OPRE Report 2016-26
May 2016

Quality Rating and Improvement Systems: Secondary Data Analyses of Psychometric Properties of Scale Development

Quality Rating and Improvement Systems: Secondary Data Analyses of Psychometric Properties of Scale Development

OPRE Report #2016-26

May 2016

Submitted by: Margaret Burchinal, University of North Carolina
Sandra L. Soliday Hong, University of North Carolina
Terri J. Sabol, Northwestern University
Nina Forestieri, University of North Carolina
Ellen Peisner-Feinberg, University of North Carolina
Louisa Tarullo, Mathmatica Policy Research
Martha Zaslow, Society for Research in Child Development, Child Trends

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

Contract Number: HHSP23320095642WC

Project Director: Louisa Tarullo
Mathematica Policy Research

Suggested citation: Burchinal, M., Soliday Hong, S., Sabol, T., Forestieri, N., Peisner-Feinberg, E., Tarullo, L. and Zaslow, M. (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.

Acknowledgements: The analysis in this report was conducted as part of the Child Care and Early Education Quality Features, Thresholds, and Dosage and Child Outcomes Project (Q-DOT). The project was conducted by Mathematica Policy Research and its subcontractors Child Trends, the University of North Carolina-Chapel Hill, and the University of Virginia. One author also received funding from the IES-funded post-doctoral fellowship R305B100028 awarded to UNC-Chapel Hill, and one author received funding from grant F32 HD076557 awarded to Northwestern University. We are grateful to the study authors, state leaders, programs and children in North Carolina and Georgia for allowing us to include the data from their pre-kindergarten evaluations in our analyses.

Disclaimer: The views expressed in this report do not necessary reflect the views or policies of our funders: OPRE, ACF, DHHS, IES, and the states of North Carolina and Georgia.

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.

Quality Rating and Improvement Systems: Secondary Data Analyses of Psychometric Properties of Scale Development

Margaret R. Burchinal,
University of North Carolina

Sandra L. Soliday Hong
University of North Carolina

Terri J. Sabol
Northwestern University

Nina Forestieri
University of North Carolina

Ellen Peisner-Feinberg
University of North Carolina

Louisa Tarullo
Mathematica Policy Research

Martha Zaslow
Society for Research in Child Development, Child Trends

Correspondence for questions concerning this article should be addressed to Margaret Burchinal, Frank Porter Graham Child Development Institute, 521 South Greensboro Street, Chapel Hill, NC 27599-8185; Burchinal@unc.edu.

Executive Summary

The results of this secondary data analysis simulating a QRIS validation using six large early care and education datasets demonstrate several issues that should be considered when constructing, validating, and making changes to existing quality ratings. First, QRIS are developed from logic models that involve multiple outcome areas such as improving children's outcomes, professionalization of the workforce, family engagement, and ECE systems building. The analyses reported here suggest that separate QRIS rating scales will be needed for each of these dimensions unless they are highly correlated. Second, selection of the quality indicators should be based on the consistency and magnitude of effects in research literature. The QRIS rating is more likely to accurately measure quality when there is good evidence that we know how to measure the included quality indicators in a manner that predicts desired outcomes for the QRIS. Third, use of validated professional guidelines for defining the cut-points in the rating scales can maintain the information in the selected quality measures as they are converted into ratings to form the QRIS score. Results from this secondary data analysis suggest that a QRIS score reflecting classroom quality based on these principles predicts small but significant gains in children's academic outcomes.

Quality Rating and Improvement Systems: Secondary Data Analyses of Psychometric Properties of Scale Development

Quality Rating and Improvement Systems (QRIS) are state or local market-based policy initiatives designed to increase access to high quality early care and education (ECE). QRIS seek to improve ECE quality by publishing the quality ratings of enrolled programs so parents can make informed child care choices and by providing quality improvement activities to those programs (Tout, Starr, Soli et al, 2010) . Each state or locality develops their rating scale from selected direct and indirect indicators of ECE quality with the goal of producing a summary rating that concisely describes the level of quality of each ECE program. The purpose of this study is to illustrate how psychometric properties of scale development can be useful in the development and refinement of QRIS ratings, using the data from six large studies of early care and education.

The Multiple Goals of QRIS

Starting about 20 years ago, states created market-based incentive systems as part of an effort to develop an ECE system within states, improve access to high quality ECE, and professionalize the ECE workforce (Tout, Zaslow, Halle & Forry, 2009; Zellman & Perlman, 2008). QRIS are designed to provide a standard way of rating ECE program quality based on multiple criteria and making the rating available to parents, as well as other key target groups such as policymakers and the programs themselves. QRIS are also intended to provide a guiding framework to both assess and improve quality systematically under one uniform system that can apply across the various early care and education sectors, including community-based child care centers and family child care homes, Head Start, and public prekindergarten. The QRIS should motivate programs to improve quality to the extent that (1) parents use the ratings in selecting

care settings, and (2) states provide quality improvement resources to programs across the full range of quality and give higher reimbursement rates from subsidies to higher quality programs. In this context, lower quality providers would face an incentive either to improve the quality of their program or leave the market (Zellman & Perlman, 2008). Thus, QRIS attempt to improve quality by affecting both the demand for high quality care and the supply.

The success of such efforts rests on the ability of rating systems to accurately identify and measure key aspects of quality that relate to the stated outcomes of the QRIS. QRIS were initiated as a means to produce a number of outcomes, including improving ECE program quality, increasing children's learning by supporting the language, cognitive and social skills development of young children enrolled in ECE programs, professionalization of the ECE workforce, increasing parent engagement, and greater collaboration across ECE systems (Zaslow & Tout, 2014). QRIS seek to achieve these outcomes through incentivizing quality improvement in the various domains of ECE quality including classroom quality, family and community engagement, professionalization of the workforce, and the development of an integrated ECE system (Zaslow & Tout, 2014). Some of these outcomes are more strongly associated with some of the QRIS quality improvement efforts and indicators – such as improving children's academic and social skills through raising classroom quality or increasing parent engagement through enhanced parent outreach in the ECE programs. To date, however, QRIS ratings combine all of the quality indicators into a single rating, which may help explain why validation of QRIS scores relative to child outcomes have yielded mixed findings (Soliday Hong, Howes, Marcella, Zucker, & Huang, 2014; Sabol & Pianta, 2014; Thornburg, Mayfield, Hawks, & Fuger, 2009). This study will draw upon evaluation theory and psychometric guidelines to examine the extent to which ECE quality appears to be multi-dimensional. If so, we

will focus on the association between a QRIS rating of one specific aspect of quality and the desired outcome associated with that aspect of quality to demonstrate that a focus on psychometric development of scales can yield ratings that are associated with one of the desired outcomes, children's academic and social skills.

Most states have developed their quality ratings through articulating a theory of change model, selecting the quality criteria to include within the model based on research evidence and then developing their rating system through a consensus process with stakeholders. Setting aside the possibility that QRIS have multiple outcomes, the model underlying most QRIS is typically aligned with researchers' conceptualization of ECE quality (NICHD ECCRN, 2003). For example, with respect to promoting child outcomes, this quality model posits that "structural" or indirect quality variables create the conditions in which ECE providers can provide high quality environments for and interactions with children (NICHD ECCRN, 2002). In turn, higher quality teacher-child interactions and child care environments lead to larger gains in children's cognitive and social skills (Mashburn et al., 2008; Pianta, Barnett, Burchinal, & Thornburg, 2009).

Based on the ECE literature, states have created quality rating systems to apply decades of evidence on the importance of ECE quality for children's development to a policy context. A recent compendium summarized the state and local QRIS rating systems in 2014 (Build Initiative & Child Trends, 2015). The authors reported that almost all QRIS ratings include some indirect measures of ECE quality such as caregiver education and training, curriculum, and program administration, and more direct measures of quality such as observations of the environment. While this paper focuses on rating of quality components most closely related to child outcomes, according to the compendium, at least half of QRIS ratings include other indirect measures such as ratio and group size, health and safety, child assessments, accreditation, community

involvement, continuous quality improvement and direct measures of teacher-child relationships and parent engagement.

The selected ECE quality measures are scored and typically grouped into categories or quality domains. Each quality variable is scored by assigning a rating that indicates lower to higher levels of quality for that variable. The quality domains typically reflect the various dimensions of ECE quality based on that state's or locality's logic model. The scores for the quality domains are computed, weighted if deemed necessary, and then combined to form a single rating using a block, point, or hybrid system. Some systems upweight some items or domains based on their logic model that considers those domains crucial for assessing ECE quality. The block system assigns an overall rating for an ECE program based on the lowest domain score, and therefore communicates that the quality within each domain is at that level or higher at that program. The points system sums the points across domains and assigns an overall score based on that sum, thereby communicating the average level of quality across domains for a program. The hybrid combines the two approaches, requiring a minimum score within each domain for a given quality score and then summing points to determine the final rating if those minimum score criteria are met. This approach communicates that a given level of quality is present in each domain, while the overall quality score may suggest higher quality across domains. The decisions regarding the quality variables to include, the cut-scores to use to create the quality ratings for each variable, the variables to use in each domain score, and how to combine scores across domains all will impact the ability of the QRIS rating to accurately describe ECE quality as related to child outcomes for a given ECE program and thus to predict outcomes across ECE programs. These issues are discussed in more detail below and examined empirically in our secondary data analysis.

The mixed findings from QRIS validation studies suggest that greater attention to psychometric properties is needed. To date, there has been some evidence showing validation using QRIS outcomes such as observed classroom quality or increased parent involvement in the ECE program and far less evidence using QRIS outcomes such as gains in child outcomes or changes in parenting. Two studies focused on validating the indirect or structural quality variables, and reported modest to moderate associations between QRIS structural or indirect quality indicators and the quality of the classroom environment and teacher-child interactions (Hestenes et al., 2014; Jeon, Buettner, & Hur, 2014). Four evaluation studies related overall QRIS ratings to child outcomes, and reported either small or no associations with child outcomes (Hestenes et al., 2014; Sabol & Pianta, 2014; Soliday Hong et al., 2014; Thornburg, Mayfield, Hawks, & Fuger, 2009; Zellman, Perlman, Le, & Setodji, 2008). Finally, a secondary data analysis also simulated the QRIS based on selected states' established systems, and related these scores to both observed classroom quality and gains in child outcomes among preschoolers in public prekindergarten programs (Sabol, Soliday Hong, Pianta, & Burchinal, 2013). They found that the simulated QRIS of only one state reliably predicted child outcomes, and that QRIS rating scale differed from the other ratings scales in two ways – it only included the quality indicators with moderately strong replicated empirical evidence and it combined the indicators using a points system.

We argue that careful attention to the selection of indicators of quality to be included in QRIS as well as to how the rating scale is computed (as a block, point or hybrid system) will be important to the ability of the QRIS to predict to desired outcomes, including child outcomes. Further, we argue that attention to the psychometric properties of the ratings will also be critical.

We turn now to a discussion of the issue of psychometric properties of rating scales that should be taken into account when creating QRIS summary scores.

Psychometric Properties of Rating Scales

Test theory provides several guidelines for constructing rating scales that, when implemented, will increase our ability to measure the underlying construct (Lambert, Nelson, Brewer, & Burchinal, 2006). Psychometrically, a rating scale provides better measurement when items are carefully selected to align with the intended outcomes, appropriately scored, and measure the same construct. These properties relate to the dimensionality of the measure, item selection, and the reliability of the scale. Such properties are important when scales are used in research to demonstrate adequate measurement of important constructs, and are even more important when they are used to guide policy and practice to ensure that decisions are being made as accurately and fairly as possible.

The first test theory guideline involves the dimensionality of the scale (Allen & Yen, 2001). Each scale should be measuring a single dimension and should include multiple scales if measuring a construct with multiple dimensions. Ideally, each scale within the instrument should consist of items that all provide strong measurement of different aspects of the same underlying construct. Only then can the scale scores reliably reflect the extent to which that construct is being manifested. For example, a math scale that consists of items measuring a reasonable variety of age-appropriate math skills should provide a good index of a child's understanding of math. When the items measure different constructs, often indexed by low correlations among the items, the scale score cannot provide a good index of a single construct. With the example of the math scale, the total score would not reflect math skills if it included

items measuring reading and attention skills, and it is unlikely that the total score would provide as good an index of math as a scale composed exclusively of math items.

Most quality indicators included in QRIS can be seen as relating to four broad dimensions (Zaslow & Tout, 2014): (1) both structural and process aspects related to children's experiences in classrooms, discussed above as the indirect and direct aspects of quality related to children's experiences in ECE and their developmental outcomes (including group size, use of a curriculum, teacher qualifications and professional development, accreditation status, use of child assessments, observed quality, and provision for special needs); (2) family engagement with the full range of families (including family partnerships, provisions for cultural and linguistic diversity, and regular communication with families); (3) professionalization of the workforce (including creating career lattices and systems such as registries for marking teacher and caregiver progress on career lattices); and (4) quality indicators related to ECE as a system and the functioning of programs within that system (including program administration, licensing compliance, and alignment with early learning standards).

If this is the case, then the total score will be a mix of all of these dimensions together and reflect each dimension less well than would separate scales. Validation would be impaired because the total score would correlate with desired outcomes less well than would separate scale scores for each dimension and the desired outcome associated with that dimension (e.g., classroom quality and child outcomes). Quality improvement efforts in specific dimensions might be successful, but may not be reflected in the QRIS ratings because the other dimensions might mask the impacts of that improvement.

Ideally, items are carefully selected as providing measurement of different aspects of that dimension. In the context of the math assessment example, this approach would involve

selecting items that measure the relevant age-appropriate math skills. Within the context of QRIS, this strategy would involve selecting indicators that produce higher quality ECE within that dimension of the QRIS and using the magnitude and consistency of those associations in the research literature to guide the selection of indicators.

