0.27**	−0.18	−0.10	0.02	0.50
Item 5: Privacy space	3.56	3.98	4.30	4.64	5.07	5.34	5.76	−0.35	−0.17	**−0.24**	−0.08	0.06	0.18	0.45
Item 6: Child−related display	3.67	3.91^	4.30	4.88	5.07	5.44	5.85	−0.10	**−0.27**	−0.26	0.03	0.07	0.23	0.30
Item 7: Gross motor space	3.89	4.04^	4.43	4.66	4.93	5.32	5.67	−0.22	**−0.30**	−0.10	−0.08	0.03	0.19	0.36
Item 8: Gross motor equipment	3.84	4.10	4.39	4.63	4.91	5.11^	5.57	−0.29	**−0.32**	−0.10	−0.06	0.02	0.10	0.46
Personal Care Routines														
Item 9: Greeting/departing	3.46	3.64^	3.84^	4.26	4.31^	4.54^	5.30	−0.23	−0.23	−0.21	−0.19	−0.09	−0.08	0.50
Item 10: Meals/snacks	3.79	4.28	3.96#	4.63	4.84	5.21	5.68	−0.48	−0.17	−0.10	−0.06	0.00	0.14	0.51
Item 11: Nap/rest	3.43	4.04	3.98^	4.81	4.91^	5.29	5.90	−0.38	−0.24	−0.15	0.08	**0.05**	0.13	0.47
Item 12: Toileting	3.73	4.33	4.10^	4.65	4.66^	5.26	5.53	−0.49	−0.16	−0.09	−0.05	−0.02	0.16	0.47
Item 13: Health	3.35	4.10	3.94^	4.54	4.60^	5.05	5.58	−0.32	**−0.39**	−0.11	−0.07	−0.04	0.08	0.54
Item 14: Safety	3.84	4.38	3.87#	4.44	4.82	5.05	5.59	−0.42	−0.16	−0.14	−0.12	0.00	0.06	0.55
Language−Reasoning														
Item 15: Books	3.20	3.76	3.69^	4.56	5.10	5.38	5.82	−0.29	−0.20	**−0.26**	−0.22	0.05	0.14	0.52
Item 16: Child communication	2.52	3.09	3.43	4.16	4.46	5.06	5.55	−0.30	−0.23	**−0.29**	−0.28	−0.09	0.13	0.45
Item 17: Language reasoning	3.19	3.90	4.30	4.66	4.99	5.23	5.81	−0.37	−0.24	−0.22	−0.08	0.04	0.12	0.51
Item 18: Informal use of language	2.75	3.10	3.61	4.35	4.75	5.10	5.53	−0.33	−0.23	**−0.31**	−0.25	−0.02	0.10	0.51
Activities														
Item 19: Fine motor	2.74	3.39	3.85	4.39	4.58	5.14	5.67	−0.34	−0.27	−0.25	−0.24	−0.05	0.12	0.54
Item 20: Art	2.96	3.74	4.09	4.80	5.29	5.48	5.94	−0.36	−0.33	−0.29	−0.02	0.11	0.22	0.48
Item 21: Music	3.35	4.01	4.43	4.84	5.32	5.68	6.13	−0.23	**−0.37**	−0.16	0.01	0.14	0.26	0.39
Item 22: Blocks	3.30	3.74	3.69^	4.38	4.97	5.48	5.95	−0.35	−0.21	−0.20	**−0.27**	0.04	0.37	**0.35**
Item 23: Sand/water	3.74	4.14	4.29^	4.75	5.19	5.43	5.82	−0.39	−0.12	**−0.19**	−0.04	0.10	0.24	0.36
Item 24: Dramatic play	3.33	3.69	4.02	4.70	5.24	5.72	6.26	−0.28	**−0.36**	−0.21	−0.09	0.15	0.36	**0.33**
Item 25: Nature/science	3.54	4.34	4.56	4.97	5.67	5.74^	6.08	−0.43	−0.22	−0.09	0.09	0.14	0.20	0.43
Item 26: Math	3.03	3.70	4.02	4.72	5.40	5.43^	6.04	−0.41	−0.17	**−0.27**	−0.08	0.16	0.17	0.45
Item 27: Multimedia use	3.53	4.25	4.14^	4.84	5.17	5.46	5.93	−0.39	−0.24	−0.14	−0.01	0.09	0.22	0.39
Item 28: Diversity acceptance	3.78	4.06	4.41	4.74	5.18	5.46	5.98	−0.24	−0.20	−0.19	−0.05	0.11	0.16	0.40
Interaction														
Item 29: Gross motor supervision	3.46	3.61^	3.81^	4.48	4.88	5.33	5.84	−0.35	−0.26	−0.22	−0.16	0.02	0.20	0.45
Item 30: General supervision	3.11	3.53	3.83	4.27	4.58	4.97	5.57	−0.39	−0.29	−0.17	**−0.19**	−0.08	0.06	0.54
Item 31: Discipline	2.94	3.45	3.68^	4.12	4.70	5.01	5.68	−0.37	−0.26	−0.23	**−0.24**	−0.05	0.09	0.53
Item 32: Staff−child interactions	3.09	3.56	3.69^	4.04	4.25^	4.57	5.27	−0.37	−0.25	−0.17	**−0.22**	−0.07	−0.07	0.54
Item 33: Child−child interactions	2.73	3.36	3.56^	4.14	4.62	4.90	5.49	−0.36	−0.28	−0.21	**−0.24**	−0.04	0.04	0.49
Program Structure														
Item 34: Schedule	3.15	4.05#	3.46	4.31	4.58	5.13	5.69	−0.27	**−0.33**	−0.20	**−0.23**	−0.04	0.10	0.58
Item 35: Free play	2.84	3.79#	3.46	4.30	4.76	5.07	5.71	−0.38	−0.25	**−0.28**	−0.24	−0.02	0.10	0.57
Item 36: Group time	3.18	3.43^	3.69^	4.18	4.65	4.86	5.48	−0.40	−0.16	**−0.26**	−0.22	−0.04	0.01	0.54

