OPRE Report 2017-92
December 2017

Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies

CCEEPRA
Child Care and Early Education Policy and Research Analysis

INQUIRE

Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies

OPRE Report #2017-92

December 2017

Kathryn Tout, Katherine Magnuson, Shannon Lipscomb, Lynn Karoly, Rebecca Starr, Heather Quick, Diane Early, Dale Epstein, Gail Joseph, Kelly Maxwell, Joanne Roberts, Christopher Swanson, and Jennifer Wenner

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

Contract Number: HHSP23320095631WC

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

Suggested citation: Tout, K., Magnuson, K. Lipscomb, S., Karoly, L, Starr, R., Quick H., ...& Wenner, J. (2017). Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies. OPRE Report #2017-92. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

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

All photos courtesy of Allison Shelley/The Verbatim Agency for American Education: Images of Teachers and Students in Action.

Acknowledgements

This report was produced as part of Child Trends' Child Care and Early Education Policy Research and Analysis (CCEEPRA) contract funded by the Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families, U.S. Department of Health and Human Services. Planning and writing of the report was initiated through the Quality Initiatives Research and Evaluation Consortium (INQUIRE), a workgroup designed to facilitate the exchange of information and resources related to the research and evaluation of quality rating and improvement systems (QRIS) and other quality initiatives. Ivelisse Martinez-Beck is the federal project officer for CCEEPRA and INQUIRE. Her expertise and guidance were greatly appreciated in the production of this report. We also acknowledge helpful reviews from Meryl Barofsky, Emily Schmitt, Mary Bruce Webb and Naomi Goldstein at OPRE.

The authors of this report are INQUIRE members who conducted state validation studies and Child Trends staff, most of whom were also an author of at least one validation study. Author affiliations are listed on the following page. We appreciate the collaboration and time that all authors contributed to this report.

We are especially grateful for the perspectives shared by the authors of the three commentaries included in the report: Martha Zaslow from the Society for Research in Child Development; Katherine McGurk and Erin Gernetzke from the Wisconsin Department of Children and Families, Division of Early Care and Education; and Elizabeth Shuey, a Society for Research in Child Development fellow at OPRE, now at the Organisation for Economic Co-operation and Development (note that Dr. Shuey´s views are her own and do not necessarily reflect those of the Organisation or its Member countries). Their insights help to put the findings in a broader context.

We appreciate the support of Rowan Hilty at Child Trends in producing tables and figures for the report.

Finally, we acknowledge the investment made by the ten states included in the synthesis to conduct high-quality research to inform continuous improvement of their Quality Rating and Improvement Systems. The partnerships developed between the state implementation teams and the research teams were productive and mutually beneficial. The validation studies were ultimately more useful and relevant because of the strong partnerships that developed.

Author Affiliations

Kathryn Tout, Child Trends – Arizona and Minnesota reports

Katherine Magnuson, University of Wisconsin – Wisconsin report

Shannon Lipscomb, Oregon State University – Oregon report

Lynn Karoly, RAND Corporation – Delaware report

Rebecca Starr, Child Trends – Minnesota report

Heather Quick, American Institutes for Research (AIR) – California report

Diane Early, Child Trends – Rhode Island report

Dale Epstein, Child Trends – Arizona report

Gail Joseph, University of Washington – Washington report

Kelly Maxwell, Child Trends – Rhode Island report

Joanne Roberts, Wellesley College – Massachusetts report

Christopher Swanson, Johns Hopkins University – Maryland report

Jennifer Wenner, Child Trends

Table of Contents

Overview

Quality Rating and Improvement Systems (QRIS) are initiatives implemented in states to promote improvement in the quality of early care and education (ECE) programs. Although systems vary in their specific features, QRIS typically include a process for measuring and rating ECE program quality, sharing ratings with parents and the public, and providing supports (including financial incentives) to help programs improve their quality. Because the number of states with a QRIS and the proportion of ECE programs participating in voluntary QRIS have increased in recent years, it is important to learn about whether and how QRIS activities are working to achieve intended goals. QRIS validation studies are one type of QRIS evaluation that examine a set of questions about how well the quality measurement and rating processes are working to differentiate meaningful levels of ECE program quality. Validation studies analyzing how QRIS ratings are associated with measures of quality and preschool children's development were required by states that received Race to the Top – Early Learning Challenge grants.

The purpose of this report is to compile and analyze findings from 10 validation studies examining quality ratings of ECE programs participating in state QRIS. The availability of recent research results addressing similar research questions in 10 different states offers a rare opportunity to synthesize findings across multiple contexts and discuss the implications for design, implementation, and future research on state ECE quality initiatives.

Looking across findings from the 10 studies, QRIS ratings appear to be a helpful tool for state early childhood systems to differentiate programs at lower and higher levels of quality. Overall, QRIS ratings reflect differences in environments, interactions, and activities between ECE programs at higher and lower rating levels. Although statistically significant, the differences in observed quality scores between QRIS rating levels were generally small. Findings for family child care programs had mixed results.