Focusing on the Goal of Improving Child Outcomes: Direct and Indirect Quality Indicators

Even when focusing on the specific aspects of QRIS ratings that are thought to contribute to child outcomes, it is important to distinguish between those indicators thought to be most directly and immediately predictive of child outcomes, and those that are important but that are indirectly related to these outcomes. The aspects of quality that are most directly related to children's development include the opportunity for diverse age-appropriate activities, inclusion of intentional learning activities as part of daily routines, and the quality of interactions that children have with the adults and peers in the classroom. Others aspects of ECE are more distal to a child's direct experiences, but are viewed as important because they can create the conditions for high quality through children's immediate experiences in ECE settings. These structural quality variables include factors such as the professional development of teachers and caregivers, use of an evidence-based curriculum, ratio and group size.

A key first step is narrowing the focus to the aspects or dimensions of QRIS that are intended to improve child outcomes (rather than the goals of family engagement, professionalization of the workforce, or ECE systems). A second step, having narrowed the focus to the aspects of QRIS intended to improve child outcomes, is to select the strongest possible quality indicators. The QRIS rating scale is more likely to provide good measurement when items are selected because the ECE literature has shown that the quality component is related

either to observed classroom quality or child outcomes across a number of studies and the magnitude of those associations are moderate to large. Thus, challenges to the development of the QRIS rating include variability in the level of evidence across the various components, combining disparate quality dimensions into a single score, and focusing only on outcomes related to a single dimension. Given the four dimensions listed above, we need to be explicit as to which indicators predict child outcomes and which predict other outcomes. QRIS developers thus should specify if there are multiple logic models in which different QRIS dimensions predict different outcomes such as improved language, cognitive and social skills in children, higher levels of parent engagement, higher levels of professionalization within the workforce, and increased collaboration and alignment across the different types of programs within the ECE system. Careful attention to the alignment between quality indicators and intended outcomes should increase the ability of QRIS domain scores to predict those aligned outcomes.

Item selection is critical for reliable and valid measurement (Lambert et al., 2006). Ideally items are selected because they provide reliable measurement of indicators of the uni-dimensional construct being measured by that scale (Allen & Yen, 2001). Within the ECE literature, there are a few quality factors that have replicated evidence that they are related to overall classroom or setting quality or to child outcomes, and many quality factors thought to be important with much less empirical evidence. A recent literature review (Burchinal, Magnuson, Powell, & Hong, 2015) provides replicated evidence with modest to moderate effect sizes for several QRIS quality factors as they relate specifically to child outcomes: (1) Curricula--using an evidence-based curriculum with aligned training or coaching is related to substantial gains in children's literacy, math, and social skills and there is some evidence that focused, sequenced curricula produce larger gains; (2) Child-teacher ratios and group sizes--settings with large

numbers of children per teacher and with larger group sizes have been reported to be lower quality and to produce larger behavior problems and smaller gains in academic skills; (3) Teaching staff qualification--settings with providers with higher levels of education have been shown to be of higher quality and to produce larger gains in academic skills (although some evidence suggests this finding might be due to other confounding factors); and (4) Program administration and leadership – settings directed by individuals with more education and ECE training and that offer higher wages and benefits to their staff have been rated as providing higher classroom quality in multiple studies, although some studies did not replicate findings for staff wages.

The ECE literature also indicates that process quality measures predict gains in children's academic and social skills during their time in the ECE setting, and sometimes at older ages (Burchinal et al., 2015). There are professional guidelines that have been developed based on this extensive literature for each of these structural and process ECE quality measures (American Academy of Pediatrics, 2011). In contrast, as we have noted, other key indicators included in QRIS are hypothesized to be related to outcomes other than child outcomes. As one example, the existence of a career lattice and reliance on a registry may be related to professionalization of the workforce, but not to children's development. In addition, there may be measures of quality that, while assumed to be related to child outcomes, have less extensive evidence in support of this relationship. Examples include child health and safety practices and support for the child's home language. There is much less empirical evidence linking ECE practices to desired outcomes, and often the existing evidence tends to be contradictory. Measurement of these important ECE quality factors need further work to ensure we are measuring them accurately. Thus, QRIS rating scales provide substantially less accurate measures of ECE quality under the following

conditions: when quality is multi-dimensional and (1) the scale combines indicators across dimensions or (2) when all indicators are related to the same quality dimension, but the scale combines indicators with both strong and weak measurement.

Optimally, the items are scored in a manner that minimizes the loss of information in going from a continuous variable to categorical ratings. Within the QRIS, optimal scoring is most likely to happen when ratings are based on professional guidelines founded on considerable research evidence. For example, staff qualifications can be measured by the proportion of teachers and caregivers in the ECE program with a bachelor's degree (BA) or higher. This component is typically turned into an indicator by assigning points on a Likert scale to specific proportions of staff with these qualification present in a program (e.g., 0 = less than 25% of providers have a BA, 1 = 25–50% have a BA, 2 = 50–75% have a BA, and 3 = 75–100% have a BA). Successful item scoring retains as much information as possible from the original quality variable in the scale. This can happen when scoring rules capture the important variability in the original variable. In our example, the staff qualifications would be successfully scored if, on average, the quality of the child care center was highest when at least 75% of teachers have a BA, slightly lower when 50-75% of the teachers have a BA, lower when 25-50% of the teachers had a BA, and lowest when less than less than 25% of teachers have a BA. One way to determine whether item scoring maintained this information is to compare the correlations between a desired outcome and both the original quality variable (e.g., proportion of teachers with a BA) and the scored quality indicator (e.g., the 4-level ranking). Loss of information in scoring quality indicators is indicated when the correlation between a desired outcome (e.g., classroom quality or child outcomes) and the continuous quality variable (e.g., years of education) is much larger than the correlation between that outcome and the categorical quality

rating indicator. Arbitrary or misaligned scoring will reduce or eliminate the ability of a quality indicator to measure that element.

Finally, the reliability of the scale should focus on the intended purpose of the data and be rigorous so that substantial measurement error does not contribute to limitations of the quality rating. For example, when reporting the child-teacher ratio of a classroom, it is important for the QRIS rating system to be clear whether the person gathering data to determine the rating is to record the number of children and teachers present during an observation or the number of children enrolled and staff members assigned to that classroom, regardless of whether they are present on the day of center is observed. Furthermore, the accuracy of reporting should be evaluated at the level the indicators are used in the QRIS rating scale. For example, if the cut-points are child-teacher ratios of less than 8:1, 8:1 to less than 10:1, 10:1 to less than 12:1, and greater than 12:1, then the accuracy of reporting will be very important around those cut-points and much less important between cut-points. A reporting error of 9.5:1 when the real ratio was 9.75:1 has no consequence, but a reporting error of 10:1 when the real ratio was 9.75:1 moves that center into a lower rating on that indicator. Similarly, if one of the Environmental Rating Scales (e.g., ECERS) is being used to assess process quality and the quality rating cut-point for a particular rating level is 5 on a scale from 1−7, then the difference between 4.7 and 4.9 is irrelevant to the program's assigned rating; however, the difference between a 4.9 and 5.1 score on that measure is much more consequential because it will result in a different assigned rating.

Auspice Differences in QRIS

Finally, the extent to which the auspice or type of ECE program affects validation of the QRIS ratings has been raised as an issue that should be considered when conducting validation studies (Sabol et al., 2013). ECE settings that are part of programs such as Head Start or public

pre-kindergartens often have to meet higher standards than other programs, and this requirement has led to concern that their inclusion in validation studies might impair the ability to validate the QRIS, due to more limited variability in the distribution of ratings across all programs. This concern is reflected, in part, in alternative pathways to star ratings for programs like Head Start and pre-kindergarten programs in many states, and examination of these alternative pathways are part of their QRIS validation in at least some of these states (QRIS Compendium, 2014). Direct examination of the validity of QRIS ratings in programs from different auspices and with differing ranges on component ratings would be extremely informative.

Present Study

This study was designed to illustrate the psychometric issues described above in the creation of a simulated QRIS rating scale, and to explore the relationship between the resulting ratings and observed classroom quality and child outcomes through secondary data analysis. To accomplish this purpose, we examined ECE quality indicators with the strongest research evidence as predictors of classroom process quality or child outcomes, including the quality variables of teaching staff and director education, child:adult ratios, group size, and curriculum. Although other quality indicators in QRIS may be predictors of other key outcomes, such as family engagement, professionalization of the workforce, or ECE systems outcomes, those outcomes were not available to us for these analyses. Therefore, they are not the focus of this study. Each of the indicators included here has been shown to predict either classroom process quality and/or child outcomes across multiple studies or is a widely used measure of ECE structural or process quality (Burchinal, Kainz, & Cai, 2011; NICHD ECCRN, 2002; Phillips, Mekos, Scarr, McCartney, & Abbott-Shim, 2000). Although director education has not been widely examined in the quality literature, we included it because QRIS ratings are center-level

ratings and director education could be a significant indicator of center-level quality, and thus important to classroom quality and child outcomes. Some evidence indicates it is an important quality indicator (e.g., Helburn, 1995). It is important to note that, whereas some QRIS directly include observations of classroom quality as an indicator contributing to the quality rating, in these analyses, observed classroom quality serves as both an outcome variable and a predictor, depending on the analysis.

We focused on structural measures of quality as predictors of observed process quality and child outcomes. We then categorized these measures, using the guidelines of the American Academy of Pediatrics and the American Public Health Association (American Academy of Pediatrics, 2011), to create quality indicators. We next examined the extent to which the selected structural quality variables and categorized ratings of each quality variable independently predict process quality and child outcomes. We created hypothetical overall quality ratings for each program based on the structural quality indicators and related them to process quality and child outcomes. We then added the process quality ratings to create hypothetical overall quality ratings based on both structural and process quality indicators and related them to child outcomes. Within this context, we examined item selection, item scoring, and dimensionality of our QRIS scale. Finally, we addressed questions about whether QRIS rating scales may be less reliable or valid when used to describe ECE settings within programs that have more stringent performance standards. The success of QRIS will depend, in part, on their ability to assign valid and reliable ratings of quality to programs and to predict child outcomes. We hope the results of these analyses can help identify what may be working and what may need further refinement within these systems.

Research Questions:

1. Is it possible to develop a QRIS rating with adequate psychometric properties using the selected structural quality indicators?
2. With the QRIS rating system developed using these psychometric properties, can we validate the individual indicators by demonstrating that both the indicators and the converted ratings are related to higher process quality and child outcomes?
3. With the QRIS rating system developed using these psychometric properties, can we validate the overall rating by demonstrating that it predicts both process quality and child outcomes?
4. Do QRIS scores predict classroom quality and child outcomes differently in the subset of programs with performance standards, such as Head Start, which might constrain the range of structural indicators of quality, as opposed to the full range of ECE programs?

Methods

We simulated QRIS ratings using secondary data from six large studies of ECE quality and children's acquisition of language, cognitive and social skills during the preschool year. We selected the studies because they included a large number of centers serving 3- and 4-year-old children (~100 or more), collected data on structural and process quality using indicators widely used in QRIS ratings, and measured child outcomes for a sample of children in those centers using widely used assessments of early academic and social skills. These studies included federal and state-funded ECE programs, as well as community-based, center-based ECE settings. The sample included two studies of Head Start: the Head Start Family and Child Experiences Survey (FACES) 2006 and 2009; two evaluations of state Pre-Kindergarten programs: the North Carolina Pre-Kindergarten Evaluation (NC Pre-K) and Georgia Pre-Kindergarten Evaluation

(GA Pre-K); and, two studies of preschool classrooms from different auspices: the preschool observational sample from the Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B) and the National Center for Research in Early Care and Education (NCRECE) professional development study.

Research Studies and Participants

Data from the six studies of ECE contributed to a meta-analysis. Participants from the six studies represented a diverse population of children and a range of ECE settings (see Table 1). ECLS-B and GA Pre-K were representative samples at the national and state levels, respectively. The Head Start FACES, and NC Pre-K studies focused primarily on low-income children and included representative samples of children enrolled in Head Start and the NC pre-kindergarten programs, respectively. NCRECE recruited a combination of Head Start and prekindergarten programs in 11 states. In the NC Pre-K and NCRECE, the pre-kindergarten programs were sometimes located in public schools and sometimes in nonprofit and for-profit community programs that met the state standards for prekindergarten. The sample participants in four out of the six studies were enrolled in a variety of center-based programs: local school systems, private non-profit and for-profit settings, and Head Start. The FACES data sets focused exclusively on children enrolled in Head Start settings. Within this variety of settings, programs in five out of the six studies were regulated by state or federal standards for ECE quality, resulting in a sample composed primarily of programs with moderate to high levels of quality. ECE programs in the GA Pre-K and NC Pre-K studies were required to follow state standards for participation in their respective state's programs, and the Head Start programs were regulated by federal Head Start guidelines. Details about these studies are provided below and in Tables 1 and 2.

National Center for Research in Early Care and Education (NCRECE). This study included preschool teachers who participated in a three-phase study evaluating the impacts of two forms of professional development—a 14-week course and year-long coaching. The first phase was a professional development course focused on improving knowledge and skills in observing interactions between children and their teachers/caregivers, the second phase was a coaching intervention that focused on teachers' actual classroom behaviors and interactions, and the third phase was a non-treatment post-intervention year. This secondary analysis employed data only from the third and final phase (post-intervention year). The recruitment process targeted large community preschool and Head Start programs across the country and included teachers in centers in New York City, NY; Hartford, CT; Chicago, IL; Stockton, CA; Dayton, OH; Columbus, OH; Memphis, TN; Charlotte, NC; and Providence, RI. During the initial recruitment, program administrators and teachers were invited to attend recruitment meetings in each location to learn about the study details. Teachers were considered eligible for participation if they were the lead teacher in a publicly funded classroom in which English was the primary language of instruction, the majority of children were eligible for kindergarten the following school year, and the settings were not primarily serving children with special needs.

