University of Illinois System

INSTITUTE *of* GOVERNMENT & PUBLIC AFFAIRS

Early Investments Policy Initiative

March 15, 2018

Rethinking high-stakes use of observational measures of preschool quality

By Rachel A. Gordon

In the past decade, the field of early childhood has seen increased high stakes use[1] of observational measures of child care and preschool quality. That is, scoring above or below a particular cut-score on these measures now has substantial financial and reputational consequences for child care centers and preschools. The laws and regulations that led to such high-stakes use called for "reliable and valid" measures.[2] However, interpretations of "reliable and valid" vary, often differing from the latest academic and professional standards for measurement.[3] Particularly important are the incentives for scale developers—and their marketing companies—to attach a static and blanket "reliable and valid" designation to a measure. Doing so is in contrast to the latest standards, which instead call for a continuous accumulation of evidence regarding multiple aspects of reliability and validity and for each potential use of an instrument to carefully weigh that full body of evidence.[4]

In this brief, we provide an example of how digging deeper into the validity evidence for one widely-used observational measure of early childhood education quality—the Early Childhood Environment Rating Scale, Revised (ECERS-R)[5]—reveals problems that have important implications for its high-stakes use. Specifically, the complex scoring of the ECERS-R results in higher scores not always reflecting higher quality. Although the scoring structure may make sense for certain uses—for instance, practitioners may resonate with the items' organization around major preschool routines and activities—it makes less sense for other uses, such as pursuing the policy goals of assuring specific aspects of quality that support school readiness.

Our analysis has important implications for practice and policy. First, we recommend that policymakers and practitioners move away from viewing a measure's reliability and validity as uniform and static. Rather, regular local validation will help increase independent evidence of measurement properties including how measures operate across local sub-contexts. Second, it's easy to lose track of all quality instruments being imperfect realizations of the true level of quality that exists in a program. The rapid movement toward high-stakes use of observational measures of quality has shined a light on certain holes in their validity evidence. Yet these imperfections do not negate the possibility that better measures could be developed to help assess the relationship

About the Author

Rachel A. Gordon is a Professor in the Department of Sociology at the University of Illinois at Chicago and professor at the University of Illinois Institute of Government and Public Affairs. Gordon's research broadly examines contextual, social, and policy factors that affect children and families. She has studied how child care and preschool quality affect child development, the measurement of children's social and emotional competencies, the relationships between youth gang participation and delinquency, the association between community context and child well-being, the causes and consequences of grandmother co-residential support for young mothers, and an evaluation of an innovative job program for young couples. Gordon is the author of two textbooks (Regression Analysis for the Social Sciences and Applied Statistics for the Social and Health Sciences) and has published her research in leading academic journals including the American Journal of Evaluation, Child Development, Criminology, Demography, Developmental Psychology, Early Childhood Research Quarterly, Journal of Marriage and Family, and Journal of Research on Adolescence.

[1] Ackerman, D. J. (2014). State-funded pre-k policies on external classroom observations: Issues and status. Princeton, NJ: Educational Testing Service.
Child Trends. (2017). The QRIS Compendium. Retrieved from http://qriscompendium.org/about/

[2] U.S. Department of Education. (2013b). Applications for new awards; Race to the Top-early learning challenge. Retrieved from https://www.federalregister.gov/articles/2013/08/30/2013-21139/applications-for-new-awards-race-to-the-top-early-learning-challenge#h-6

[3] Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

[4] Ibid.

[5] Harms, T., Clifford, R. M., & Cryer, D. (1998). Early Childhood Environment Rating Scale, Revised Edition. New York: Teachers College Press. Available from http://www.ersi.info/ecers.html

between what happens within the walls of preschool classrooms and children's progress toward school readiness.

Background on ECERS-R and Its High-Stakes Use

The ECERS-R was not designed specifically for its current high-stakes uses. Rather, the instrument was "based on a checklist of items for improving the quality of environments in early childhood classrooms that Harms (one of the instrument's creators) had compiled during nearly 20 years of teaching and observation."[6]

The ECERS[7] measure was first published in 1980, following the rapid growth of maternal employment and child care in the 1970s. The first revision—the ECERS-R[8]—was released in 1998 and was widely adopted in research, practice and policy. A third revision was recently released—the ECERS-3[9]—and we discuss below the ongoing efforts to address the validity of the ECERS-3 for high-stakes use.