The studies yielded inconsistent evidence of small positive associations between ratings and patterns of children's development. Some selective positive associations were found in some states, but not across all developmental domains examined, nor across all measures within a domain. Three of six studies found evidence that QRIS ratings were associated with some measures of executive function, and four found selective associations between ratings and measures of social-emotional development.

Results documenting observed quality at medium and low levels across many QRIS programs highlight the need for continued investment and innovation in quality improvement supports for ECE programs.

Overall, the report is intended to update state administrators and other stakeholders about the effectiveness of current QRIS quality ratings in distinguishing meaningful levels of quality. The report also addresses issues of interest to researchers conducting evaluations of quality initiatives.

Executive Summary

The purpose of this report is to compile and analyze findings from 10 validation studies examining ratings of early care and education (ECE) programs participating in state Quality Rating and Improvement Systems (QRIS). The availability of recent research results addressing similar questions in 10 different states offers a rare opportunity to synthesize findings across multiple contexts and discuss the implications for design, implementation, and future research on state ECE quality initiatives. The report is intended to update state administrators and other stakeholders about the effectiveness of current QRIS quality ratings in distinguishing meaningful levels of quality. The report also addresses issues of interest to researchers conducting evaluations of quality initiatives.

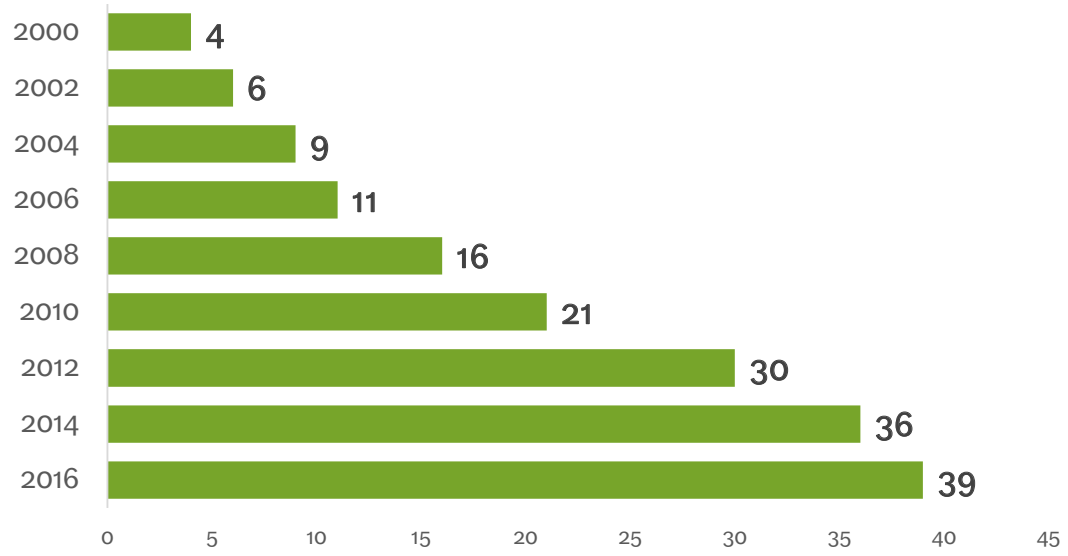
Background on QRIS and the Purpose of Validation Studies

QRIS are initiatives implemented in 39 states[1] to promote improvement in the quality of ECE programs. Although systems vary in their specific features, QRIS typically include a process for measuring and rating ECE program quality, sharing ratings with parents and the public, and providing supports (including financial incentives) to help programs improve their quality.

The first QRIS became operational in 1997. Growth in state QRIS was relatively slow and steady in the early 2000s, until the pace of growth increased in 2008. Rapid growth from 2010 to 2014 was due in part to the Race to the Top – Early Learning Challenge Grants awarded to 20 states, which required a QRIS (ES Figure 1).

[1] Note that California and Florida each have multiple QRIS operating at a county or local level

ES Figure 1. Growth in State QRIS



Source: QRIScompendium.org, October 2017

A review of trends in QRIS participation reveals significant growth over time in the overall number and proportion of eligible programs participating. In 2010, more than half of the 22 QRIS implemented had fewer than 1 of 3 eligible center-based programs enrolled in the QRIS (Tout et al., 2010). In 2016, program participation rates in many QRIS included over half of the eligible programs in the state or local area: 22 of 41 QRIS reported that more than 50 percent of eligible center-based programs were participating, and 16 QRIS reported that more than half of eligible family child care programs were participating in the QRIS (Friese, Starr, & Hirilall, forthcoming).

With the recent expansion in QRIS participation, it is important to take stock of QRIS as a system framework for quality improvement. Evaluation can play an important role in the process of examining the components of QRIS—ratings, outreach strategies, quality improvement supports, grants and awards for programs, websites with ratings and information about selecting ECE programs—and their effectiveness in supporting desired outcomes.