Once teachers consented to participate, they were randomized at the community level into the course or control group for the first phase. Of the 440 recruited teachers, the 357 who completed that phase of the study and 73 newly recruited teachers were randomized in the second phase to the MyTeachingPartner (Pianta et al., 2014) coaching intervention or a control group. Of these teachers, 222 remained in their classrooms for an additional year and agreed to be followed into the third and final year of the study. This third, follow-up year with no intervention provided the sample for this study. The study included all teachers enrolled in this

follow-up year, regardless of their prior treatment conditions. Analyses accounted for treatment by including treatment dummy variables.

The majority of teachers included in these analyses (55%) worked in Head Start programs, and a smaller portion worked in public schools (35%). Teachers were experienced, with an average of 11 years of experience teaching preschool age children. They had diverse educational backgrounds (Associate's (AA) degree or less = 37%; BA degree = 44%; MA degree or higher =19%). Teachers' reported ethnicity as follows—47% reported African American, 33% White, and a smaller number reported Hispanic/Latino (12%) or other ethnicity (8%). Within a classroom, the majority of children were at or below the poverty line. Table 1 presents descriptive information on teachers and classrooms.

Up to four children per classroom of participating teachers were recruited in the follow-up year. Teachers were given child recruitment packets in the fall and asked to send them home to the parents of all children in their classrooms. The packets included a letter explaining the project, a parental consent document, a family contact form, and a family questionnaire. Teachers collected the completed packets from parents. Data collectors retrieved the returned consent packets from the teacher on the morning of the school/center visit and randomly selected and assessed four children from this group. The children's language and literacy skills were assessed by trained data collectors and their behavior rated by their teachers in the fall and spring. The children were 4.7 years of age on average in the fall of their preschool year, diverse (42% African American, 35% Hispanic/Latino), and about half had mothers with more than a high school education (see Table 1).

Family and Child Experiences Survey (FACES) 2006 and 2009. The Head Start FACES was first launched in 1997 as a periodic longitudinal study of program performance.

Successive nationally representative samples of Head Start children, their families, classrooms, and programs provide descriptive information on the population served; staff qualifications, credentials, and attitudes; Head Start classroom practices and quality measures; and child and family outcomes. The FACES data come from a battery of direct child assessments across multiple developmental domains in the fall and spring; interviews with children's parents and teachers about the child in the fall and spring; interviews with children's parents, teachers, and program managers about their backgrounds in the fall; and direct observations of classroom quality in the spring. All analytic results have been computed using sampling weights, so reported findings are representative of the cohort of children enrolled in Head Start settings in 2006 and 2009.

The FACES 2006 sample included 60 programs, 135 centers, 410 classrooms, 365 teachers, and 3,315 children who entered Head Start as 3- or 4-year olds in fall 2006 (West et al., 2010). Classroom observations were conducted in a representative sample of 335 classrooms attended by 3- and 4-year-old children enrolled in their first year of Head Start in Fall 2006. Children were assessed in fall of their first year, spring of each year they attended Head Start, and the spring of kindergarten. Approximately two-thirds (63%) of children in the sample were 3 years old at enrollment, and the others were 4 years old or older. Just over a third of children were Hispanic/Latino; another third were African American. On average, children were 4.49 years of age (SD = .5) at initial fall assessment (see Table 1).

The FACES 2009 sample included 60 programs, 129 centers, 486 classrooms, 439 teachers, and 3,349 children who entered Head Start as 3- or 4-year olds in fall 2009 (Malone et al., 2013). Children enrolled into FACES 2009 if they were entering Head Start for the first time in Fall 2009, and were assessed in fall and spring of each year they attended Head Start and in

the spring of Kindergarten. Sixty-one percent of children in the sample were 3-year-olds, and the remaining children were 4-year-olds or older. More than a third of children were Hispanic/Latino, and another third were African American. On average, children were 4.0 years of age (SD = .6) at fall assessment (see Table 1).

Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B): Preschool child care sample. The ECLS-B is a nationally representative sample of 14,000 children born in the United States in the year 2001 (U.S. Department of Education, 2007). The children were recruited from diverse socioeconomic and racial/ethnic backgrounds, with oversamples of Asian and Pacific-Islander children, American Indian and Alaska Native children, Chinese children, twins, and low-birth-weight children. About a fifth of the recruited sample reported using a language other than English at home. The children were followed from their recruitment at under 9 months old into kindergarten. As is required by ECLS-B data usage agreements, sample sizes are rounded to the nearest 5. All analysis results have been computed using the sampling weights, so reported findings are representative of the cohort of children born in the U.S. in 2001. A 25% random subset of families with 10+ hours/week of ECE was selected by ECLS-B to have their ECE setting observed to obtain quality ratings, with oversampling for low-income and diverse families. Children who lived in Alaska, Hawaii, or on Indian reservations, or who attended an ECE setting in which neither Spanish nor English was spoken, were not eligible for the observation study. The preschool ECE sample was 4.5 years at assessment on average and diverse: 23% African American, 19% Hispanic/Latino, and 39% of the children had mothers with a high school education or less. Child outcomes were collected in the winter of the child's last preschool year, so the outcomes sample for this study includes only children with 3 months or more in the ECE setting, to allow a minimum amount of time for children to have exposure to

that setting. The child's language and social skills from the 24-month assessments were included as controls in this sample.

North Carolina Pre-Kindergarten Evaluation (NC Pre-K). The North Carolina Pre-Kindergarten (NC Pre-K) program (formerly known as the More at Four Program) is a state-funded initiative providing a classroom-based educational program for at-risk 4-year-olds, designed to help them be more successful when they enter kindergarten (for full details, see Peisner-Feinberg, 2013; Peisner-Feinberg & Schaaf, 2007; 2008). Children are eligible for NC Pre-K based on family income (up to 75% of state median income or 300% of federal poverty status) and other risk factors (limited English proficiency, identified disability, chronic health condition, and developmental/educational need). The program first targets "unserved" children (those not already being served in a preschool program) and then "underserved" children (those in a program but not receiving child care subsidies and/or those in lower quality settings). NC Pre-K provides funding for pre-K classrooms at a variety of different types of programs, including public schools, Head Start, and community child care centers (both for-profit and nonprofit). Local sites are expected to meet a variety of program guidelines and standards around curriculum, training and education levels for teachers and administrators, class size and student-teacher ratios, North Carolina child care licensing requirements, and provision of other program services.

An independent evaluation study has been conducted each year, including various samples of classrooms and children, depending on the research design for that year. Several of the evaluation studies involved examination of program quality and gains in child outcomes during pre-kindergarten, based on a random sampling of classrooms each year. These studies included observations of the quality of pre-kindergarten classroom practices; fall and spring

assessments of children's language, literacy, math, and behavior skills; teacher surveys of demographic information; and use of statewide administrative data. This secondary analysis included data from three of these cohorts: 99 classrooms and 514 children from the 2003–2004 academic year, 57 classrooms and 478 children from the 2005–2006 year, and 50 classrooms and 321 children from the 2007–2008 year. Pre-K children were, on average, 4.5 years of age (range = 4.0–5.1 years) at fall assessment. At the time of enrollment, 37% of children were African American, 33% were European American, and 24% were Hispanic/Latino; 51% were male.

Georgia Pre-Kindergarten Evaluation (GA Pre-K). Georgia's Pre-K (GA Pre-K) is one of the few state-funded universal pre-kindergarten programs in the United States, with the aim of providing pre-kindergarten services to all 4-year-olds whose families want their children to participate in the program, regardless of family income level. In the 2011–2012 school year, GA Pre-K served more than 94,000 different children in a variety of settings across the state, including local school systems, private settings, and blended Head Start/Georgia's pre-kindergarten classrooms. The 2011–2012 evaluation study included observations of classroom quality in a random sample of 100 of Georgia's pre-kindergarten classrooms; fall and spring assessments of the language, literacy, math, and behavioral skills of a sample of 571 children attending these classrooms over the pre-kindergarten program year; teacher and parent surveys of demographic information; and use of statewide administrative data (Peisner-Feinberg, Schaaf, & LaForett, 2013). A sample of 100 classrooms was randomly selected from the 3,922 Georgia pre-K classrooms operating in August 2011, excluding 100 classrooms participating in an intensive professional development intervention study. Parents were given the option to opt out of the study; the families of 4.3% of enrolled students refused to participate, including all parents in one of the selected classrooms. On average, six children were randomly selected from each

classroom for assessment. GA Pre-K children at the fall 2011 assessment were 4.7 years of age on average, diverse (39% African American, 14% Hispanic/Latino), and most had mothers with more than a high school education (see Table 1).

Exclusion Criteria across Studies

We excluded children from the study based on the following criteria: the child moved out of the center during the year; the child could not be assessed in English or Spanish (i.e., did not pass a language screener in the spring and spoke a language other than English or Spanish); the child attended the program only once per week (in the FACES 2009 data); the child was not assessed at the endpoint of the preschool year (winter for ECLB-B and spring for all other studies), or the child had not attended the ECE program for at least 3 months prior to the endpoint assessment. The last condition applied only to the ECLS-B, in which some children had their preschool outcomes collected too early in the academic year to allow for classroom experiences to affect the child's development.

Measures

Quality measures. Quality measures were selected for this study using three criteria. First, they had to measure aspects of quality that are included in QRIS ratings scales (Tout et al., 2010; Build Initiative & Child Trends, 2014). The recent QRIS Compendium lists 14 aspects of ECE quality, each of which are included in at least 33% of the current QRIS ratings scales and most of which are included in more than half of the currents scales. Second, they had to have replicated evidence indicating that they were predictors of either observed classroom quality or child outcomes and be included in QRIS logic models as factors that lead to improved child outcomes. Third, they had to be measured in the selected studies. There were seven quality aspects that met all three criteria: process quality measures of child care environment and of

teacher-child interactions and structural quality measures of teacher and director education, child-teacher ratios and group size, and curricula. These are described below.

Process quality indicators. The studies varied in how they measured process quality in the classrooms. The Early Childhood Environmental Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998) is widely used as a measure of global quality, and was included in ECLS-B, FACES, NC Pre-K, and GA Pre-K. The ECERS-R is a well-established measure of the global classroom quality environment that assesses both structural and process quality. It includes seven general areas: personal care routines, furnishings and displays for children, language-reasoning experiences, fine and gross motor activities, creative activities, social development, and adult needs. Scores on each of 45 items can range from 1 to 7, with the overall mean score used as a global measure of the developmental appropriateness or quality of the classroom. To be consistent with other research, we did not include the adult needs items in the overall classroom quality scores. An overall score from 1 to 2.9 is considered poor quality; scores from 3 to 4.9 are considered medium to good quality; and scores of 5 or greater are considered good to excellent quality. The total scale has good internal consistency ($r = .921$: Harms et al., 1998), and raters must meet a criterion of at least 85% agreement within one point on ratings during training.

The Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2004, 2008) measures the quality of teacher-child interactions (here called an interaction-specific measure). It includes ratings on 10 items, scored on a 1−7 scale from low to high, which combine into scores on three overarching domains of classroom quality. The first domain, Emotional Support, encompasses four items: Positive Climate (the emotional connection among children and teachers); Negative Climate (expressed negativity, such as anger and hostility);

Teacher Sensitivity (responsiveness to children's concerns); and Regard for Student Perspectives (accommodations for children's points of view). The second domain, Classroom Organization, includes three items: Behavior Management (how effectively behavior is monitored or redirected); Productivity (how well time is organized to maximize learning activities); and Instructional Learning Formats (how well teachers facilitate children's engagement to maximize learning opportunities). The final domain, Instructional Support, incorporates three items: Concept Development (how teachers foster higher-order thinking skills); Quality of Feedback (how well teachers extend learning in their responses to children); and Language Modeling (facilitation of language). The scale has demonstrated good inter-rater reliability, ranging from 78.8% to 96.9% agreement within one point, with an average across all items of 87.1% agreement within one point. The domain scores show high internal consistency ($.83 < \alpha < .90$). The full CLASS was administered in NCRECE, FACES 2009, NC Pre-K, and GA Pre-K. The Instructional Support scale domain of the CLASS was used in FACES 2006.

Structural indicators of quality. Each of the studies collected information on five structural indicators frequently used in QRIS: teacher education, classroom child:adult ratios, presence of a curriculum, class size, and director education. In each of the studies, the teacher whose classroom was observed reported on her final degree and whether she had a degree or coursework in early childhood or a related discipline. She also reported whether a curriculum was used in the classroom and, if so, what the curriculum was, and whether there was training provided for it. The number of children and adults were observed when classrooms were observed, so the class size and child:adult ratio were computed based on observed presence of children and adults at one point in time. The director also reported on her final degree and whether she had a degree or coursework in early childhood or a related discipline.

Simulated QRIS ratings. To simulate QRIS ratings, we converted the scores on selected quality measures into quality ratings for each classroom, computed the average rating per center for each quality measure, and created the QRIS total from the center-level indicator ratings. First, each quality indicator was converted into quality ratings at the classroom level by assigning it a 0, 1, or 2 using the criteria discussed below and listed in Appendix A. This conversion was done at the classroom level because the criteria for group size and ratio were different depending on the age of children in the classroom. Second, we combined the quality indicators for each quality measure to form a center-level rating by computing the mean of the classroom-level quality ratings for that quality indicator. The number of classrooms with data varied from a single classroom per center in ECLS-B to having multiple classrooms in the Head Start and Pre-K studies. The center-level quality ratings for each quality measure were rounded so values were 0, 1, or 2. The quality ratings for the five structural variable and one or two process variables were calculated through computing a mean score of the quality ratings (i.e., simulating a points system), and this mean constituted our QRIS total rating.

