

Best Practices in Creating and Adapting Quality Rating and Improvement System (QRIS) Rating Scales



Best Practices in Creating and Adapting Quality Rating and Improvement System (QRIS) Rating Scales

OPRE Research Brief #2016-25

May 2016

Submitted by: Margaret Burchinal, University of North Carolina
Louisa Tarullo, Mathematica Policy Research
Martha Zaslow, Society for Research in Child Development and Child Trends

Submitted to: Ivelisse Martinez-Beck, PhD., Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract number: HHSP23320095631WC

Project director: Kathryn Tout
Child Trends
7315 Wisconsin Avenue
Suite 1200W
Bethesda, MD 20814

This report is in the public domain. Permission to reproduce is not necessary.

Suggested citation: Burchinal, M., Tarullo, L. & Zaslow, M. (2016). *Best practices in creating and adapting Quality Rating and Improvement System (QRIS) rating scales*. OPRE Research Brief #2016-25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This document was prepared to accompany other resources on evaluation of Quality Rating and Improvement Systems (QRIS) and other quality improvement initiatives developed by the Quality Initiatives Research and Evaluation Consortium (INQUIRE).

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at <http://www.acf.hhs.gov/programs/opre/index.html>.



Overview

Ratings produced in Quality Rating and Improvement Systems (QRIS) are intended to provide meaningful information to parents and policymakers about the quality of early care and education programs. Given this central purpose of ratings, it is important to establish and follow guidelines for creating QRIS ratings that can promote their validity and integrity. The purpose of this brief is to demonstrate how using the principles of scale development can support the development of QRIS ratings.

The brief summarizes an analysis that uses the data from six large studies of early care and education to simulate state QRIS ratings. The results suggest that QRIS ratings can achieve their desired goal of predicting gains in child outcomes when attention is paid to the psychometric principles of scale development including: dimensionality (ensuring that a scale represents one, not multiple dimensions), selecting items with strong evidence, and scoring items using established criteria for cut points. The analysis provided significant validation —albeit modest, in terms of the strength of associations—of almost all of the carefully selected quality measures of classroom experiences. This finding was observed when measures were analyzed as individual quality indicators and when they were combined into a summary QRIS rating of classroom experiences. It also held true whether focusing only on structural indicators, or including both structural and process indicators of quality.

The analysis provides an example but not a blueprint for developing QRIS ratings. The authors conclude that ratings can be strengthened by using a careful approach to construction of the ratings that takes into account the content and evidence base for selected quality indicators.



Best Practices in Creating and Adapting Quality Rating and Improvement System (QRIS) Rating Scales

Quality Rating and Improvement Systems (QRIS) are state or local policy initiatives designed to increase availability of and access to high quality early care and education (ECE). QRIS address two concerns: (1) consumer education, by publishing the summary rating of quality of enrolled programs so parents can make informed child care choices, and (2) support of continuous ECE program quality improvements, and ultimately children's developmental outcomes, by providing quality improvement resources to those programs (Tout, Starr, Soli et al, 2010). Each state or locality develops its rating scale from selected indicators of ECE quality with the goal of producing a summary rating that concisely describes the level of quality of an ECE program (Zellman & Perlman, 2008).

The purpose of this brief is to illustrate how attending to principles of scale development can be useful in the development and refinement of QRIS rating scales. It uses the data from six large studies of early care and education to simulate QRIS information, thereby demonstrating analytic approaches that can be carried out with actual state data.

QRIS rating scales are typically developed by ECE stakeholders within a state or community, using research evidence and professional expertise to identify the quality indicators to be included. While developers consider the *content* and the *structure* of the rating scale (for example, whether they will use a block or point system for scoring the indicators), less attention is paid to principles of scale development. This step is important because the QRIS rating is, in fact, a scale. It combines measures of different aspects of quality into a single score and sets increasing levels of quality to create more points or levels. As such, adherence to psychometric principles of scale development is necessary to ensure that the scale measures effectively what it is designed to measure (Lambert et al., 2006).