When originally developed, the ECERS-R reflected the prevailing concept of best practice in early care and education—developmentally appropriate practice—which includes: a predominance of child-initiated activities selected from a wide array of options; a "whole child" approach that integrates physical, emotional, social and cognitive development; and, highly trained teachers who enable development by being responsive to children's age-related and individual needs (Bryant, Clifford, & Peisner, 1991; Copple & Bredekamp, 2009; Cryer, 1999; Harms et al., 1998).[10] In an interview reflecting on the scale, Harms encapsulated the "whole child" perspective as follows: "in order to provide care and education that will permit children to experience a high quality of life while helping them develop their abilities, a program must provide for the three basic needs of children: a) protection of their health and safety, b) building positive relationships, and c) opportunities for stimulation and learning from experience … It takes all three to create quality care. No one component is more or less important than the others, nor can one substitute for another."[11]

The organization of the ECERS-R items reflects this "whole-child" perspective, as well as the practitioner-focused origins of the scale. Many items encompass the ways in which child care center directors and teachers structure the care setting, including different areas of the classroom (indoor space, gross motor space, space for privacy), events of the day (meals/snacks, greeting/departing, nap/rest), activities (blocks, music, art), and time use (schedule, free play, group time). The scale developers note that this structure makes it easy for observers to "collect information that is likely to be found under similar circumstances."[12]

These origins differ from current high-stakes use. That is, the ECERS-R was not developed specifically to assure that child-care settings funded by state and federal government are of sufficiently high quality in order to narrow gaps in children's school readiness. Rather, the scale was adopted for high-stakes use because it was one of the most well-known measures at the time this policy interest emerged. These high-stakes uses are quite consequential. For instance, a common strategy has been to write into policy particular measures of the quality of early childhood classrooms, like the ECERS-R, and to penalize or reward programs with certain average scores on these measures. By 2017, approximately three-quarters of the 41 states that relied upon an observational classroom assessment in their Quality Rating and Improvement Systems (Child Trends, 2017), and in 2012-2013, 19 states used the ECERS-R for monitoring their state pre-kindergarten programs.[13] Depending on the state, these ratings can influence public perceptions of centers' and schools' performance—through publicized star or medal rankings, similar to restaurant or movie reviews—and can determine the amount of money the state provides to subsidize children's care. These high stakes warrant close scrutiny of the accumulated evidence regarding the ECERS-R. We now turn to one important piece of evidence: the validity of its standard scoring structure.

Description of the ECERS-R Scoring Structure

Understanding the ECERS-R standard scoring procedure is critical to grasping one of its key limitations for high stakes use.

Each ECERS-R item is scored on a 1-to-7 scale with odd-value labels from 1 = Inadequate quality to 3 = Minimal quality to

[6]Frank Porter Graham Child Development Institute. (2003). A whole new yardstick. Early Developments, Vol. 7 [Rating early childhood environments], 8–11. Available from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/early-developments/FPG_EarlyDevelopments_v7n2.pdf

[7]Harms, T., Clifford, R. M., & Cryer, D. (1980). Early Childhood Environment Rating Scale. New York: Teachers College Press. See also http://www.ersi.info/scales_history.html

[8]Harms, T., Clifford, R. M., & Cryer, D. (1998). Early Childhood Environment Rating Scale, Revised Edition. New York: Teachers College Press. Available from http://www.ersi.info/ecers.html

[9]Harms, T., Clifford, R. M., & Cryer, D. (2015). Early Childhood Environment Rating Scale, Third Edition. New York: Teachers College Press. Available from http://www.ersi.info/ecers3.html

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

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

Cryer, D. (1999). Defining and assessing early childhood program quality. Annals of the American Academy of Political and Social Science, 563, 39–55. doi:10.1177/0002716299563001003