Teacher and director education. Using APHA/APA criteria, the education levels of the teachers and director were scored, creating a quality component rating for each. Separate indicators described teacher and director education levels using the following criteria: (0) high school degree or equivalent, and the teacher/director had not attended a college or university; (1) the teacher/director had attended some college or had attained a child development associate's degree (CDA), an associate's degree, or a bachelor's degree without either a major or related coursework in early childhood or related fields; and (2) the teacher/director had attained a bachelor's degree or higher and had taken coursework related to early care and education. For the Georgia Pre-K study, we collected no information on the education level of the director or

principal; because principals are required to have coursework in early childhood, we assigned all programs the highest QRIS rating for director education.

Group size. The class size was scored was scored using APHA/APA criteria to create a quality component rating for group size. These criteria vary depending on the age of the children in the classroom. The following criteria were used for scoring for classrooms with at least one 3-year-old: (0) more than 18 children, (1)15 to 18 children, and (2) 14 or fewer children. For classrooms in which all children are 4-years-old or older, the following criteria were applied: (0) more than 20 children, (1) 15 to 20 children, and (2) 16 or fewer children.

The ratio of the number of children to number of teachers in the classroom was scored using APHA/APA criteria to create a quality component rating for the classroom ratio. These criteria also vary depending on the age of the children in the classroom. The following criteria were used for scoring for classrooms with at least one 3-year-old: (0) greater than 9:1; (1) less than or equal to 9:1 and greater than 7:1; and (2) less than or equal to 7:1. For classrooms in which all children are 4-years-old or older, the following criteria: (0) greater than 10:1; (1) less than or equal to 10:1 and greater than 8:1; and (2) less than or equal to 8:1.

Curriculum. The use of curricula was scored based on evidence from research studies. As discussed above, research indicates that use of a published curriculum is related to classroom quality and child outcomes, and use of content-specific curriculum and/or providing teacher with training on the published curriculum are both related to larger gains in child outcomes. Accordingly, the curriculum quality component ratings reflected the following: (0) program did not report using a curriculum or reported using one that was locally developed or unpublished; (1) program only used a published global curriculum not specific to a particular content area; and (2) program used a content-related curriculum (most commonly a literacy curriculum), provided

training or technical assistance related to its curriculum, and/or reported using its global curriculum a great deal. Less information about the curriculum was available in the ECLS-B data set, so we used the information available to score the curriculum indicator: (0) no curriculum, (1) curriculum, and (2) curriculum and related training.

Classroom environment quality. The global quality of the classroom environment was measured by ECERS-R total score, and the quality component rating for global environmental quality was created using the guidelines of the ECERS-R developers. These criteria are based on the classroom total score on the ECERS-R: (0) score less than 3, (1) score greater or equal to 3 or less than 5, (2) score of 5 or greater.

Teacher-child interactions. The quality of the interactions between the teachers and children in the classroom were assessed using the CLASS, and the quality component rating for Teacher-Child Interactions was created using criteria developed for the Head Start Designation Renewal System (DRS; Office of Head Start, 2011). The DRS provided one of the few widely used rankings of CLASS scores. Each of the three CLASS domain scores was rated, and then a composite was created from the three ratings. The Emotional Support domain was scored as: (0) less than 4, (1) 4–5, and (2) greater than 5. The Instructional Support was scored as: (0) less than 2, (1) 2–3, and (2) greater than 3. Classroom Organization was scored as (0) less than 3, (1) 3–4, and (2) greater than 4. The lowest domain rating was identified and became the overall Teacher-Child Interactions rating.

Overall QRIS ratings. The quality component ratings were combined to create QRIS scores. This process involved several steps. First, most of the quality components were collected at the classroom level, so they were combined to create center-level ratings. A within-center mean was computed for each of the classroom quality component ratings. Second, these

mean ratings were rounded so the center-level ratings were constrained to have values of 0, 1, or 2. Third, we combined the center-level ratings to form QRIS scores. The mean of the center-level quality ratings for teacher education, director education, ratio, group size, and curriculum was computed to form a QRIS rating of structural quality for that center. The mean of teacher education, director education, ratio, group size, curriculum, ECERS, and CLASS was computed to form QRIS rating of structural and process quality for that center.

Support for all children and their families. We also looked at three additional quality components that we thought measured another quality dimension- family engagement and support for all children and their families. Across the six studies, we had information about the quality of family involvement, quality of support for dual language learners and their families, and quality of support for children with special needs. There are not professional standards for these quality components, and we used standards in existing QRIS (Build Initiative and Child Trends, 2015). As for the QRIS composite, we assumed a rating of 0 meant low quality, 1 meant moderate quality, and 2 meant high quality.

Quality of *family involvement* was measured, and a component rating was created using the follow criteria. At level 0, programs had to have a handbook for families, offer an orientation to the program, and meet with the family at least once a year. Level 1 added a requirement for two regularly scheduled conferences a year; level 2 added a requirement for written notes from parent conferences, individual reports on the child's development, and high parent satisfaction ratings on parent surveys.

Quality of support for *dual language learners* was also measured. No requirement for specialization related to dual language learners was required at level 0. At level 1, programs had

to use one or more of the children's home languages in the classroom, and at level 2 they also had to implement a culturally sensitive curriculum.

Quality of support for *children with special needs* was examined. Programs did not have to demonstrate any practices specific to children with special needs at level 0. At level 1, programs had to use a published screening or child assessment tool for children with special needs. To receive a score of 2, programs had to use their assessment/screening tool to guide planning for learning activities and/or identify activities for individual children with special needs to do at home.

Preschool child outcomes. All studies administered individualized direct assessments of children's language and academic skills with widely used assessment measures and asked teachers to rate children's social-emotional skills with questionnaires. Multiple studies relied on the same measures of language and academic skills, but each used a different measure of social-emotional skills.

Literacy. Four of the studies (FACES 2006, FACES 2009, NC Pre-K cohort 3, and GA Pre-K) administered the Letter-Word Identification subtest measure of pre- and emerging reading from the Woodcock-Johnson Test of Achievement III (Woodcock, McGrew, & Mather, 2001). The Letter-Word subtest measures letter and word identification skills. The child is initially asked to identify letters. Further items require the child to read and pronounce written words correctly. The Woodcock-Johnson III internal consistency coefficients for the 3- to 5-year-old age group range from .97 to .99 for the Letter-Word Subtest, according to the measure's authors. The test has internal consistency reliability coefficients of .92 for 3-year-olds and .91 for 4-year olds. In the NC Pre-K, GA Pre-K, FACES 2006 and 2009 studies, a subset of children were also assessed in Spanish, using comparable subscales on the Batería III Woodcock-Muñoz

(Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005). For these studies, the higher of the standardized scores in Spanish or English was used. The internal consistency coefficients for the Woodcock-Muñoz 3- to 5-year-old age group range from .84 to .98 for the Letter-Word Subtest, according to the measure's authors.

The NCRECE study administered the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007). The Print Knowledge subscale measures literacy skills in alphabet knowledge and written language. Internal consistency reliabilities are .85 or higher, and inter-rater reliabilities are .96 or higher. Finally, ECLS-B administered an early literacy assessment (alpha = .92) that consisted of 74 items measuring early literacy and language skills, including letter knowledge, word recognition, print conventions, and phonological awareness (Najarian, Snow, Lennon, Kinsey, & Mulligan, 2010).

Language skills. Most studies included a measure of receptive vocabulary. The Peabody Picture Vocabulary Test, 3rd edition was administered in the NCRECE and NC Pre-K studies (PPVT; Dunn & Dunn, 1997), and the Peabody Picture Vocabulary Test, 4th edition (PPVT; Dunn & Dunn, 2007) was administered in the FACES 2006 and FACES 2009 studies. In the PPVT assessment, children are shown a set of four pictures and asked to select the picture that best represents the meaning of a word spoken by the examiner. Internal consistency reliability tends to be high, ranging from .92 to .98. For the NC Pre-K cohorts 2 and 3, and for FACES 2006 and 2009 studies, all children were given the test in English. The Spanish language assessment of receptive vocabulary was the Test de Vocabulario en Imagenes Peabody (TVIP; Dunn, Lugo, & Dunn, 1997). Based on the PPVT-R (an earlier version of the PPVT), TVIP contains 125 translated items to assess the vocabulary of Spanish-speaking and bilingual

students. For this study, we used the higher of the standardized PPVT or TVIP scores as the child's language score.

GA Pre-K used the Woodcock-Johnson III Picture Vocabulary subtest, which measures oral language development and lexical (word) knowledge. The task requires the child to identify pictured objects. Although a few receptive items are offered at the beginning of the test, this is primarily an expressive language task at the single-word level. Internal consistency coefficients for the Picture Vocabulary subtest for the 3- to 5-year-old age group range from .80–.89. A subset of children in the GA Pre-K study was also assessed in Spanish using a comparable subscale on the Batería III Woodcock-Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005). For this study, we used the higher of the standardized scores in Spanish or English. The internal consistency coefficients for the Woodcock-Muñoz 3- to 5-year-old age group range from .88 to .93 for the Picture Vocabulary Subtest, according to the measure's authors.

Mathematics. Four of the studies (FACES 2006, FACES 2009, NC Pre-K, and GA Pre-K) administered the math skills measure from the Woodcock-Johnson Tests of Achievement III (Woodcock, McGrew, & Mather, 2001)—the Applied Problems Subtest. This subtest examines the child's ability to analyze and solve math problems. The internal consistency coefficients for the 3- to 5-year-old age group range from 92 to .94, according to the measure's authors; for 4-year-olds, the internal consistency reliability coefficients are .91. In the NC Pre-K cohorts 2 and 3, GA Pre-K, FACES 2006 and 2009 studies, based on a screening procedure, a subset of children were assessed in Spanish using comparable subscales on the Batería III Woodcock-Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005). For these studies, either the standardized score in Spanish or English was used. The internal consistency coefficients for the

Woodcock-Muñoz 3- to 5-year-old age group range from .90 to .98 for the Applied Problems Subtest, according to the measure's authors. The mathematics assessment scale for the ECLS-B study included items from ECLS-K math assessments from the Woodcock-Johnson and Peabody Individual Achievement test (PIAT), with items added from the Test of Early Mathematics Ability – Third Edition (TEMA-3). The math assessment ($\alpha = .92$) consists of 58 items focused on number sense, properties, operations, and probability. The present secondary data analyses utilized the item response theory (IRT) scores calculated by the ECLS-B for these assessments (Najarian, Snow, Lennon, Kinsey, & Mulligan, 2010).

Social-emotional adjustment. Various measures designed to assess social skills and behavior problems were also administered across the different studies. In fall and spring in NC Pre-K, teachers completed the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) at prekindergarten (pre-K). The Social Skills Questionnaire from the SSRS is composed of 38 items describing child behavior, each rated on a 3-point scale reflecting how often the child exhibited each behavior. The total score reflects levels of perceived social competence, with internal consistency of .90, test-retest reliability of .75 to .88, and moderate concurrent and predictive validity to other indices of social competence. Teachers in FACES 2006 and 2009 completed a 12-item adaptation of the SSRS measuring social skills (Malone et al., 2013); alphas ranged from .88 to .89. Teachers in GA Pre-K completed a revised version of the SSRS, the Social Skills Improvement System Rating Scales (SSIS-RS; Gresham & Elliott, 2008). The Social Skills score from the SSIS-RS is composed of 46 items, with test-retest reliability of .81 to .84. For the ECLS-B, social skills and problem behaviors were assessed through parent reports, using select items from the SSRS and the Preschool and Kindergarten Behavior Scales – Second

Edition (PKBS-2; Merrell, 2003). We conducted factor analyses to create both positive social-emotional and problem behavior composite scores.

Analytic Plan

We conducted all analyses using the quality measures in both their original form (e.g., continuous scores on the ECERS) and after they were converted to ratings based on the QRIS matrix. We conducted three sets of analyses: (1) process quality predicted by structural quality as both continuous quality scores and as a converted rating, (2) child outcomes predicted by structural and process quality as both continuous quality scores and as a converted rating, and (3) process quality predicted by our QRIS total rating based only on structural quality indicators, and child outcomes predicted by our QRIS total rating based on both structural and process quality indicators.

To guide item selection and scoring of a QRIS rating, we first selected quality indicators and used existing evidence to identify potential cut-points and rating criteria to divide the continuous quality ratings into discrete ratings. We examined the psychometric properties and dimensionality of the quality indicators by conducting a reliability analysis and an exploratory factor analysis. Next, to look at the validity of the structural component of the rating, we examined the association between the ratings and children's experiences in early care and education or process quality. Finally, we examined the association between the structural and process quality indicators and overall ratings to children's outcomes to assess whether children learned more in classrooms rated as having higher quality.

We fit two- or three-level mixed models. All analyses accounted for nesting of children in classrooms at the first level by estimating random classroom intercepts. Analyses either accounted for nesting of classrooms in centers through estimating random intercepts at the center

level or through accounting for clustering using sandwich estimators. We conducted value-added models that included fall pre-test child outcome scores as a covariate in all studies but the ECLS-B. For that study, the 24-month Bayley Mental Development scores were included as the pre-test in analyses of the preschool literacy and math outcomes and the 24-month behavior scores, composed of 7 items from the Bayley Behavior Rating Scales, were included as the pre-test in analyses of the social skills outcomes (BSID-II; Bayley, 1993). Additional covariates in the child outcomes models included the following: child gender, child race, maternal education, family income, home language, whether the child had an individualized education plan (IEP), and treatment conditions for the one study with treatment (NCRECE). We applied sample weights to the ECLS-B and FACES data.