Attempts to validate QRIS rating scales thus far have provided little evidence that children's early learning and development were enhanced if they attended ECE programs with higher ratings (Hestenes et al., 2014; Sabol, Soliday Hong, Pianta, & Burchinal, 2013; Sabol & Pianta, 2014; Soliday Hong et al., 2014; Thornburg, Mayfield, Hawks, & Fuger, 2009; Zellman, Perlman, Le, & Setodji, 2008). This limited prediction of child outcomes from QRIS rating scales may be, in part, due to lack of attention to principles of scale development when the QRIS ratings were developed. Attention to these principles will enhance the ability of QRIS ratings to achieve the goal of predicting to child outcomes.

This brief will illustrate the importance of using psychometric principles in QRIS rating scale development, by (a) articulating the specific principles that are of importance in this process, and (b) providing an example that involves using these principles to create and test a hypothetical QRIS rating score. It is important to note that the example is illustrative and includes many—but not all—possible QRIS indicators that may be available at the state level. Accordingly, the purpose is to illustrate how the principles can be put to use, not to suggest that this example should be used as a template.

Further details about the analyses described in this brief are available in a longer technical report (Burchinal et al., 2016).

Principles

Scale development that takes into account psychometric issues points to three principles that are especially relevant for QRIS rating scales: dimensionality, item selection, and item scoring. A definition of each of these principles is provided below, along with a brief discussion of its relevance to development or revision of QRIS rating scales.

Dimensionality: ECE quality has many different aspects, such as the learning environment, qualifications of staff, program administration and management, and family engagement. While related to each other, these different aspects are thought to represent separate *dimensions* of program quality, including: (1) children's experiences in the setting, including teacher-child interaction, health and safety provisions, materials and activities, group size, child to adult ratio, teacher qualifications, use of a curriculum, and use of child assessments; (2) family engagement practices to support the full range of families, including family partnerships, provisions for cultural and linguistic diversity, and regular communication with families; (3) program administration practices, including workplace policies and procedures and supports for staff; and (4) engagement with the ECE system through licensing compliance, alignment with early learning standards, or development of a quality improvement plan that incorporates ECE system resources.

QRIS ratings typically include indicators of these different aspects of quality. The multiple indicators are measured, scored, and combined into a rating using various methods that reflect the priorities of the developers.¹ A principle of scale development is that a scale should measure one construct, and thus represent a single dimension; combining multiple dimensions in a single scale score reduces the ability of that scale to predict the outcomes associated with any of the different dimensions. Multiple scale scores are needed when multiple dimensions are being measured. Thus, it is important to test whether more than one dimension of quality is encompassed by a QRIS. It is advisable to have more than one rating scale if there are multiple dimensions of ECE quality.

Item Selection: Items within the QRIS rating are the individual measures of quality that are scored in specific ways to become the quality indicators that are combined to form the overall rating. A principle

¹ In a "block" QRIS structure, all indicators identified for a quality level must be achieved before indicators at the next quality level can be scored. In a "point" structure, points are awarded for quality indicators achieved. The points are summed and a quality level is assigned based upon a selected cut-point for the summed scores. For further information and examples of how QRIS structures produce different rating distributions, see Tout, Chien, Rothenberg & Li, 2014.

of scale development is that item selection involves finding the individual indicators with strong evidence that they measure important aspects of the construct of interest and that they measure them well. Items in a QRIS should be selected based both on the evidence of their importance (as established in the research literature or in statements of best practice developed by experts) and on the evidence that we can measure those indicators successfully. Both types of evidence are needed because poor measurement of important items will dilute the ability of the QRIS quality rating to predict child outcomes.