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

[11]Frank Porter Graham Child Development Institute. (1999). Building blocks. Early Developments, Vol. 3 [From process to product in early childhood development], 11-12. Available from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/early-developments/FPG_EarlyDevelopments_v3n3.pdf

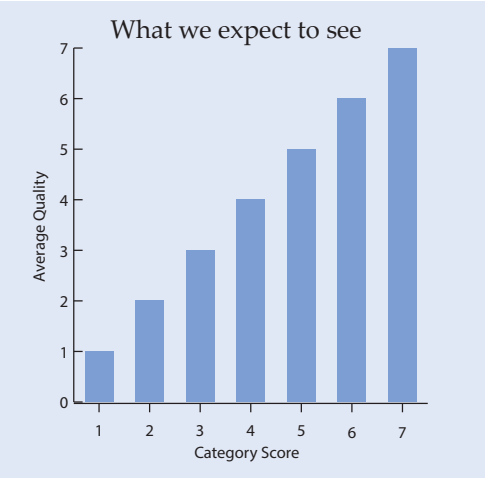
[12]Cryer, D., Harms, T., & Riley, C. (2003). All about the ECERS–R. Lewisville, NC: Kaplan.

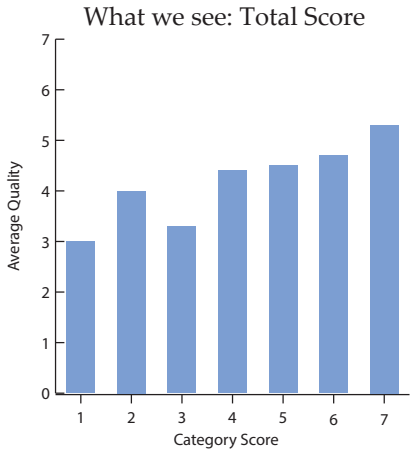
[13]Ackerman, D. J. (2014). State-funded pre-k policies on external classroom observations: Issues and status. Princeton, NJ: Educational Testing Service.

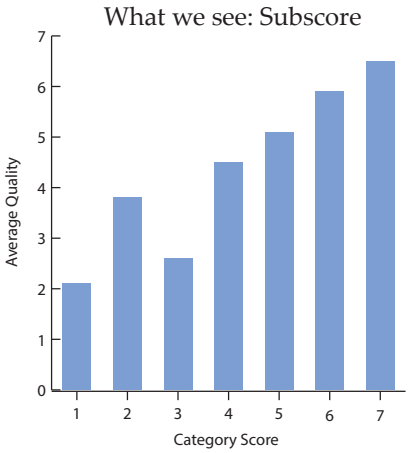
Child Trends. (2014). The QRIS Compendium. Retrieved from http://qriscompendium.org/about/

Figure 1: Examples of What we expect to see vs. What we see in the ECERS-R and Rasch Measures[1]