We standardized all continuous variables—predictors and outcomes—to have a within-sample mean of 0 and a standard deviation of 1, so regression coefficients can be interpreted as effect sizes. We accounted for missing data by imputing 40 data sets using a Markov chain Monte Carlo method (MCMC; Schafer, 1997), conducting analyses within each imputed data set, and combining results across imputed data sets. We conducted analyses of the data from each study, and combined the coefficients using meta-analysis. In the meta-analysis, we computed the weighted mean coefficient and standard error based on the coefficients, standard errors, and sample sizes within each study.

Results

First, we described the families, children, and centers in the studies, and then we explored the psychometric properties of the quality ratings through examining the dimensionality, validating the selected quality components, and then creating and validating the QRIS rating. The descriptive statistics on the children and families are reported in Table 1 and on the ECE

centers/programs in Table 2. All descriptive statistics in the first two tables are unweighted so that we can show means and standard deviations in a similar manner across all studies. Most of the studies focused on programs that serve families with lower incomes, and this is reflected in Table 1 in the proportion of parents with high school or less education and the mean family incomes. Most of the studies included ECE programs with higher standards (Head Start and state Pre-Ks), and, as shown in Table 2, the mean quality indicator scores across studies were mostly in the moderate to high range, with programs scoring either a 1 or 2 on all quality indicators out of a possible range of 0 to 2. The highest quality indicator score across studies was director education, with most programs meeting the highest quality level (and thereby the highest rating score) for this indicator. The lowest two quality indicators were related to process quality, although most programs still received a moderate quality score of 1.

Psychometric Properties of the Rating Scale

To address the psychometric properties of the scale, we first calculated a measure of reliability among all of the selected seven items within our QRIS rating designed to predict child outcomes (see Table 3). A set of items with a Chronbach's alpha of .7 or above typically is considered to have high internal consistency. The alphas in each study suggest that the indicators are not highly correlated and probably not measuring a single dimension, a suggestion seen even before turning the continuous variables into ratings. Adding the process quality ratings to the structural quality ratings increased the alphas in all studies. We also explored the addition of other QRIS indicators that stakeholders commonly request for inclusion in QRIS ratings, such as family engagement, special needs, and cultural and linguistic diversity. At least when using a simplified version of these indicators, inclusion of these factors increased the alphas only slightly in some studies and substantially reduced them in others.

We also conducted an exploratory factor analysis with varimax rotation to examine the dimensionality of the QRIS ratings across studies (see Table 4). We considered quality indicators with a loading of .30 and above to be related to one factor, and used an eigenvalue of 1 to identify the appropriate number of factors across studies. The studies varied between two and four factors within their quality ratings.

Although the factor structure varied somewhat across studies, there was rather consistently a factor that included teacher education, director education, and curriculum; a factor that included ratio and group size; and a factor that included the observed classroom quality measures. In about half of the studies, the observed classroom measures loaded on either the first or second factor. All studies that included information related to the exploratory factors—family involvement, dual language learners, and special needs—resulted in a factor structure in which these items formed a separate factor, indicating low levels of associations with the other indicators of quality.

In summary, these 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 and are likely to dilute associations with outcomes that might be seen with individual quality variables or unidimensional scales.

Using Process Quality to Validate Structural Quality Indicators

The next set of analyses examined the validation of selected structural quality variables and their scoring based on professional guidelines using observed classroom quality. The first set of analyses asked whether the selected quality variables were significant predictors of ECERS-R and CLASS scores, and results are shown on the left side of Table 5.

The first set of analyses examined the individual structural quality indicators as predictors of process classroom quality. Table 5 shows the partial association between continuous measures of structural quality and classroom process quality, aggregated to the center level. The left side of the table shows the associations between each structural indicator and process quality, with both treated as continuous variables. The right side of the table shows the association between each structural indicator after it was categorized into a rating, (i.e., 0, 1 or 2) and the process quality measure as a continuous variable. Analyses adjusted for site, treatment, and basic sample characteristics of centers, and applied sample weights where appropriate.

The left side of Table 5 shows the regression coefficients and standard errors from the analysis of data from each project and from the meta-analysis that combined the coefficients across projects. As shown, teacher and director education were both significant predictors of three to four of the process quality indicators in the meta-analyses. Teacher education, but not director education, was a significant predictor of CLASS Emotional Support. Lower child:adult ratios were significant predictors of higher ECERS and CLASS Emotional Support scores, but group size was not a significant predictor of process quality in the meta-analyses. Significant effect sizes for associations between structural quality variables and the four process quality measures ranged from small (.13) to moderate (.20) when Cohen's standards for partial correlations were applied to the meta-analysis results.

The right side of Table 5 shows the results from analyses of the structural quality indicators after they were categorized into ratings. The coefficients in these analyses provide the expected change in process quality, with a rating one point higher on this three-point scale. Teacher education, director education, and child:adult ratios were significant predictors of at

least two of the four process quality measures. Curriculum was a significant predictor of CLASS Instructional Support only, and group size was not reliably related to any quality measure.

Translating the structural quality measures into QRIS indicators using the APHA guidelines resulted in a strengthened association between teacher and director education with the ECERS. This pattern also held true for child:adult ratio across all process quality measures, perhaps because child age was taken into account and smaller child:adult ratios might be differentially related to process quality for younger versus older children. In general, contrary to expectations, we did not see stronger associations with the continuous structural quality variables than with the categorized quality indicators. Additionally, the magnitude of the association with classroom process quality across indicators was strengthened when they were combined into an overall quality rating.

Table 5 also shows the results from analyses of the pseudo-QRIS rating we constructed based on the structural quality indicators that have been shown to have the strongest association with process quality and child outcomes. We predicted the four process quality measures from this QRIS rating. The meta-analyses suggested that the QRIS rating was significantly related to all process quality measures, with large effect sizes ($.53 < d < .65$).

In summary, these analyses suggested that four of the five selected structural quality variables and the QRIS rating based on structural indictors appeared to be validated through showing reliable associations with process measures of ECE quality and that the use of the professional guidelines maintained, and sometimes enhanced, these associations.

Using Child Outcomes to Validate the Structural and Process Indicators

The next set of analyses examined associations between child outcomes and both the structural and process quality indicators**.** Two-level HLM analyses were conducted using the

data from each project. The child's spring scores were predicted from the quality indicator, accounting for the child's fall score, family and child covariates, and the nesting of children in classrooms. Results from these analyses were combined using meta-analysis. Results are shown in Table 6. As in Table 5, the left side of Table 6 shows the results of the structural quality indicators and process quality scores as continuous variables; the right side of the table shows the results of the structural quality indicators and process quality scores as categorized quality ratings.

The results from the meta-analyses using each structural quality measure as a continuous and as a scored categorized rating were examined and contrasted. The meta-analyses indicated that a higher level of teacher education was a significant predictor of higher levels of language, literacy, and math skills as a continuous variable, and of literacy skills as a categorized rating. A higher level of director education was significantly related to higher levels of language, literacy, and math skills as both a continuous variable and a categorized rating. Larger child:adult ratio predicted lower social skills as a continuous variable. Higher ratings (indicating smaller group sizes) on the categorical indicator were negatively related to math skills. Having a curriculum was related to higher levels of children's social skills. Effect sizes were small ($.03 < d < .07$), even when reliably different from zero.

The meta-analyses also examined the extent to which the process quality measures predicted child outcomes. The ECERS-R total score was not reliably related to any of the child outcomes. Higher scores on two of the CLASS scores (Instructional Support and Classroom Organization) were significantly related to higher levels of literacy skills when the CLASS scores were examined as continuous variables. A single categorical rating combining all three

CLASS scores was also significantly related to higher literacy and language skills. Again, effect sizes were quite small, even when statistically significant ($.03 < d < .07$).

In summary, these analyses provided some, albeit small and less consistent, evidence of validation for the individual quality variables, and suggested that associations continued to be observed when the continuous variables were categorized using the professional guidelines.

Using Child Outcomes to Validate the QRIS Ratings

The final set of analyses predicted the child outcomes from the QRIS rating, both with and without the process quality components. The results are shown in Table 7. A higher QRIS rating based only on the structural indicators significantly predicted higher language, literacy, and math skills, whereas a higher QRIS rating based on both structural and process quality indicators significantly predicted higher language and literacy (but not math) skills. Again, effect sizes were quite small, even when statistically significant ($.06 < d < .07$).

QRIS in All Programs and in Programs with Performance Standards

We conducted a final set of analyses to examine the generalizability of the combined results across studies for various types of ECE programs. We compared results from the preschool centers included in a nationally representative study (ECLS-B preschool follow-up) with results from studies of Head Start programs that adhered to Head Start program standards, which truncated the low end of the range of quality observed in these programs (FACES 2006 and 2009). We conducted the same set of analyses described above separately to relate the QRIS standards and overall ratings to observed classroom quality and gains in child outcomes. We used the meta-analysis program to combine findings across the FACES 2006 and 2009 samples, and contrast findings from the combined FACES sample with those from the ECLS-B sample.

Results indicated almost no evidence that QRIS quality indicators and simulated QRIS scores related differently to observed quality and child outcomes in nationally representative studies of Head Start and of all children in the U.S. in child care. Despite large differences in the characteristics of the families of the children in the FACES and ECLS-B studies (see Table 1), and in program requirements, with the Heart Performance Standards setting the requirements for the programs in FACES, and varying licensing requirements across states setting the requirements for the programs in the ECLS-B, only two pair-wise comparisons showed different patterns of association in the 31 comparisons conducted (tables available upon request). Comparing the ECLS-B to the FACES findings, teacher education was a stronger predictor of ECERS scores in FACES than ECLS-B, whereas smaller group size quality indicator scores were a stronger negative predictor of social skills in ECLS-B than FACES. The lack of consistent differences in associations across the two sets of studies occurred despite less variability in FACES director education, group size, and ECERS scores when compared to ECLS-B.

Discussion

Among other goals, QRIS are designed to provide easily accessible information about the quality of early care and education programs and an incentive for programs to provide higher quality. Despite efforts to design QRIS ratings using structural and process quality measures that have been shown to be associated with children's learning, little empirical evidence supports the hypothesized association between aggregated QRIS ratings and children's learning. This study contributes to that literature by illustrating the application of psychometric principles of scale development in our simulated QRIS rating scale. We provide evidence supporting the validation of the selected quality indicators and demonstrate that the rating scale describing classroom

quality based on structural quality measures is related to observed classroom process quality, and the rating scale based on both structural and process classroom quality measures is related to gains in child outcomes, even if those gains are small.

Psychometric Properties

We attempted to construct our QRIS ratings using the various psychometric properties discussed above: dimensionality, item selection, and item scoring. We constructed QRIS ratings in this study by first reviewing the research literature and professional recommendations related to quality indicators. We chose to include those ECE quality components assumed to be related to a single dimension of ECE quality—classroom process quality. Analyses examining dimensionality suggested that even the structural measures of classroom quality were multidimensional and thus were only modestly related to each other. The selected structural and process classroom quality measures had average correlations, as indicated by Cohen's alpha, of .28 to .55 across the six studies, thus providing minimal evidence that they can be combined meaningfully into a single scale. Furthermore, the factor analyses indicated that the indices of other dimensions of ECE quality, such as parent engagement, practices supportive of children with special needs, and cultural and linguistic sensitivity, defined independent dimensions of ECE quality. Thus, results of these analyses suggested that multiple QRIS ratings would be needed to adequately capture the various dimensions of quality. This finding is especially important because the consequence of combining all of the various quality indicators into a single rating would be to weaken the rating scale's ability to predict the intended outcomes.

The creation of our QRIS ratings also considered item selection and item scoring. We selected structural quality measures that have been shown in the literature to be related to observed classroom quality—focusing on structural quality measures that showed significant

modest to moderate associations to observed process quality across multiple studies. The findings suggest that all but one of the selected indirect or structural classroom quality measures—teacher and director education, child:adult ratio, and curriculum —were related to observed process quality, and/or to gains in child outcomes. Group size, in contrast, was not related in the anticipated direction in any analysis, raising questions about whether it should be included in the QRIS. The magnitude of those associations of the other quality variables was similar when the structural quality variables were examined as both continuous variables and a scored categorized rating, suggesting little loss of information when the variables were scored using the professional guidelines. In a few cases, the associations were actually stronger when we used the categorized quality indicators, perhaps because they took the age of the children in the classroom into account when scoring structural variables such as child:adult ratios. Furthermore, findings suggested most of the selected direct or process classroom quality measures were related to the child outcomes. We believe these findings support combining careful selection of items through consideration of the strength of the evidence in the ECE literature and careful scoring of those items based on professional recommendations developed from the ECE literature.

Perhaps most importantly, these findings indicated that our hypothetical quality rating was related to gains in child outcomes, providing evidence for validating this rating scale. It was interesting that the QRIS rating based on the structural measures was related to gains on three of the child outcomes, whereas the rating based on the structural and process measures was related to gains on only two of the child outcomes. This finding is difficult to understand based on a logic model that posits that structural measures have their impact on child outcomes through

impact on process quality, but it does reflect the very limited associations between both the ECERS and CLASS and gains in child outcomes in these studies.

We did not examine reliability directly in this study, but did use the quality measures from studies in which the data collectors received training and had to meet the developer's standards prior to gathering the data. Future work on this issue is warranted to examine issues such as the following: (1) variability in ratings of process quality across classrooms and across time within classrooms; (2) variability in observed group sizes and ratios, and the relative contribution of using observed rather than enrollment data.

Finally, analyses across auspice suggested that the validation of the hypothesized QRIS rating was not negatively affected when applied in an auspice with higher standards than when applied to a wide variety of ECE settings. The final set of analyses compared the validation of the ratings we created in a representative sample of preschool children in the ECLS-B and in a representative sample of children in Head Start in FACES. No evidence emerged suggesting the QRIS ratings were less predictive of both observed quality and gains in child outcomes in a study of a program with higher performance standards than in a study representing all types of ECE in the US. The higher standards in Head Start appeared to reduce variability in at least some of the quality measures, but not the associations with the desired outcomes. These findings are reassuring given the inclusion of programs like Head Start and public pre-kindergarten in state QRIS.