Item Scoring: Another principle of scale development is that item scoring should also be based on the strength of the evidence. Within the context of a QRIS, item scoring involves turning the quality measure into a quality indicator by deciding how the information from the measure will be summarized and included in the summary rating scale. For example, a QRIS may create an indicator of the overall quality of the classroom environment by assigning 3 points for a total score of 4.5 and above on the Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Cryer, & Clifford, 1998), 2 points for a total score of 4.0 to 4.49, and 1 point for a total score of 3.5 to 3.99. Using cut-points that have been demonstrated to differentiate higher and lower quality programs (in other words, evidence-based item scoring) can strengthen the ability of a QRIS to predict to child outcomes because they create categories that are meaningfully different (for example, low, medium, and high quality according to ECERS-R). In contrast, use of arbitrary cut points (for example, those that have not been tested statistically) can transform important and meaningful information from measures into scores that cannot discriminate between higher and lower quality programs.

These principles raise questions about issues that could reduce the ability of the QRIS rating to predict desired outcomes. These questions include:

1. Should there be one or multiple summary ratings? Is there really one underlying dimension of ECE quality included in the QRIS, or is there more than one?
2. Is there good evidence that the selected quality indicators are related to an ECE quality dimension? Do QRIS ratings include indicators that might weaken the way the QRIS rating is functioning because they are not well measured or not based on strong research?
3. Has a cutoff for the highest score on each indicator been chosen in such a way that the highest score really indicates higher quality based on the research?

Example

To illustrate the potential importance to QRIS of attending to principles of scale development, we conducted an analysis taking the principles into account. The analysis used existing data on ECE programs to create a QRIS rating scale; the rating scale was not developed using actual QRIS data. The results indicate that a simulated QRIS rating scale developed using psychometric principles predicts gains in child outcomes.

This research brief contains a summary of the analysis and results. Detailed methods and tables can be found in the full research report upon which this brief is based.

Studies that were included: Studies were included in this analysis if they provided data on the quality of a large number of centers serving 3- and 4-year-old children (~100 or more), if the data included measures of both structural and process quality that are widely used in QRIS ratings, and if the data included measures of child outcomes for preschool-age children based on widely used assessments of early academic and social skills. These studies included federal and state-funded ECE programs, as well as community-based ECE provided in center settings. The sample included two studies of Head Start: the Head Start Family

and Child Experiences Survey (FACES) 2006 (West et al., 2010) and 2009 (Malone et al., 2013); two evaluations of state pre-kindergarten programs: the North Carolina Pre-Kindergarten Evaluation (NC Pre-K; Peisner-Feinberg, 2013) and the Georgia Pre-Kindergarten Evaluation (GA Pre-K; Peisner-Feinberg et al., 2014); and two studies of classrooms in center settings from different auspices: the preschool observational sample from the Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B; U.S. Department of Education, 2007) and the National Center for Research in Early Care and Education (NCRECE; Hamre et al., 2012) professional development study.

Quality measures: Quality measures were selected for the analysis if: (1) they were quality indicators typically included in QRIS ratings scales (Tout et al., 2010; Build Initiative & Child Trends, 2015); (2) there was replicated evidence showing that the quality measure was related to observed classroom quality or child outcomes; and (3) QRIS logic models portrayed the quality indicator as influencing child outcomes. Seven quality measures met all three criteria and were included in the analysis: process quality measures of the ECE environment (ECERS-R; Harms, Cryer, & Clifford, 1998) and of teacher-child interactions (Classroom Assessment Scoring System-CLASS; Pianta, LaParo, & Hamre, 2008) and structural quality measures of teacher and director education, child-teacher ratios and group size, and curricula (use and type). Multiple studies in the research literature on ECE indicate that each of the process quality measures included in these analyses is at least a modest predictor of child outcomes. Similarly, multiple studies indicate that each of the selected structural quality measures is a moderate to strong predictor of observed classroom quality.