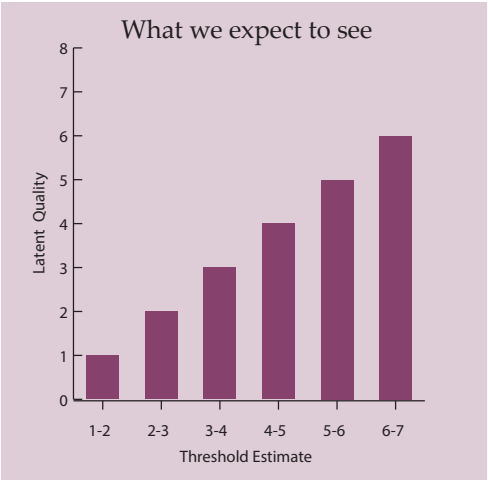
ECERS-R Raw Scores

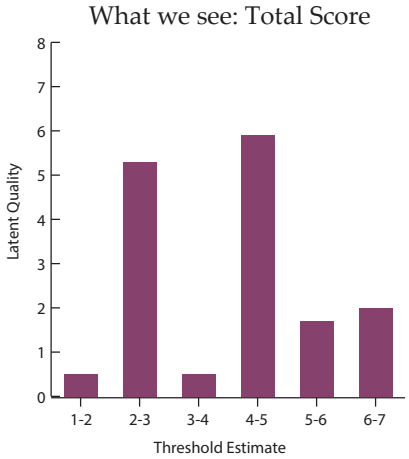






ECERS-R Rasch Measures





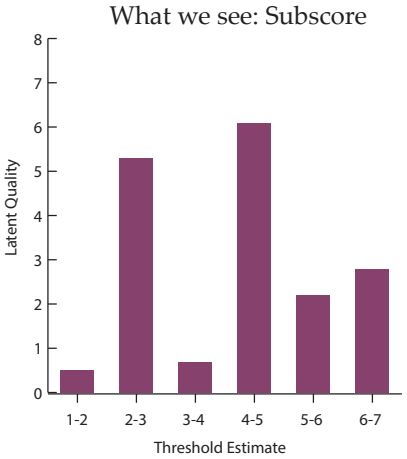
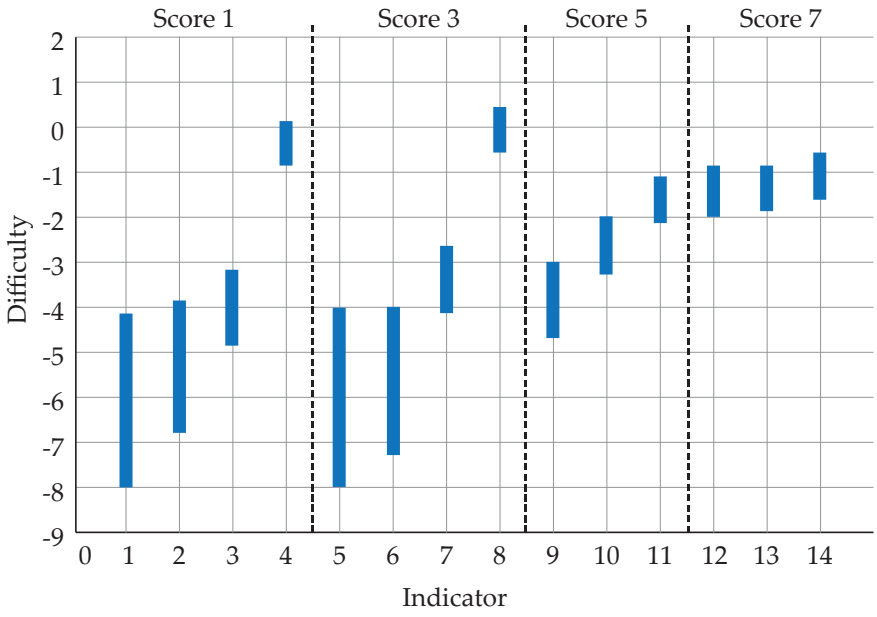


Figure 2: ECERS-R 10: Meals/Snacks[2]



Indicator Key

1. Acceptable nutritional value
2. Appropriate meal schedule
3. Positive atmosphere
4. Sanitary condition
5. Well-balanced meals
6. Schedule appropriate
7. Nonpunitive atmosphere
8. Sanitary conditions usually maintained
9. Eat independently
10. Pleasant atmosphere
11. Staff sits with children
12. Conversation
13. Child-sized utensils
14. Children help

Note: Indicator labels for "Score 1" reflect their reverse scoring for analysis. Indicators that permitted "not applicable" scores were omitted due to high levels of missing data.

[1]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. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3681422/

[2]Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N. C., Kaestner, R., & Korenman, S. (2015). New evidence about the validity of the ECERS-R for evaluations of preschool programs aimed at improving school readiness. Early Education and Development, 26, 1086-1110. Available at http://www.tandfonline.com/doi/full/10.1080/10409289.2015.1036348

5 = Good quality to 7 = Excellent quality. Observers look for several indicators that are listed under each odd-numbered category for each item (for instance, "most furniture is sturdy and in good repair," "most staff-child interactions are pleasant and helpful," and "books are organized in a reading center"). In the standard scoring, indicators for lower scores must be met before indicators of higher scores are evaluated. That is, observers stop scoring when they reach a response category at which an indicator is not observed.