Implications

In summary, this replication and meta-analysis provides evidence that individual structural measures of quality commonly used in Quality Rating and Improvement Systems may represent a meaningful assessment of the process quality of early care and education programs, if

careful consideration is given to the strength of the evidence behind the quality measures selected and the psychometric properties of scale development. In addition, individual structural and process quality measures are related to small gains in children's learning and well-being. As with any good scale, results indicated that translating quality measures into QRIS indicator scores and aggregate ratings created a classroom-level quality rating that was more highly associated with children's learning and development than individual indicators.

As states continue to engage in the process of developing and refining their Quality Rating and Improvement Systems, three implications of this study are worth considering. First, the structure of quality rating systems should mirror our theory of change and the overall level of evidence that supports that theory. Second, it is crucial to compute separate scales for different quality dimensions if those dimensions are not highly correlated. Third, it is important to select the quality indicators of each quality dimension based on the strength and consistency of the research evidence. Fourth, the scoring of those indicators should also be based on the research evidence. Although it is important to have a theory of change for each indicator, it is equally important to differentiate between quality indicators with strong and weak (or nonexistent) evidence, and feature those indicators with good evidence and form separate ratings for each of the different ECE quality dimensions. We want evidence that combining these indicators results in a prediction at least as good as when separate indicators are considered. In summary, when careful attention is paid to the psychometric properties of the rating scale, the QRIS approach can describe the information about program quality in a manner that is related to gains in child outcomes; we anticipate that QRIS ratings of other dimensions that are developed using similar psychometric processes would be able to predict additional important ECE outcomes, such as parent engagement and community involvement or systems development.

Limitations

A number of important limitations to this study exist. First, these were secondary data analyses of preschool data collected for other purposes (i.e., not for QRIS validation or evaluation), so the findings need to be interpreted accordingly. In particular, our analyses only looked at preschool early care and education—typically for children in their last year of preschool. We only have data on indicators like teacher education and ratios for the classrooms included in the study, not all classrooms as would be collected in a QRIS. Thus, generalizing findings to typical programs in the United States that serve 0- to 4-year-olds is not warranted.

Second, causal inferences cannot be drawn from our analyses. We attempted to address issues of selection bias in analyses of child outcomes by including a host of covariates, including the child's baseline scores in assessments of their development in multiple developmental domains. Nonetheless, we cannot conclude attending a higher quality center led to gains in academic skills because we saw associations between gains in outcomes and our quality indicators or the overall QRIS ratings.

Third, the studies included in this work come primarily from evaluations or descriptive studies of programs with standards that tend to reflect higher levels of quality (e.g., Head Start, state pre-K), so it is possible that studies using the full range of programs might yield different, and likely stronger, associations. However, our comparison of findings from the ECLS-B and FACES analyses yielded few differences.

Summary

Findings from this study indicate that careful attention to the dimensionality, item selection, and item scoring in QRIS ratings can improve their measurement of ECE quality. The ECE quality indicators and overall QRIS ratings developed through attending to these

psychometric principle were validated through statistically significant associations with gains in academic outcomes in this meta-analysis of the preschool data in six large child care studies.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

American Academy of Pediatrics (2011). *Caring for our children: National health and safety performance standards; Guidelines for early care and education programs. 3rd Edition.* Elk Grove Village, IL:Author. Available Online: http://nrckids.org.

Bayley, N. (1993). *Bayley Scales of Infant Development – second edition*. San Antonio, TX: Psychological Corporation.

Bryant, D. (2010). *Observational measures of quality in center-based early care and education programs*, OPRE Research-to-Policy, Research-to-Practice Brief OPRE 2011-10c. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Build Initiative and Child Trends. (2015). *A Catalog and Comparison of Quality Rating and Improvement Systems (QRIS)* [Data System]. Retrieved from http://qriscompendium.org/

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*. Washington, DC: Brooks Publishing.

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early childcare and education. In M. H. Bornstein & T. Leventhal (Eds.) *Handbook of child psychology and developmental science* (7th ed.,Vol. 4, pp.223-267). Hoboken, NJ: Wiley.

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., & Elliott, S. N. (2008). *Social skills improvement system (SSIS) rating scales*. Bloomington, MN: Pearson Assessments.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale, revised edition*. New York, NY: Teachers College Press.

Helburn, S.W. (1995). *Cost, quality, and child outcomes in child care centers. Technical report.* Denver: Department of Economics, Center for Research in Economics and Social Policy, University of Colorado at Denver.

Hestenes, L. L., Kintner-Duffy, V., Wang, Y. C., La Paro, K., Mims, S. U., Crosby, D., et al. (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

Jeon, L., Buettner, C.K., & Hur, E. (2014). Examining pre-school classroom quality in a statewide Quality Rating and Improvement System. *Child Youth Care Forum, 43*, 469–487. DOI 10.1007/s10566-014-9248-z

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*, 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.

Lugo-Gil, J., Sattar, S., Ross, C., Boller, K., Kirby, G., & Tout, K. (2011). *The Quality Rating and Improvement System (QRIS) evaluation toolkit. OPRE Report# 2011-31*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Malone, L., Carlson, B. L., Aikens, N., Moiduddin, E., Klein, A. K., West, J., et al. (2013, April). *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.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732-749. DOI: 10.1111/j.1467-8624.2008.01154.x

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.

NICHD Early Child Care Research Network. (2002). Child-care structure→Process→Outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science, 13*, 199–206. DOI: 10.1111/1467-9280.00438

NICHD Early Child Care Research Network. (2003). Does quality of child care affect child outcomes at age 4 1/2? *Developmental Psychology, 39*, 451–469.

Office of Head Start, (2011). *Report to Congress on the Proposed Head Start Program Designation Renewal System*. Washington, DC: Administration for Children and Families, U.S. Department of Health and Human Services. http://eclkc.ohs.acf.hhs.gov/hslc/data/rc/Head_Start_Proposed_Designation_Renewal_System.pdf

Peisner-Feinberg, E. S. (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 of NC Pre-K Evaluation Findings 2005–2014.pdf

Peisner-Feinberg, E. S. & Schaaf, J.M. (2007). *Evaluation of the North Carolina More at Four Pre-kindergarten Program: Children's Outcomes and Program Quality in the Fifth Year*. Chapel Hill, NC: FPG Child Development Institute. Downloaded on January 21, 2015 from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/MAF_Yr5_exec_summary.pdf

Peisner-Feinberg, E. S. & Schaaf, J.M. (2008). *Evaluation of the North Carolina More at Four Pre-kindergarten Program: Performance and Progress in the Seventh Year (2007-2008).* Chapel Hill, NC: FPG Child Development Institute. Downloaded on January 21, 2015 from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/MAF_Yr7_full_report.pdf

Peisner-Feinberg, E. S., Schaaf, J. M., & LaForett, D. R. (2013). *Children's growth and classroom experiences in Georgia's Pre-K Program: Findings from the 2011–2012 evaluation study.* Chapel Hill: The University of North Carolina, FPG Child Development Institute. Downloaded on January 21, 2015 from http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval2011-2012Report.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-4-2014.pdf

Phillips, D., Mekos, D., Scarr, S., McCartney, K., & Abbott-Shim, M. (2000). Within and beyond the classroom door: Assessing quality in child care centers. *Early Childhood Research Quarterly, 15*, 475–496. DOI: 10.1016/S0885-2006(01)00077-1

Pianta, R. C., Burchinal, M., Jamil, F. M., Sabol, T., Grimm, K., Hamre, B. K., et al. (2014). A cross-lag analysis of longitudinal associations between preschool teachers' instructional

support identification skills and observed behavior. *Early Childhood Research Quarterly*, *29*, 144–154.

Pianta, R. C., La Paro, K., & Hamre, B. (2004). *Classroom assessment scoring system: Pre-Kindergarten*. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

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

Pianta, R. C., Barnett, W. S., Burchinal, M., & Thornburg, K. R. (2009). The effects of preschool education: What we know, how public policy is or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest, 10*, 49–88. DOI: 10.1177/1529100610381908

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

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

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*. Available online at http://dx.doi.org/10.1016/j.ecresq.2014.05.001

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., Zaslow, M., Halle, T. & Forry, N. (2009). *Issues for the Next Decade of Quality Rating and Improvement Systems. OPRE Issue Brief #3.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Retrieved 1-19-15 from: http://www.acf.hhs.gov/sites/default/files/opre/next_decade.pdf

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *The Child Care Quality Rating System (QRS) assessment: Compendium of quality rating systems and evaluations.* Washington, DC: Child Trends. Retrieved 1-19-15 from http://www.researchconnections.org/content/childcare/federal/inquire-products.html

West, J., Aikens, N., Carlson, B., Meagher, C., Malone, L., Bloomenthal, A., et al. (2010, August). *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.

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.

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

Zaslow, M. & Tout, K. (October, 2014). *Reviewing and clarifying goals, outcomes and levels of implementation: Toward the next generation of Quality Rating and Improvement Systems*

(QRIS). OPRE Research Brief #2014-75. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved 1-19-15 from http://www.acf.hhs.gov/sites/default/files/opre/qris_conceptual_framework.pdf

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.

Zellman, G. L., & Perlman, M. (2008). *Child care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Santa Monica, CA: Rand Corporation.

Table 1. *Descriptive Statistics for Child and Family Characteristics, and Child Assessments.*

		ECLS-B[a] (n = 700)	NCRECE - Year 3 (n = 1210)	FACES 2006 (n = 2710)	FACES 2009 (n = 1986)	GA Pre-K (n = 571)	NC Pre-K[b] (n = 1313)
Background							
Maternal education							
< High School	%	14%	24%	37%	33%	4%	
High School	%	25%	34%	32%	30%	21%	
Some College	%	33%	28%	25%	23%	47%	
College	%	28%	15%	6%	6%	28%	
Family Income/1000	M (SD)	58 (53)	24 (22)	22 (15)	22 (15)		
Income-to-Needs Ratio	M (SD)		1.07 (1.04)	2.73 (1.43)	2.53 (1.35)		
Child gender (male)	%	52%	37%	51%	49%	50%	51%
Child age in months	M (SD)	54.06 (4.19)	56.90 (5.86)	52.80 (6.55)	48.02 (6.56)	56.92 (3.53)	54.36 (3.52)
IEP		10%			7%	4%	4%
4-year-old Cohort	%			39%	45%		
Spanish Assessment	%	0%		23%	25%	6%	12%
Not English-Home Lang	%	19%	16%	30%	26%	27%	23%
Child ethnicity/race							
African American	%	23%	42%	34%	34%	39%	37%
Hispanic	%	19%	35%	38%	35%	14%	24%
White	%	41%	14%	20%	22%	42%	33%
Other	%	17%	9%	8%	8%	5%	6%
Child Outcomes							
Language Fall	M (SD)		85.05 (19.21)	85.04 (14.96)	88.18 (14.41)	99.63 (12.09)	87.38 (17.46)
Spring	M (SD)		89.19 (17.05)	86.84 (13.91)	91.01 (14.09)	100.26 (10.46)	91.17 (16.24)
Literacy Fall	M (SD)	51.35 (0.45)	95.41 (15.11)	93.76 (17.23)	95.51 (18.38)	100.03 (13.52)	93.93 (11.81)
Spring	M (SD)	27.09 (11.17)	102.2 (14.98)	98.08 (16.95)	100.67 (17.24)	102.18 (12.22)	97.19 (11.31)
Math Fall	M (SD)	51.35 (0.45)		88.83 (17.15)	89.38 (14.63)	100.96 (13.50)	94.86 (13.73)
Spring	M (SD)	30.91 (9.99)		89.51 (14.63)	90.02 (15.09)	103.70 (12.03)	98.43 (11.57)
Social Skills Fall	M (SD)	3.95 (0.03)		15.55 (4.73)	15.35 (4.68)	97.02 (14.97)	100.71 (15.64)
Spring	M (SD)	3.86 (.56)		17.50 (4.58)	17.46 (4.41)	100.19 (14.67)	108.84 (14.91)

[a.] ECLS-B "fall" tests are the 24 months Bayley Mental Developmental Index for cognitive outcomes andd Behavior Rating Scale for social outcomes.

b. In NC Pre-K, the the WJLW was administered only in the third cohort (n=321 children).