Quality Measures Included in the Simulated QRIS Rating Scale

Process Quality

ECERS-R

CLASS

Structural Quality

Teacher education

Director education

Child-teacher ratio

Group size

Curriculum

Measures of other aspects of ECE quality were available in some, but not all, studies included in the analysis. These were measures of family engagement, inclusion of children with special needs, and inclusion of the home language and culture for dual language learners. The research evidence linking these with observed quality or child outcomes is more mixed: studies vary in their approaches to measuring these aspects of quality and show less consistency in examinations of associations with observed quality or child outcomes.

Dimensionality: A factor analysis² was carried out to examine whether the selected measures of classroom quality (both measures of process and structural quality) and measures of other aspects of ECE quality (family engagement, inclusion, and diversity) either contribute to a single dimension or reflect multiple dimensions.

Results revealed that the indicators of classroom quality and the indicators of other aspects of ECE quality including family involvement, provisions for dual language learners, and inclusion of children with special needs, loaded on separate factors. This result suggests that these aspects of quality constitute separate dimensions of quality. In addition, some evidence suggested that process quality might define a separate dimension from structural quality. While not completely consistent across the analyses of the data from the different studies, there was sufficient evidence in our analyses that they loaded on a single factor that we chose to combine the quality indicators of classroom experiences into a single scale or dimension. These indicators included teacher education, director education, use of a curriculum, ratio, group size, and observed classroom quality measures.

In summary, the findings suggest that ECE quality is multidimensional, and use of a single scale is unlikely to adequately and precisely represent quality across all of the various dimensions. These findings suggest that other quality scales might be needed to represent measures of the quality of practices to support

² Exploratory principle factor analysis with varimax rotation using an eigenvalue of one and factor loadings of .3 to determine which quality measures loaded on each factor.

cultural diversity, inclusion, family engagement, and perhaps other aspects that were not included in our data, like health and safety practices, and program administration and management. Creating a single rating encompassing multiple dimensions is likely to dilute associations with outcomes that might be seen with individual quality variables or unidimensional scales. We focus in our further analyses on the indicators of classroom experiences (including both structural and process quality indicators), as these appear to contribute to a single dimension.

Item Selection: Each of the selected classroom quality measures was then examined to determine the extent to which these structural and process quality measures predicted gains in child outcomes.³ As noted, item selection is based on consensus that the item is conceptually important (established through research and statements of professional practice) and that it can be measured effectively. The goal here was to provide further examination of the first criterion: that the quality indicators are important because they are related to child outcomes.⁴

We examined the associations between each of the process and structural quality measures and child outcomes. Results, shown in the top half of Table 2, indicated modest and significant associations for CLASS Instructional Support and Classroom Organization with scores on assessments of children's early reading skills; of director and teacher education with scores on assessments of children's early language, math, and reading skills; and of child-to-adult ratios and curricula with teacher or parent reports of children's social skills. ECERS-R total, CLASS Emotional Support, and group size were not significantly related to any the child outcomes.

In summary, these analyses demonstrated that most (but not all) of the selected indicators appear to fulfill the criterion of item selection that they predict to child outcomes.

Item Scoring: Next, we conducted an analysis to understand how two different scoring approaches would affect the association between the quality measures and child outcomes. In the previous analysis, we examined the associations between measures and outcomes using continuous scores. In this analysis, we assigned a score that created three levels—0, 1 or 2—for each of the quality measures using the scoring criteria described in Table 1. The scoring approach examined here approximates a QRIS rating with three possible levels. We used the developers' guidelines for ECERS-R and widely used criteria for defining quality for the CLASS. For the CLASS, we created a single indicator from the three CLASS domain scores as described in Table 1. For teacher and director education, ratio, group size, and curricula, we used the joint professional guidelines from the American Pediatrics Association and American Public Health Association (2011) to determine the top level, and we used licensing criteria from states with rigorous standards to establish the middle level. Classroom-level data were averaged to compute a center-level quality score of 0, 1, or 2 for each quality variable.⁵