On the one hand, this standard scoring approach might be seen as beneficial. It reduces response burden, in that observers do not have to consider indicators above the stop point. It may also reflect a philosophical perspective that centers should not get credit for higher-level aspects of quality that they are doing well (e.g., being warm and responsive in their interactions with children) if they are not doing lower-level aspects of quality well (e.g., assuring basic cleanliness and safety). This is consistent with taking a "whole child" perspective and measuring global quality.[14]

On the other hand, there are ways in which the approach might not be beneficial. For instance, mixing together indicators of different aspects of quality might limit the alignment of ECERS-R scores with particular domains of children's development (e.g., early reading, math, or social skills), therefore limiting their validity for high-stakes policy uses focused on supporting specific aspects of school readiness. It's also the case that, until recently, validity evidence for the developers' placement of indicators was lacking, even as a measure of global quality. That is, the stop-scoring approach was based on the scale developers' experiences in classrooms and understanding of the literature[15] rather than empirical evidence showing that indicators placed at lower category levels (e.g., 1 and 3) actually reflected lower levels of the underlying dimension of quality than did indicators placed at higher category levels (e.g., 5 and 7). We describe next the possible implications of indicators' true placement differing from the developers' expectations, as well as how we filled this gap in the validity evidence related to the ECERS-R.

Potential Implications of the ECERS-R Scoring Structure

As noted above, being sure ECERS-R scores offer an accurate reflection of child care classroom quality is important given the very high-stakes ways in which being above or below a particular score has become an important factor in determining recognition (e.g., four star or gold quality status) and funding (e.g., higher reimbursement levels). For instance, in Illinois' Excelerate Quality Rating and Improvement System if a licensed child care center chose the ECERS-R pathway to demonstrate sufficiently high classroom quality for the Gold Circle of Quality, then none of its classrooms could score below a 4 during on-site ECERS-R ratings by a state-approved assessor.[16]

What exactly are policymakers hoping these scores will reflect? Although quality is a multidimensional construct—and policymakers and administrators sometimes articulate multifaceted and broad goals for investments in early care and education—raising children's school readiness is one key policy goal. The goal is typically to assure that public dollars flow to settings with sufficiently high quality in order to go beyond promoting children's health and safety (which is often achieved through basic licensing standards) and to also optimize children's cognitive and socio-emotional development (i.e., to "close the school readiness gap," U.S. Department of Education, 2013; to support "positive child development and later achievement.")[17] The "Pre-K Now" initiative sponsored by the Pew Charitable Trusts similarly focused on advancing "high-quality, voluntary pre-kindergarten for all three-and four-year-olds" by pointing to evidence that "high-quality pre-k is an essential catalyst for raising school performance."[18] Quality Rating and Improvement Systems (QRIS) for child care emerged to encourage quality across all types of care settings in the late 1990s, and had spread to three-quarters of the states by 2014.[19] An umbrella organization—the QRIS National Learning Network (2015)—stated that their aim was likewise to "elevate the quality of care in state early care and education systems and to support and improve children's development."[20]

Given these policy goals, evidence is needed regarding the possibility that the standard ECERS-R scoring procedure weakens its scores' signal of quality, especially for the aspects of quality that most strongly support children's academic school readiness. In fact, close scrutiny of the instrument's organization around events of the day reveals the ways in which items mix different aspects of quality, likely relevant for multiple domains of children's school readiness. For instance, Item 10 "Meals/snacks" contains not only indicators of nutrition and sanitation but also indicators of the amount of conversation that takes place during meals and the tone of staff-child interaction. Under the standard

[14]Clifford, R. M., Reszka, S. S., & Rossbach, H. (2010). Reliability and validity of the Early Childhood Environment Rating Scale. Retrieved from http://ers.fpg.unc.edu/sites/ers.fpg.unc.edu/files/ReliabilityEcers.pdf.

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

[15]Clifford, R. M., Reszka, S. S., & Rossbach, H. (2010). Reliability and validity of the Early Childhood Environment Rating Scale. Retrieved from http://ers.fpg.unc.edu/sites/ers.fpg.unc.edu/files/ReliabilityEcers.pdf.