Table 2. *Descriptive Statistics for Classroom Observational Quality at the Center-Level, and QRIS Ratings.*

		ECLS-B[a] (n = 1450)	**NCRECE**[b] (n = 139)	**FACES 2006**[c] (n = 127)	**FACES 2009**[d] (n =108)	**GA Pre-K**[e] (n = 96)	**NC PRE-K**[f] (n = 158)
Classroom Structural Quality							
Teacher Education	M (SD)	15.00 (1.99)	15.79 (1.63)	15.80 (1.27)	15.84 (1.61)	15.69 (1.19)	15.81 (1.02)
BA+	%	54%	59%	44%	46%	95%	86%
Degree in ECE	%	97%	83%	95%	95%	60%	92%
Director Education	M (SD)	16.00 (2.12)	17.53 (1.41)	18.56 (1.75)	18.78 (2.31)		17.04 (1.93)
BA+	%	68%	99%	94%	89%		84%
Degree in ECE	%	92%	77%	87%	71%		91%
Child-Teacher Ratio	M (SD)	7.15 (2.92)	9.61 (3.73)	7.00 (2.11)	6.71 (1.88)	9.28 (1.57)	6.88 (1.89)
Group Size	M (SD)	14.10 (4.32)	17.05 (3.18)	17.07 (2.36)	17.02 (2.07)	18.88 (3.18)	13.79 (1.99)
Classroom Process Quality							
ECERS-R Total Score	M (SD)	4.53 (1.02)		3.61 (.52)	4.41 (.66)	3.63 (.56)	4.76 (.78)
CLASS Emotional Support	M (SD)		5.34 (.83)		5.36 (.37)	5.52 (.82)	5.79 (.62)
CLASS Instructional Support	M (SD)		2.41 (.69)	1.93 (.43)	2.31 (.58)	2.79 (.78)	2.97 (.86)
CLASS Classroom Organization	M (SD)		5.14 (.78)		4.70 (.48)	5.16 (.75)	5.29 (.65)
QRIS Scores							
QRIS Indicator Ratings							
Teacher Education	M (SD)	1.44 (.65)	1.52 (.49)	1.37 (.44)	1.42 (.42)	1.60 (.49)	1.60 (.53)
0 (no college)	%	9%	1%	3%	0%	0%	2%
1 (CDA/some college)	%	39%	45%	57%	56%	40%	34%
2 (BA/BS & ECE)	%	52%	54%	40%	44%	60%	64%
Director Education	M (SD)	1.58 (.56)	1.77 (.42)	1.82 (.39)	1.64 (.53)		1.76 (.53)
0 (no college)	%	4%	0%	0%	2%		5%
1 (CDA/some college)	%	35%	23%	18%	29%		13%
2 (BA/BS & ECE)	%	62%	77%	82%	69%		82%
Child: Adult Ratio	M (SD)	1.54 (.72)	.75 (.75)	1.35 (.66)	1.54 (.52)	.88 (.74)	1.68 (.52)
0 (> 10:1)	%	13%	30%	0%	6%	34%	3%
1 (10:1 – 8:1)	%	20%	42%	18%	48%	45%	26%
2 (<= 8:1)	%	52%	28%	82%	46%	22%	72%

		ECLS-B[a]	NCRECE[b]	FACES 2006[b]	FACES 2009[b]	GA Pre-K[c]	NC PRE-K[d]
Group Size[e]	M (SD)	1.66 (.60)	1.75 (.45)	.78 (.56)	1.05 (.05)	0.77 (.72)	1.92 (.25)
0 (>20)	%	6%	23%	41%	6%	40%	0%
1 (17-20)	%	22%	53%	44%	82%	43%	6%
2 (<=16)	%	72%	24%	15%	12%	17%	94%
Curriculum	M (SD)	1.67 (.69)	1.89 (.39)	1.96 (.21)	1.81(.36)	.81 (.51)	1.00 (.18)
0 (none)	%	13%	3%	1%	0%	24%	1%
1 (global, no training	%	8%	7%	1%	17%	71%	96%
2 (literacy/ TA for curric)	%	79%	90%	98%	82%	5%	2%
CLASS	M (SD)		1.78 (.85)	.51 (.45)	1.08 (.29)	1.36 (.49)	1.49 (.39)
0 (IS<2,CO<3,ES<4)	%		5%	52%	0%	6%	0%
1 (2<IS<3, 3<CO<5,4<ES<6)	%		64%	44%	89%	51%	41%
2(IS>3, CO>5, ES >6)	%		31%	3%	11%	43%	59%
ECERS (Total)	M (SD)	1.25 (.60)		.89 (.29)	1.18 (.41)	.87 (.36)	1.35 (.50)
0 (1-3)	%	8%		8%	4%	14%	1%
1 (3-5)	%	58%		90%	78%	85%	60%
2 (5-7)	%	34%		2%	17%	1%	38%
QRIS Overall Ratings							
QRIS Structural Quality Rating	M (SD)	1.58 (.32)	1.51 (.29)	1.46 (.28)	1.60(.29)	1.22 (.34)	1.62 (.21)
0	%	0%	0%	0%	0%	0%	0%
1	%	37%	55%	52%	33%	84%	25%
2	%	63%	45%	48%	67%	16%	75%
QRIS Structural & Process Quality	M (SD)	1.52 (.31)	1.52 (.26)	1.37 (.23)	1.50 (.23)	1.18 (.28)	1.53 (.20)
0	%	0%	0%	0%	0%	1%	0%
1	%	33%	45%	87%	47%	86%	36%
2	%	67%	54%	13%	53%	12%	64%

a. In the ECLS-B dataset 1 classroom was observed per center.

b. In the NCRECE, FACES 2006, & FACES 2009 datasets 1-8 classrooms were observed per center and center means are reported as the average score across classrooms.

c. In the Georgia Pre-K dataset 1-2 classrooms were observed per center and center means are reported as the average score across classrooms.

d. In the NC Pre-K dataset 1-6 classrooms were observed per center and center means are reported as the average score across classrooms. CLASS scores and average group size were collected for Cohort 3 only (n=49 centers).

e. Group size cut-offs listed for 4-year-olds

Table 3. Internal consistency: Intercorrelations among continuous measures of quality and QRIS ratings.

	Chronbach's Alpha: Internal Consistency					
	ECLS-B	NCRECE	FACES 2006	FACES 2009	GA Pre-K	NC PRE-K
Structural Quality Ratings[a]	.15	.22	.24	.41	.28	.15
Structural & Process Quality Ratings[b]	.28	.39	.34	.55	.37	.31
Structural &Process Quality, Family Engagement Rating[c]	.33	.26	.38	.44		

[a] Structural quality ratings includes ratings of teacher education, director education, class size, child/teacher ratio, and curriculum

[b] Structural & process quality includes ratings of teacher education, director education, class size, child/teacher ratio, and curriculum process, ECERS-R Total, and the combined CLASS domains

[c] Structural quality, process quality, and family engagement factors include ratings of teacher education, director education, class size, child/teacher ratio, and curriculum process, ECERS-R Total, combined CLASS domains, support for family involvement, support for DLLs, & services for children with special needs.

Table 4. Dimensionality of QRIS ratings.

	ECLS-B	NCRECE	FACES 2006	FACES 2009	GA Pre-K	NC PRE-K
Using selected indicators						
Number of Factors	2	2	3	2	3	3
Factor 1	Teacher Ed, Director Ed, ECERS	Teacher Ed, Director Ed, Curriculum, CLASS	Teacher Ed, Curriculum	Director Ed, Ratio, Group Size	Teacher Ed, Curriculum	Teacher Ed, Director Ed
Factor 2	Ratio, Group Size	Ratio, Group Size	Director Ed, Ratio, Group Size	CLASS, ECERS	Ratio, Group Size	Curriculum, CLASS
Factor 3			CLASS, ECERS		CLASS, ECERS	Ratio, ECERS
Adding other indicators						
Number of factors	3	3	4	3		
Additional factor	DLL, Parent Involvement	DLL	DLL, Family Engagement, Special Needs	DLL, Family Engagement		

Table 5. *Estimated Effect Sizes for Linear Associations between Continuous Structural Measures of Center Quality, QRIS Quality Indicator Ratings by Study and Center Continuous Process Quality.*

	Continuous Structural Variables				Categorical Structural Ratings			
	ECERS Total Score	CLASS ES	CLASS IS	CLASS CO	ECERS Total Score	CLASS ES	CLASS IS	CLASS CO
QRIS Indicators								
Teacher Education								
Meta-analysis	***.14*** (.03)***	***.10+ (.05)***	***.17*** (.05)***	***.14**(.05)***	***.30*** (.05)***	***.06 (.13)***	***.30** (.11)***	***.19 (.12)***
ECLS-B	.23*** (.05)				.38*** (.06)			
NCRECE		.25* (.11)	.35** (.11)	.30**(.11)		.23 (.21)	.46* (.21)	.40* (.20)
FACES 2006	.02 (.09)		.03 (.09)		.11 (.22)		.39+ (.21)	
FACES 2009	.09 (.10)	-.02 (.10)	.29 (.11)*	.04 (.08)	.19 (.28)	-.04 (.32)	.83*(.34)	.10 (.25)
GA Pre-K	.15 (.10)	.09 (.10)	.18+ (.10)	.09 (.10)	.11 (.21)	-.17 (.21)	-.11 (.21)	-.32 (.21)
NC PRE-K	.04 (.07)	.09 (.15)	-.02 (.15)	.29* (.14)	.11 (.13)	.37 (.37)	.14 (.37)	1.05** (.34)
Director Education								
Meta-analysis	***.15*** (.04)***	***.10 (.08)***	***.10 (.06)***	***.19**(.07)***	***.30*** (.06)***	***.30* (.14)***	***.37** (.12)***	***.42** (.13)***
ECLS-B	.18*** (.05)				.33*** (.08)			
NCRECE		.25* (.12)	.23 (.12)	.28* (.11)		.28 (.22)	.22 (.22)	.54** (.21)
FACES 2006	.32** (.11)		.17 (.11)		.64* (.25)		.05 (.25)	
FACES 2009	.13 (.14)	-.14 (.16)	.03(.14)	-.02 (.12)	.82*** (.21)	.46+ (.23)	.90***(.20)	.38* (.19)
NC PRE-K	.04 (.07)	.07 (.15)	-.18 (.15)	.33* (.14)	-.03 (.12)	.08 (.29)	-.03 (.29)	.28 (.28)
Group Size								
Meta-analysis	***.03 (.04)***	***-.04 (.06)***	***-.05 (.05)***	***-.04 (.06)***	***.09 (.06)***	***.14 (.10)***	***.06 (.08)***	***.14 (.10)***
ECLS-B	.12* (.05)				-.06 (.08)			
NCRECE		-.10 (.09)	-.06 (.09)	-.15 (.09)		.18 (.14)	.09 (.14)	.22 (.14)
FACES 2006	-.21* (.09)		-.01 (.09)		.55*** (.15)		.25 (.15)	
FACES 2009	.04 (.11)	-.09 (.12)	-.04 (.10)	-.08 (.11)	.78 (.47)	.44 (.44)	1.44** (.52)	.60 (.36)
GA Pre-K	-.18 (.11)	-.04 (.11)	-.06 (.11)	.01 (.13)	.04 (.12)	.03 (.11)	.13 (.11)	.05 (.13)
NC PRE-K	.28* (.14)	.19 (.15)	-.11 (.15)	.24+ (.14)	-1.61 (.71)	-.69 (.72)	.32 (.72)	-.15 (.72)
Child:Adult Ratio								
Meta-analysis	***-.17*** (.03)***	***-.20**(.05)***	***-.07+ (.04)***	***-.10+ (.06)***	***.26*** (.05)***	***.23** (.06)***	***.21** (.06)***	***.18* (.08)***
ECLS-B	-.25***(.06)				.28*** (.08)			
NCRECE		-.15 (.09)	-.08 (.09)	-.10 (.09)		.19 (.12)	.15 (.13)	.19 (.12)
FACES 2006	-.14+ (.07)		.02 (.07)		.33** (.11)		.16 (.13)	
FACES 2009	-.15(.11)	-.27+(.10)	-.27* (.12)	-.13 (.10)	.25 (.25)	.28 (.29)	.33 (.26)	.34 (.21)
GA Pre-K	-.18+ (.11)	-.09 (.11)	-.11 (.11)	-.01 (.13)	.34* (.14)	.21 (.14)	.28* (.14)	.06 (.15)
NC PRE-K	-.11 (.07)	-.34* (.14)	-.10 (.15)	-.15 (.15)	.03 (.14)	.41+ (.24)	.31 (.25)	.23 (.25)

Table 5 (continued). *Estimated Effect Sizes for Linear Associations between Continuous Structural Measures of Center Quality, QRIS Quality Indicator Ratings by Study and Center Continuous Process Quality.*

Continuous Structural Variables				Categorical Structural Ratings			
ECERS Total Score	CLASS ES	CLASS IS	CLASS CO	ECERS Total Score	CLASS ES	CLASS IS	CLASS CO
Curriculum							
Meta-analysis				***.02 (.07)***	***.17 (.14)***	***.33* (.14)***	***.05 (.14)***
ECLS-B[a]				-.04 (.08)			
NCRECE					-.05 (.24)	.00 (.26)	-.20 (.24)
FACES 2006				-.11 (.39)		.75+ (.38)	
FACES 2009				.53 (.38)	.24 (.46)	.27 (.48)	.11 (.41)
GA Pre-K				.17 (.20)	.31 (.20)	.40* (.20)	.18 (.20)
NC PRE-K				.00 (.37)	.18 (.60)	.49 (.60)	.31 (.60)
QRIS Rating							
Meta-analysis				***.62*** (.11)***	***.59** (.19)***	***.53*** (.14)***	***.64*** (.18)***
ECLS-B				.75*** (.17)			
NCRECE					.65* (.31)	.60 (.32)	.84** (.30)
FACES 2006				.52** (.18)		.30 (.19)	
FACES 2009				1.51**(.49)	.86+ (.49)	2.17*** (.55)	1.04* (.40)
GA Pre-K				.71** (.30)	.36 (.30)	.52+ (.30)	-.00 (.31)
NC PRE-K				-.01 (.32)	.93 (.69)	.72 (.70)	1.49* (.66)

Note: Covariates included site and treatment group if relevant. Sample weights applied in ECLS-B and FACES 2006 and 2009.
+p<.10 *p<.05 **p<.01 ***p<.001.