Next, we examined the associations between these categorized quality indicators and child outcomes. The results of these analyses, shown in the middle rows of Table 2, were compared to the results of the analyses using the continuous quality variables, shown in the top rows of Table 2. Our purpose in comparing the findings using the two different scoring approaches —continuous quality variables scores and “leveled”

3 Observed quality was examined in regression analyses that included site and, if relevant, treatment as covariates. Child outcomes were examined using hierarchical linear models that accounted for nesting of children in classrooms and included the child's fall scores, family characteristics, site, and, if relevant, treatment as covariates. Multiple imputation was used to account for missing data. Results from the analyses within each study were combined using meta-analysis.

4 Improving child outcomes is one, but not the only, goal of QRIS. We focus on it in this brief because child outcomes should be most strongly affected by higher quality classroom experiences. Child outcomes are also the focus of many policymakers. However, we want to recognize that there are other goals for QRIS (e.g., increasing family engagement, professionalization of the ECE workforce, creating connections across the ECE system).

5 Center-level quality scores were computed for each quality variable as the mean of the classroom categorized data for a given indicator across classrooms within a center.

quality indicator scores —was to assess how similar or different they were. If findings obtained using the categorized quality indicators are similar to those obtained using the continuous measures, we would feel confident that the scoring of the items was acceptable (in other words, the categorized quality indicator scores would appear to be maintaining most of the information in the continuous quality variable scores).

Findings revealed small, but statistically significant, associations between almost all selected continuous quality variables and child outcomes, and a similar but somewhat stronger pattern of association between the categorized quality indicator and child outcomes. The CLASS Instructional Support variable was significantly related to pre-literacy, and the CLASS combined rating was related to both language and pre-literacy scores. Director and teacher education was related to language, pre-literacy skills, and math as continuous quality variables and to language and pre-literacy skills as categorized quality indicators. Director, but not teacher, education also related to math as a categorized indicator. Ratio was related to social skills as both a continuous and categorized variable (a marginally significant finding). Group size was related —but in the wrong direction —to math skills as a categorized, but not continuous, measure. Curriculum was related to social skills, and was only measured as a categorized variable.

In summary, these analyses provided some support for almost all of the selected quality variables. Results from analyses involving continuous and categorized quality measures tended to yield similar findings — suggesting the professional guidelines provided meaningful and useful cut-points for creating the quality indicator ratings. Of note were findings indicating that ECERS-R total, CLASS Emotional Support, CLASS Classroom Management, and group size did not show reliable associations with any of the child outcomes in the anticipated directions.

QRIS Summary Rating Score and Child Outcomes: Lastly, we created a QRIS summary rating score. The QRIS summary rating score included both structural and process quality measures and was related to the child outcomes. We created a center-level overall QRIS rating score by calculating the mean of all seven indicators —the two process quality indicators and the five structural quality indicators. Predictions to child outcomes were analyzed using the summary QRIS rating scores (see bottom rows of Table 2). The overall QRIS rating score was a significant and modest predictor of two of the four child outcomes: language and reading skills, and a “marginal” predictor of math.

In summary, the QRIS rating focusing on classroom experiences (including both structural indicators process indicators) appeared to be validated in analyses that showed significant (although modest) associations with child outcomes.

Conclusions

The analyses using simulated ratings suggest that QRIS ratings can achieve their desired goal of predicting gains in child outcomes when attention is paid to the psychometric issues of dimensionality, selecting items with strong evidence, and scoring items using established criteria for cut points. Analyses provided significant validation —albeit modest, in terms of the strength of associations—of almost all of the carefully selected quality measures of classroom experiences. This finding was observed when measures were analyzed as individual quality indicators and when they were combined into a summary QRIS rating of classroom experiences. It also held true whether focusing only on structural indicators, or including both structural and process indicators of quality.