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

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

[16]Excelerate Illinois. Gold Circle of Quality. Oct. 2, 2017. Retrieved from http://www.excelerateillinoisproviders.com/docman/resources/2-gold-excelerate-illinois-chart/file

[17] Improving Head Start for School Readiness Act of 2007. Public Law 110-134). 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. Author: Press Release (April 16, 2013).

[18]Pew Charitable Trusts. (2014). Pre-K Now. Retrieved from http://www.pewtrusts.org/en/archivedprojects/pre-k-now

[19]Child Trends. (2014). The QRIS Compendium. Retrieved from http://qriscompendium.org/about/

[20]Quality Rating and Improvement Systems. Visit their website at http://qrisnetwork.org/

stop scoring, aspects of quality that might be most strongly associated with children's cognitive or social development cannot be separated from those that may be most strongly associated with children's health and safety.

The ways this mixing of different aspects of quality could combine with the stop scoring to mute the ECERS-R signal of quality can be illustrated by analogy. Imagine a measure of child development whose items mixed together indicators of whether a child was consistent in her personal hygiene, friendly in her peer interactions, and extensive in her spoken vocabulary. If rating her higher in each of these three domains depended upon her ratings in the other areas, as in stop scoring, then the scale scores would reflect some mixture of the different aspects of herself, rather than any pure indication of each or their composite. This issue would be especially salient to the extent that the three aspects of development—health, social, academic—are distinct rather than fully overlapping. That is, under stop scoring, some of the children in a lowest item category might be low in all three areas whereas others might be low in one but high in one or both of the others. Thus, the stopped scores indicate some complex interplay between the content and number of the indicators.

Empirical Evidence Regarding the ECERS-R Scoring Structure

We started our empirical work on the ECERS-R scoring structure by considering the basic question of whether higher overall quality was associated with higher item categories. An intuitive way to do this was to see whether a classroom's average scores across all items were higher in each successive category for each item. The top/left graph in Figure 1 illustrates what we expected to see. In our research with a nationally representative sample of preschool-aged children, we found instead that average quality was unexpectedly lower in some higher categories.[21] The top/middle section of Figure 1 shows an example of what we actually saw for one of the ECERS-R items, where there is an unexpected dip at the third category. In other words, classrooms that were scored a "2" on this item averaged higher overall quality than did classrooms scored a "3."

Although the ECERS-R scale developers often describe the ECERS-R as a measure of global quality—and its total score is frequently used—the developers also provide a way to average subsets of items in subscales. We found the lack of steady progression of quality even when we used the subscale associated with an item, rather than the total score. The top/right graph in Figure 1 shows an example, with the dip at the third category being even more pronounced for the item's subscale score than for the total score.

This issue of category order can be more formally tested with psychometric models, and we verified the problem with these more rigorous tests. One set of psychometric models (partial credit model) estimate thresholds that reflect the point on the underlying dimension of quality at which a classroom would have a 50:50 chance of being scored in the higher or each pair of adjacent categories. We found at least one set of these thresholds was out of order for every ECERS-R item. The bottom panel of Figure 1 provides an example of what we expected to see (on the left) and what we actually saw (in the middle and right) for the same item considered above.

In recent work, we replicated and extended these results using eight different large-scale datasets representing a diverse array of types of care (Head Start, state pre-kindergarten, and community-based child care), research teams (smaller investigator-collected data and large survey firms) and demographics (including lower income samples that are often the target of public funding).[22] We found at least one instance of threshold disorder for the majority of items in nearly every dataset. When we pooled the datasets together in order to offer more precise estimation, we were able to pinpoint the most common locations of threshold disorder, finding one common instance around category 3 (as in the example shown above) and another common instance around category 5. This may reflect an added nuance to the ECERS-R stop scoring—the different ways in which indicators are assessed for scoring even versus odd items (odd scores require all of their associated indicators to be met; even scores require half or more of the indicators of the next odd-numbered score to be met).