Table 6. *Estimated Effect Sizes for Linear Associations between Continuous Measures of Center Quality, QRIS Indicator Ratings by Study, and Child Outcomes, Controlling for Child, Family, and Program Characteristics.*

	Continuous Structural Process Variables				Categorical Structural Process Ratings			
	Language	Literacy	Math	Social Skills	Language	Literacy	Math	Social Skills
QRIS Indicators								
Teacher Education								
Meta-analysis	***.03** (.01)***	***.07*** (.01)***	***.04** (.01)***	***-.01 (.01)***	***.02 (.02)***	***.13*** (.03)***	***.04 (.03)***	***-.04 (.04)***
ECLS-B		.06 (.05)	.13* (.06)	-.01 (.08)		.18+ (.09)	.20* (.10)	-.11 (.11)
NCRECE	.04 (.03)	.09 (.05)			.06 (.05)	.14 (.08)		
FACES 2006	.04+ (.02)	.06* (.03)	.07** (.03)	-.04 (.03)	.03 (.05)	.12* (.06)	.12* (.06)	-.10 (.07)
FACES 2009	.02 (.03)	.08** (.02)	-.01 (.02)	0 (.02)	-.02 (.08)	.22** (.07)	-.02(.06)	-.02 (.07)
GA Pre-K	-.01 (.03)	.01 (.03)	.06* (.03)	.02 (.04)	-.05 (.06)	-.04 (.07)	-.00 (.06)	-.05 (.09)
NC PRE-K	.04+ (.02)	.16* (.07)	.12* (.06)	-.00 (.04)	.04 (.04)	.43* (.17)	.00 (.05)	.04 (.08)
Director Education								
Meta-analysis	***.04** (.01)***	***.04** (.01)***	***.05** (.01)***	***.01 (.02)***	***.07** (.03)***	***.08** (.03)***	***.09** (.03)***	***.05 (.04)***
ECLS-B		.14** (.05)	.11* (.05)	-.00 (.08)		.14* (.07)	.09 (.08)	-.05 (.13)
NCRECE	.04 (.02)	.08** (.03)			.04 (.08)	.09 (.11)		
FACES 2006	.08*** (.02)	.01 (.02)	.07** (.02)	.03 (.03)	.24*** (.05)	.10 (.06)	.23*** (.06)	.09 (.07)
FACES 2009	.02 (.03)	.04 (.02)	.02 (.03)	-.04 (.03)	.08 (.08)	.03 (.05)	.00 (.05)	.05 (.05)
GA Pre-K	-.00 (.02)	.10 (.07)	.01 (.03)	.05 (.04)	-.03 (.04)	.12 (.13)	.07 (.05)	.04 (.08)
Group Size								
Meta-analysis	***-.02 (.01)***	***.02 (.02)***	***.02 (.02)***	***-.00 (.02)***	***.04+ (.02)***	***-.00 (.02)***	***-.10** (.03)***	***.00 (.03)***
ECLS-B		.08+ (.05)	.13** (.04)	.08 (.08)		-.17* (.07)	-.18* (.07)	-.21+ (.11)
NCRECE	-.02 (.03)	.00 (.04)			.05 (.04)	-.00 (.06)		
FACES 2006	-.04+ (.02)	.01 (.03)	.02 (.03)	-.01 (.03)	.08* (.04)	.08+ (.05)	.05 (.05)	.01 (.06)
FACES 2009	.02 (.03)	.00 (.03)	-.02 (.03)	-.02 (.02)	-.03 (.09)	-.08 (.08)	.11(.07)	.06 (.06)
GA Pre-K	-.03 (.03)	.00 (.03)	.00 (.03)	.04 (.05)	.01 (.04)	.02 (.04)	-.07 (.06)	.00 (.05)
NC PRE-K	-.00 (.04)	.16* (.06)	.04 (.05)	.09 (.07)	-.29 (.21)	-.44 (.35)	-.25 (.27)	-.02 (.37)

Table 6 (continued). *Estimated Effect Sizes (SE) for Linear Associations between QRIS Indicator Ratings by Study and Child Outcomes, Controlling for Child, Family, and Program Characteristics.*

	Continuous Structural Process Variables				Categorical Structural Process Ratings			
	Language	Literacy	Math	Social Skills	Language	Literacy	Math	Social Skills
Ratio								
Meta-analysis	***-.01 (.01)***	***.00 (.01)***	***.01 (.02)***	***-.04* (.02)***	***.02 (.02)***	***.00 (.02)***	***.00 (.01)***	***.04+ (.03)***
ECLS-B		.02 (.04)	-.00 (.06)	-.05 (.07)		-.01 (.07)	-.02 (.08)	.06 (.10)
NCRECE	-.03 (.03)	.01 (.04)			.05 (.04)	.01 (.05)		
FACES 2006	-.03 (.02)	.01 (.03)	-.02 (.03)	-.05 (.03)	.03 (.03)	.01 (.04)	.01 (.04)	.04 (.04)
FACES 2009	-.02 (.03)	-.01 (.02)	-.02 (.03)	-.07* (.03)	.04 (.07)	.01 (.05)	.04 (.06)	.10+ (.05)
GA Pre-K	-.02 (.03)	-.01 (.04)	.01 (.03)	.05 (.05)	.02 (.04)	.02 (.05)	-.00 (.04)	-.06 (.06)
NC PRE-K	.03 (.02)	.12+ (.07)	.07** (.03)	-.02 (.04)	-.05 (.04)	-.08 (.12)	-.07 (.06)	.09 (.08)
Curriculum								
Meta-analysis					***-.04 (.04)***	***.00 (.04)***	***.04 (.04)***	***.12* (.05)***
ECLS-B						.00 (.06)	.15+ (.08)	.02 (.11)
NCRECE					-.09 (.07)	-.05 (.10)		
FACES 2006					.02 (.13)	-.02 (.15)	.09 (.15)	.17 (.16)
FACES 2009					-.08 (.09)	.00 (.11)	-.08 (.07)	.24** (.09)
GA Pre-K					-.03 (.06)	-.04 (.06)	-.02 (.06)	.03 (.08)
NC PRE-K					.04 (.11)	.98*** (.24)	.45** (.14)	.28 (.22)
ECERS								
Meta-analysis	.01 (.01)	.02+ (.01)	-.00 (.01)	.02 (.02)	.03 (.03)	.01 (.03)	.00 (.03)	.07+ (.04)
ECLS-B		.02 (.04)	.06 (.04)	.04 (.09)		.02 (.06)	.10 (.08)	.15 (.13)
NCRECE	.02 (.02)	.03 (.02)	-.00 (.02)	.01 (.03)	.08 (.07)	.08 (.08)	-.05 (.08)	.10 (.10)
FACES 2006	.03 (.03)	-.01(.03)	-.01 (.02)	.02 (.03)	-.01(.07)	-.14*(.06)	-.05 (.05)	.01 (.07)
FACES 2009	.04 (.03)	.04 (.03)	.02 (.03)	-.02 (.04)	.09 (.08)	.19* (.09)	.09 (.08)	-.04 (.12)
GA Pre-K	-.03 (.02)	.05 (.07)	-.05 (.03)	.11* (.05)	-.01 (.05)	.11 (.18)	-.00 (.07)	.20* (.10)
NC PRE-K								

Table 6 (continued). *Estimated Effect Sizes for Linear Associations between QRIS Indicator Ratings by Study and Child Outcomes, Controlling for Child, Family, and Program Characteristics.*

	Continuous Structural Process Variables				Categorical Structural Process Ratings			
	Language	Literacy	Math	Social Skills	Language	Literacy	Math	Social Skills
CLASS – combined								
Meta-analysis					***.08* (.03)***	***.13** (.04)***	***.06 (.05)***	***.07 (.07)***
NCRECE					.11* (.05)	.20** (.08)		
FACES 2009					.06 (.10)	.04 (.09)	.00 (.11)	.11 (.11)
GA Pre-K					.07 (.06)	.09 (.07)	.07 (.06)	.03 (.09)
NC PRE-K[e]					.04 (.10)	.36* (.18)	.09 (.14)	.14 (.20)
CLASS – Instructional Support								
Meta-analysis	***.03 (.02)***	***.06*** (.01)***	***.01 (.02)***	***.03 (.02)***	***.05* (.02)***	***.07** (.03)***	***.05 (.03)***	***.04 (.04)***
NCRECE	.04 (.03)	.13***(.04)			.06 (.03)	.13* (.05)		
FACES 2006	.03 (.02)	.02 (.03)	.00 (.02)	.02 (.03)	.04 (.05)	-.03 (.06)	-.00 (.06)	-.00 (.07)
FACES 2009	.03(.03)	.05+(.02)	-.01(.04)	.04(.03)	-.07 (.18)	-.07 (.11)	.09 (.11)	.17 (.14)
GA Pre-K	.06* (.03)	.06* (.03)	.03 (.03)	.03 (.04)	.06 (.04)	.08+ (.05)	.06 (.04)	.02 (.06)
NC PRE-K[e]	-.02 (.04)	.09 (.07)	.02 (.05)	.00 (.07)	.05 (.06)	.24* (.10)	.06 (.08)	.13 (.12)
CLASS – Emotional Support								
Meta-analysis	***.02 (.02)***	***.01 (.02)***	***-.02 (.02)***	***.02 (.02)***	***.05+ (.03)***	***.08* (.03)***	***.03 (.04)***	***.04 (.05)***
NCRECE	.04 (.03)	.10*** (.04)			.05 (.05)	.09 (.07)		
FACES 2009	-.02(.03)	-.03(.02)	-.04+(.02)	.02(.03)	.04 (.06)	.09 (.05)	.01 (.07)	.06 (.07)
GA Pre-K	.06* (.03)	.04 (.03)	.03 (.03)	.01 (.04)	.08 (.05)	.06 (.06)	.07 (.05)	.01 (.08)
NC PRE-K[e]	-.04 (.04)	.05 (.07)	-.06 (.05)	-.00 (.07)	-.03 (.08)	-.04 (.15)	-.04 (.11)	.07 (.16)
CLASS – Classroom Organization								
Meta-analysis	***.02 (.02)***	***.05** (.02)***	***-.01 (.02)***	***.02 (.02)***	***.06+ (.03)***	***.07+ (.04)***	***.02 (.04)***	***.02 (.06)***
NCRECE	.05 (.03)	.11** (.04)			.13* (.05)	.17* (.07)		
FACES 2009	-.01(.03)	.00(.03)	-.03(.03)	.03(.03)	.04 (.07)	-.05 (.07)	-.03 (.07)	.03 (.08)
GA Pre-K	.02 (.03)	.05 (.03)	.02 (.03)	.03 (.04)	-.01 (.06)	.05 (.07)	.03 (.06)	.03 (.09)
NC PRE-K[e]	-.01 (.04)	.14* (.07)	.04 (.05)	-.06 (.07)	.03 (.09)	.33* (.16)	.12 (.12)	-.04 (.17)

Note: All models accounted for nesting of children within centers, except the ECLS-B analyses where only one child per classroom/center participated in the study. Covariates included the fall or prior score on the outcome, gender, race/ethnicity, mother's education, family income, home language, and if relevant, language of assessment, IEP status, site and treatment group. Sample weights were applied in ECLS-B and FACES. The pre-test score for all child outcomes models in the ECLS-B child care sample is the 24 month Bayley mental score (*M*=49.11, *SD*=10.71).
+p<.10 *p<.05 **p<.01 ***p<.001.

Table 7. *Estimated Effect Sizes for Linear Associations between QRIS Ratings by Study and Child Outcomes, Controlling for Child, Family, and Program Characteristics.*

	Language	Literacy	Math	Social Skills
QRIS Rating (Structural Indicators)				
Meta-analysis	***.07*** (.02)***	***.06* (.03)***	***.07* (.03)***	***.03 (.03)***
ECLS-B		.13 (.12)	.23 (.18)	-.21 (.23)
NCRECE	.09 (.08)	.09 (.11)		
FACES 2006	.07** (.02)	.05* (.03)	.08** (.03)	.02 (.03)
FACES 2009	.00 (.14)	.22 (.13)	.00 (.10)	.23* (.10)
GA Pre-K	-.01 (.09)	-.00 (.10)	-.02 (.08)	-.11 (.12)
NC PRE-K	.14 (.20)	.51 (.33)	.16 (.25)	.14 (.36)
QRIS Rating (Structural Indicators+ Process Quality)				
Meta-analysis	***.07*** (.02)***	***.07* (.03)***	***.05+ (.03)***	***.03 (.03)***
ECLS-B		.14 (.12)	.30 (.18)	-.08 (.22)
NCRECE	.15 (.09)	.18 (.13)		
FACES 2006	.07** (.02)	.05+ (.03)	.05* (.03)	.02 (.03)
FACES 2009	.07 (.15)	.15 (.15)	.05 (.13)	.27* (.12)
GA Pre-K	.05 (.11)	.09 (.12)	.04 (.11)	-.12 (.16)
NC PRE-K	.08 (.24)	.81* (.38)	.25 (.30)	.30 (.42)

Note: All models accounted for nesting of children within centers, except the ECLS-B analyses where only one child per classroom/center participated in the study. Covariates included the fall or prior score on the outcome, gender, race/ethnicity, mother's education, family income, home language, and if relevant, language of assessment, site and treatment group. Sample weights were applied in ECLS-B and FACES.
+p<.10 *p<.05 **p<.01 ***p<.001.

Appendix A.
Rating Matrix Criteria for QRIS Ratings across and Within Studies.

	Teacher Education	**Child:Adult Ratio**	**Curriculum**	**Director Education**	**Group Size**	**ECERS (Total)**	**CLASS**	**Family Involvement**	**DLL**	**Special Needs**
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 CO < 3 ES < 4	1. Family Handbook 2. Family Orientation 3. Family met with staff at least once/year	none	none
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 = 2-3 CO = 3-5 ES = 4-6	Level 1 plus: 4. Regularly scheduled parent conferences at least twice/year	Use of home language(s).	Use of published screening or assessment tool.
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+ CO = 5+ ES = 6+	Level 2 plus: 5. Program keeps written notes from parent conferences 6. Individual reports on child's development 7. High parent satisfaction	Level 2 plus: Curriculum is culturally sensitive.	Level 2 and use assessment tool to: 1. Guide planning for learning activities. and/or 2. Identify activities for child to do at home.

*Note: In the ECLS-B dataset, the Curriculum Indicator is coded as: (1) no curriculum, (2) curriculum, and (3) curriculum and training.
*Note: In the FACES 2006 dataset, the CLASS Instructional Support domain is used in place of the full CLASS scales.