The analyses provide an example —not a blueprint—for using such an approach. For example, the simulated summary rating did not include some additional quality indicators with strong evidence, because the selected datasets lacked measures of these aspects of quality. In addition, other QRIS foci, such as family engagement and cultural responsiveness that are less established in the research literature, likely need

similar attention to psychometric principles to determine how QRIS can most validly represent these important aspects of ECE quality. It is likely that these quality dimensions would comprise separate QRIS ratings that would be validated in relation to different desired outcomes. However, we were not able to test these hypotheses with available data.

In conclusion, attention to these psychometric principles in creating or adapting QRIS scales should improve their ability to be validated with desired outcomes.

Table 1: Categorizing Variables Measuring Quality of Classroom Experiences

Quality Level	Teacher Education ^a	Child:Adult Ratio ^a	Curriculum	Director Education ^a	Group Size ^a	ECERS-R	CLASS ^a
0	No College/ HS	3 years: > 9:1 4 years: > 10:1	No Curriculum	No College/ HS	3 years: > 18 4 years: > 20	< 3	IS < 2 or CO < 3 or ES < 4 or
1	Some College or CDA	3 years: ≤ 9:1 4 years: ≤ 10:1	Global Curriculum only	Some College or CDA	3 years: ≤ 18 4 years: ≤ 20	3-5	IS = 2+ CO = 3+ ES = 4+
2	College (BA/ BS) & ECE	3 years: ≤ 7:1 4 years: ≤ 8:1	Literacy Curriculum and/or Technical Assistance	College (BA/BS) & ECE	3 years: ≤ 14 4 years: ≤ 16	5+	IS = 3-7 & CO = 5-7 & ES = 6-7

^a Note: Classrooms were assigned to the middle quality level (1) if the criteria for the top level were not met and criteria for the middle level were met

Table 2. Estimated Effect Sizes:^a Results from meta-analyses relating process and structural quality variables to child outcomes

	Language ^b	Pre-Literacy ^c	Math ^d	Social Skills ^e
Center Quality Indicators as Continuous Variables				
<i>ECERS-R Total</i>	.01 (.01)	.02+ (.01)	-.00 (.01)	.02 (.02)
<i>CLASS Instruct Support</i>	.03 (.02)	.06*** (.01)	.01 (.02)	.03 (.02)
<i>CLASS Emotion Support</i>	.02 (.02)	.01 (.02)	-.02 (.02)	.02 (.02)
<i>CLASS Classroom Org</i>	.02 (.02)	.05** (.02)	-.01 (.02)	.02 (.02)
<i>Director Education</i>	.04** (.01)	.04** (.01)	.05** (.01)	.01 (.02)
<i>Teacher Education</i>	.03** (.01)	.07*** (.01)	.04** (.01)	-.01 (.01)
<i>Child:adult Ratio</i>	-.01 (.01)	.00 (.01)	.01 (.02)	-.04* (.02)
<i>Group Size</i>	-.02 (.01)	.02 (.02)	.02 (.02)	-.00 (.02)
Scored Center Quality Indicators				
<i>ECERS-R Total</i>	.03 (.03)	.01 (.03)	.00 (.03)	.07+ (.04)
<i>CLASS combined</i>	.08* (.03)	.13** (.04)	.06 (.05)	.07 (.07)
<i>Director Education</i>	.07** (.03)	.08** (.03)	.09** (.03)	.05 (.04)
<i>Teacher Education</i>	.02 (.02)	.13*** (.03)	.04 (.03)	-.04 (.04)
<i>Child:Adult Ratio</i>	.02 (.02)	.00 (.02)	.00 (.01)	.04+ (.03)
<i>Group Size</i>	.04+ (.02)	-.00 (.02)	-.10** (.03)	.00 (.03)
<i>Curriculum</i>	-.04 (.04)	.00 (.04)	.04 (.04)	.12* (.05)
Overall QRIS Rating	.07*** (.02)	.07* (.03)	.05+ (.03)	.03 (.03)

Note: +.1 < p < .05; * p < .05; ** p < .01; *** p < .001

^a The effect sizes can be interpreted as analogous to partial correlations. For correlations, .10 is viewed as modest, .30 as moderate, and .5 as large.