Note. Values are within−category raw score averages and category−total point−biserial correlations. For means, # indicates reversals based on confidence intervals; ^ indicates overlapping confidence intervals. For point−biserial correlations, bold with underline means the value is smaller than the preceding correlation. Results are from the analysis of the first or only waves of the eight datasets stacked together, n = 4,048 classrooms.

Figure 1

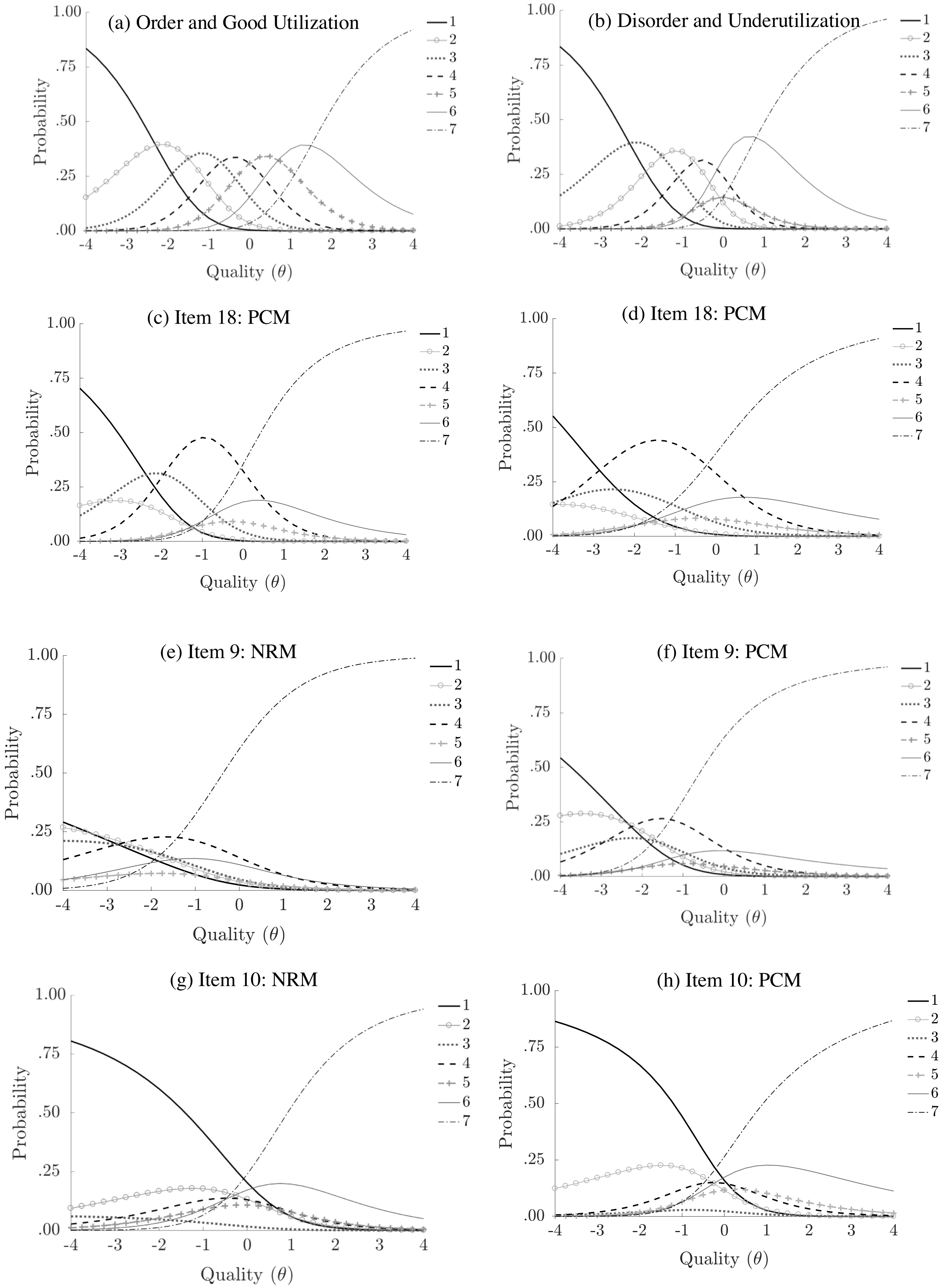
Visual Representation of ECERS−R Stop−scoring Guidelines



Note. Indicators of Category 1 are negatively oriented. Indicators of Categories 3, 5, and 7 are positively oriented.

Figure 2

Category Probability Curves for Example and Select ECERS Items



Note. NRM indicates the nominal response model. PCM indicates the partial credit model.