Our most recent work also used another type of psychometric model (the nominal response model) which helped us differentiate between whether a category is fully out of order—as described earlier—versus being underused or redundant. Each issue suggests measurement error but in different ways, and the partial credit model featured in earlier studies cannot distinguish among these reasons. By stacking the eight data sets together, we had the sample size needed to estimate the more complicated kind of model that can distinguish these issues. We found all three issues were evident in the ECERS-R items. Every item had some underused and redundant categories, and one-fifth of items had categories that were fully out of order.

What does this mean? Underuse could happen simply because a sample is unusual (i.e., the sample happened to omit classrooms at a certain scale level). In representative samples like ours, such chance omissions should be less of an issue. Instead, underuse may reflect inefficiency in scale construction. In the ECERS-R, some indicators are mirror images of each other, especially between the 1st and 3rd categories, which may lead the higher category to be skipped over (e.g., "greeting is often neglected" at the 1st category and "most children greeted warmly" at the 3rd). Redundancy might also happen in this case, especially given slight wording differences ("warmly"). Redundancy can also happen when similar indicators are placed at two adjacent categories (e.g., "pleasant social atmosphere" and "meals and snacks are time for conversation").

[21]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. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3681422/

[22]Fujimoto, K. A., Gordon R. A., Peng, F., & Hofer, K. G. (2018). Examining the category functioning of the ECERS-R across eight data sets. AERA Open, 4(1), 1-16.

We further probed these results by taking advantage of a set of studies that had observers rate the ECERS-R in a slightly different way.[23] Rather than stop-scoring, the observers rated every indicator associated with every category of every item. We used this information to help us test our hypothesis that disorder might reflect the mixing of different aspects of quality combined with the stop-scoring approach. In other words, if the indicators associated with each category of an item reflect progressively higher levels of a single aspect of quality, then stop scoring would not be a problem. However, if indicators do not progress in this way (sometimes indicators of higher levels of quality are placed at lower categories) stop scoring may produce category disorder. This lack of progression may be especially likely when an item's indicators mix different aspects of quality.

We indeed saw this result, particularly for the items in the ECERS-R "Personal Care Routines" subscale. Figure 2 illustrates the result for the ECERS-R Meals and Snacks item, where sanitary conditions were estimated to be substantially more difficult than the remaining indicators (that is, the 4th and 8th indicators are positioned higher in the graph shown in Figure 2). In other words, the developers' placed sanitary conditions at categories 1 and 3 even though they were estimated to be higher on the quality dimension than any other of the items' indicators, including those placed at categories 5 and 7. Said another way, sanitary conditions were a difficult hurdle for many classrooms to overcome in order to be recognized for other activities, like staff and children sitting together and engaging in conversation during the meal.

Importance of the Results

Intuitively, category disorder is problematic. If higher categories of items do not consistently reflect more of the construct being measured then summary scores do not provide a good signal of that construct. In the high-stakes context, these noisy signals are all the more problematic because they are compared with absolute cutoffs to make consequential decisions.

The greater noise in the scores would also make them less likely to associate in predictable ways with other constructs. Evidence is indeed accumulating that ECERS-R scores have small associations with the important outcome often featured in policy—children's school readiness. These small associations have been documented in a number of recent studies, including our own work[24] and work by other research teams.[25]

For instance, in the nationally representative sample used in our first study described above, we found that the ECERS-R total and subscale scores were not significantly associated with children's reading and math outcomes and were significantly associated with just a few of children's social and emotional outcomes.[26] All associations were also small in size (less than one-tenth of a standard deviation, a size considered "small" by social scientists and in relation to school readiness gaps). Although numerous factors may be responsible for these small associations (e.g., some children's limited hours and weeks of exposure to preschool; possible limitations with the reliability and validity of measures of child development), the problem with the ECERS-R standard stop scoring is likely one important culprit.

Takeaway Points

Policymakers' and scale developers' use of quality measures are generally well intentioned. Policymakers want to be sure that public dollars flow to high quality settings, in part so that children flourish in their care. Scale developers want to help teachers understand and grow the quality of their classrooms. Developers also want to help researchers document what conditions support or impede quality as well as how higher quality leads to greater child development.