^b **Language measures.** Most studies used Peabody Picture Vocabulary Test- Ed 3 or 4 (Dunn & Dunn, 1997, 2007) to measure language. One study used the Woodcock Johnson III Picture Vocabulary Subscale (Woodcock, McGrew, & Mather, 2001). One study did not measure language. All studies that measure language tested Spanish-speaking children in Spanish if not proficient in English, and we used the scores from the Test de Vocabulario en Imagenes Peabody (Dunn, Lugo, & Dunn, 1997) or Batería III Woodcock-Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005) in these analyses, along with a dummy variable indicating whether the Spanish language assessment was used for that child.

^c **Literacy measures.** All but two of the studies used the Woodcock Johnson III Letter-Word Identification Subscale (Woodcock et al., 2001). The other studies used the Print Knowledge Scale of the Test of Preschool Early Literacy (Lonigan, Wagner, Torgesen, & Rashotte, 2007) and an early literacy assessment that consisted of 74 items from major early literacy measures that assessed letter knowledge, word recognition, print conventions, and phonological awareness (ECLS-B; Najarian, Snow, Lennon, Kinsey, & Mulligan, 2010).

^d **Math measures.** All but two of the studies used the Woodcock Johnson III Applied Problems Subscale (Woodcock et al., 2001). One study (NCRECE) did not measure math skills and the other used a math assessment with 58 items drawn from major early math measures and focused on number sense, property, operations, and probability (ECLS-B; Najarian et al., 2010).

^e **Social-emotional adjustment measures.** Most studies relied on teacher ratings, using the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) or the updated version, Social Skills Improvement System Rating Scales (SSIS-RS; Gresham & Elliott, 2008). FACES and ECLS-B relied on parent ratings, using selected items from the SSRS and the Preschool and Kindergarten Behavior Scales – Second Edition (Merrell, 2003).

References

- American Academy of Pediatrics, American Public Health Association, National Resource Center for Health and Safety in Child Care and Early Education (2011). *Caring for our children: National health and safety performance standards; Guidelines for early care and education programs*. 3rd Edition. Elk Grove Village, IL: American Academy of Pediatrics; Washington, DC: American Public Health Association. Available Online: <http://nrckids.org>.
- Build Initiative & Child Trends. (2015). *A Catalog and Comparison of Quality Rating and Improvement Systems (QRIS)* [Data System], Retrieved from <http://griscompendium.org/>
- Burchinal, M., Soliday Hong, S., Sabol, T., Forestieri, N., Peisner-Feinberg, E., Tarullo, L., Zaslow, M. & Martinez-Beck, I. (2016). *Quality Rating and Improvement Systems: Secondary data analyses of psychometric properties of scale development*. OPRE Report #2016-26. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Dunn, L. M., & Dunn, D. M. (1997). *PPVT-III: Peabody picture vocabulary test*. Minneapolis, MN: NCS Pearson.
- Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Minneapolis, MN: NCS Pearson.
- Dunn, L. M., Lugo, P., & Dunn, L. M. (1997). *Vocabulario en imágenes Peabody (TVIP)*. Circle Pines, MN: American guidance service.
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system (SSRS)*. Circle Pines, MN: American Guidance Service.
- Gresham, F., & Elliot, S. N. (2008). *Social skills improvement system (SSIS) rating scales*. Bloomington, MN: Pearson Assessments.
- Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., ... & Scott-Little, C. (2012). A course on effective teacher-child interactions effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, 49(1), 88-123.
- Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale, revised edition*. New York, NY: Teachers College Press.
- Hestenes, L. L., Kintner-Duffy, V., Wang, Y.C., La Paro, K., Mims, S.U., Crosby, D., Scott-Little, C., & Cassidy, D.J. (2014). Comparisons among quality measures in child care settings: Understanding the use of multiple measures in North Carolina's QRIS and their links to social-emotional development in preschool children. *Early Childhood Research Quarterly* (2014). DOI: 10.1016/j.ecresq.2014.06.003
- Lambert, R. G., Nelson, L., Brewer, D., & Burchinal, M. (2006). Measurement issues and psychometric methods in developmental psychology. In K. McCarthy, M. Burchinal, & K. L. Bub (Eds.), *Best practices in quantitative psychology for developmentalists. Monograph of the Society for Research in Child Development*, 71(3), 24-41.
- Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *TOPEL: Test of preschool early literacy*. Austin, TX: Pro-Ed.
- Malone, L, Carlson, L., Aikens,N., Moiduddin, E., Klein,K., West, J., Kelly,A., Meagher, C., Bloomenthal, A. Hulse, L., & Rall, K. *Head Start Family and Child Experiences Survey: 2009 User's Manual*. Report submitted to the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Washington, DC: Mathematica Policy Research, April 2013