Yet calls for "reliable and valid" measures in laws and regulations—as well as requirements for similar certifications in journal articles and grant proposals—can lead scale developers and scale users to reduce (ideally large) bodies of evidence to shorter sound bites. Not only can some important limitations of evidence get lost by this reduction, but doing so also encourages scale developers and scale users to treat a measure's reliability and validity as uniform and static. Once a scale gets an endorsement of "reliable and valid" it may be difficult or undesirable to reconsider.

This situation varies from contemporary standards of measurement, which encourage a continuous amassing of evidence related to an instrument's reliability and validity and constant re-assessment of this evidence for each use. Indeed, the same evidence may lead to different conclusions for different uses. Whereas the issue of category disorder that we highlighted above is problematic for many research and policy uses, it may be less problematic if the goal is an efficient scoring strategy for focusing teachers' attention on characteristics believed to reflect the highest level of global quality. Ultimately, the onus is on scale users to locally validate measures—and for independent external validations to be regularly and transparently shared—

[23]Gordon, R. A., Hofer, K. G., Fujimoto, K. A., Risk, N. C., Kaestner, R., & Korenman, S. (2015). New evidence about the validity of the ECERS-R for evaluations of preschool programs aimed at improving school readiness. Early Education and Development, 26, 1086-1110. Available at http://www.tandfonline.com/doi/full/10.1080/10409289.2015.1036348

[24]Abner, K., Gordon, R.A., Kaestner, R. and Korenman, S. (2013). Does child care quality mediate associations between the type of care and child development. Journal of Marriage and Family, 75(5), 1203-1217. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24068846

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. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3681422/

[25]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.

Layzer, J. I., & Goodson, B. D. (2006). The quality of early care and education settings—Definitional and measurement issues. Evaluation Review, 30(5), 556-576.

[26]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. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3681422/

in order to better accumulate evidence related to how measures operate across local contexts and for various purposes.

For all uses, it is also important to remember that quality instruments are imperfect realizations of underlying constructs. If we view measures as living, breathing tools, in constant need of reflection and refinement, then we can avoid "throwing the measure out with the evidence" when an instrument is revealed to work differently than intended. In many ways, the growing recognition of smaller-than-expected associations between observational quality scores and children's development has spurred such reflection regarding what quality is all about and how it is best measured. The recent release of the ECERS-3 is an important step in this direction. Although it retains the standard scoring and much of the structure of the ECERS-R—and recent results show small associations with children's outcomes based on its standard scoring[27]—alternative scoring structures are in development.[28] In the meantime, we encourage policymakers and practitioners to collaborate with researchers to carefully consider the full body of evidence related to high-stakes use of quality measures and to build and share more local evidence related to such uses. •

Acknowledgments

The research summarized in this brief results from a series of collaborative research projects led by Professor Gordon that examine the psychometric properties of widely-used measures of preschool and child care quality, including projects funded by from the Institute of Education Sciences (IES), U.S. Department of Education, through Grants R305A090065 and R305A130118, as well as by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), through Grant R01HD060711. The footnotes in this brief reference articles that contain more details about these projects and the collaborative teams. The opinions expressed are those of the author and do not necessarily represent views of IES, the U.S. Department of Education, or NICHD.

Contact Rachel A. Gordon at ragordon@uic.edu

The Institute of Government and Public Affairs (IGPA) is a public policy research organization at the University of Illinois. IGPA's mission is to improve public policy and government performance by: producing and distributing cutting-edge research and analysis, engaging the public in dialogue and education, and providing practical assistance in decision making to government and policymakers. The institute's work not only advances knowledge, but also provides real solutions for the state's most difficult challenges.

To learn more, visit igpa.uillinois.edu.

[27]Gordon, R. A., Hofer, K. G., Aloe, A. M., Wilson, S., Peng, F., Gaur, D., & Lambouths, D. III. (2018). Early childhood classroom quality and preschool-aged children's vocabulary: A meta-analytic study of heterogeneity and moderation of effects.

[28]Institute of Education Sciences. Funded Research Grants and Contracts. Large-Scale Psychometric Assessment of the ECERS-3. Retrieved from http://ies.ed.gov/funding/grantsearch/details.asp?ID=1715