Merrell, K. W. (2003). *Preschool and Kindergarten behavior scales (PKBS-2)*. Austin, TX: PRO-ED.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005). *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

Najarian, M., Snow, K., Lennon, J., Kinsey, S., & Mulligan, G. (2010). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B): Preschool-kindergarten psychometric report*. Washington, DC: U.S. Department of Education.

Peisner-Feinberg (2013) *North Carolina Pre-Kindergarten Program Evaluation: Summary of Research 2002-2013*. . Chapel Hill: The University of North Carolina, FPG Child Development Institute. Downloaded on October 15, 2014 from <http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/Summary%20of%20NC%20Pre-K%20Evaluation%20Findings%205-2014.pdf>

Peisner-Feinberg, E.S., Schaaf, J.M., LaForett, D.R., Hildebrandt, L.M., & Sideris, J. (2014). *Effects of Georgia's Pre-Kindergarten Program on children's school readiness skills: Findings from the 2012–2013 evaluation study*. Chapel Hill: The University of North Carolina, FPG Child Development Institute. Downloaded on October 15, 2014 from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval_RDDReport%203-4-2014.pdf

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System Manual: Pre- K*. Baltimore, MD: Brookes.

Sabol, T. J. & Pianta, R. C. (2014). Validating Virginia's quality rating and improvement system among state-funded Pre-Kindergarten programs. *Early Childhood Research Quarterly*, online first publication. DOI: 10.1016/j.ecresq.2014.03.004

Sabol, T. J., Soliday Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can ratings of Pre-K programs predict children's learning? *Science*, 341, 845–846.DOI: 10.1126/science.1233517

Soliday Hong, S.L., Howes, C., Marcella, J., Zucker, E., & Huang, Y. (2014). Quality Rating and Improvement Systems: Validation of a Local Implementation in LA County and Children's School-Readiness. *Early Childhood Research Quarterly*.

Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009, October). *The Missouri quality rating system school readiness study*. Columbia, MO: Center for Family Policy & Research.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *The child care Quality Rating System assessment: Compendium of Quality Rating Systems and evaluations*. Washington, D.C.: Office of Planning, Research and Evaluation.

U.S. Department of Education, National Center for Education Statistics. (2007). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) 9-Month—Preschool Restricted-Use Data File and Electronic Codebook (CD-ROM)*. (NCES 2008-034). Washington, DC: Author.

West, J, Aikens,N., Carlson, B., Meagher, C.,Malone, L., Bloomenthal, A., Kelly, A., Rall, K., & Zota, R. *Head Start Family and Child Experiences Survey: 2006 User's Manual*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, August 2010.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Zellman, G.L., Perlman, M., Le, V., & Setodji, C.M. (2008). *Assessing the validity of the Qualistar early learning quality rating improvement system as a tool for improving child-care quality*. Santa Monica, CA: RAND.