Validation studies are one type of QRIS evaluation that examine a critical but relatively narrow set of questions about how well the quality measurement and rating processes are working to differentiate meaningful levels of ECE program quality in a QRIS (Zellman & Fiene, 2012). Researchers in the Quality Initiatives Research and Evaluation Consortium (INQUIRE) have described validation as a collection of research activities that contributes to improvement of the QRIS. If QRIS ratings are associated positively with external measures of quality and gains in children's development, program administrators and policymakers have initial evidence suggesting that the ratings are helpful in differentiating quality; a lack of associations indicates that the ratings need revisions (assuming that the study design and methods were appropriate for the questions of interest). Validation does not result in a yes or no answer about effectiveness of the QRIS (e.g., Lahti et al., 2013; Tout & Starr, 2013; Zellman & Fiene, 2012).[2] Validation analyses can include a review of evidence for the quality indicators included in the QRIS, an examination of scoring on the quality indicators, and an assessment of how the overall ratings are associated with external measures of quality and patterns of children's development. Validation studies analyzing how QRIS ratings predict measures of quality and children's development were required by states that received RTT-ELC grants.

[2] INQUIRE is designed to facilitate the identification of issues and the development and exchange of information and resources related to the research and evaluation of QRIS and other quality initiatives. INQUIRE is funded through the Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families (ACF) and managed through a contract with Child Trends.

While validation was a new activity for some RTT-ELC grantees, validation studies had been conducted as early as 1999. Karoly (2014) conducted a literature review of 14 validation studies conducted between 1999 and 2011 (not including the 10 studies in this synthesis) and found a small positive correlation between ratings and quality scores, suggesting that the ratings were generally capturing meaningful, albeit small, differences in quality on the measures. Although some studies found significant associations between ratings or components of ratings and child development measures, others did not. Overall, there was weak, inconsistent evidence of a link between QRIS ratings and child development in the previous studies.

The current paper builds on Karoly's (2014) synthesis by examining findings from 10 state QRIS validation studies conducted after 2013 and completed by August 2017, in Arizona (Epstein et al., 2017), California (Quick et al., 2016), Delaware (Karoly et al., 2016), Maryland (Swanson et al., 2017), Massachusetts (Wellesley Centers for Women & UMass Donahue Institute Applied Research & Program Evaluation, 2017), Minnesota (Tout et al., 2016), Oregon (Lipscomb et al., 2017), Rhode Island (Maxwell et al., 2016), Washington State (Soderberg et al., 2016), and Wisconsin (Magnuson & Lin, 2015; Magnuson & Lin, 2016).[3] Whereas the studies in the original review varied significantly in their sample sizes and methods, the 10 new studies included in this synthesis were remarkably similar in the overall designs and methods used.

This synthesis focuses on study results for two core validation questions: 1) To what extent are QRIS ratings associated with measures of observed quality? 2) To what extent are QRIS ratings associated with gains in children's development? These questions align with those required by RTT-ELC grants. This report focuses on the core validation questions and the consistency of findings across the states. Overall, the synthesis is intended to provide QRIS administrators, stakeholders, and researchers with a comprehensive look at findings across states to inform next steps in QRIS design, implementation, and evaluation.

Approach to the Synthesis

The process used to produce this report included collaboration across the research teams for the ten states included in the synthesis.[4] The group developed the plan for the synthesis and the outline for the report, and contributed analysis and writing. The collaborative process for the synthesis included several steps of document review, meetings with research teams, and review and discussion of findings across studies. Descriptive data and findings from multivariate analyses were mostly included exactly as they appeared in the state reports; however, for certain analyses, some state teams produced new tables to align their analysis and facilitate comparison across states.

Characteristics of QRIS in the Synthesis

The QRIS included in the synthesis share some common characteristics, although (as with all QRIS across the nation) each has a unique profile of design and implementation features. Most of the 10 QRIS included in the synthesis were within the first three to five years of an implementation transition (to a statewide system or to a new structure for the QRIS, in response to receipt of the RTT-ELC grant) at the time the validation studies were conducted. The density of participation among eligible center-based and family child care programs in each QRIS ranged from under 10 percent in California to around 80 percent in Delaware and Wisconsin. Five of the QRIS (Maryland, Massachusetts, Minnesota, Washington, and Wisconsin) had uneven distributions of rated programs, with one level representing over half of the program ratings. In the other five states, ratings were more equally distributed across rating levels.

Two QRIS (Massachusetts and Minnesota) had four rating levels; the remaining eight had five rating levels. A quality rating is generated in different ways depending on the rating structure chosen for the QRIS. Some structures designate quality indicators at each QRIS level and require that all indicators are met to achieve the level and move to the next (block structure). Others assign points to indicators and award a rating based

[3] Nine of the ten studies were conducted by state RTT-ELG grantees. Arizona is not an RTT-ELC grantee, but its validation study addressed the first validation question and was conducted during the same time period as the other state studies.
[4] Although timing and release of the state reports varied, the research teams began meeting in 2015 to discuss experiences with recruitment, measures selection, and analysis strategies.

on the number of points achieved. Some QRIS use a hybrid of these methods and have a mix of points and required indicators. Six of the QRIS included in the synthesis (Arizona, California, Delaware, Minnesota, Washington, and Wisconsin) used a hybrid structure, while four (Maryland, Massachusetts, Oregon, and Rhode Island) used block structures. The indicators of quality used in each QRIS tapped common domains of quality, including supports for child health and wellness, staff qualifications, and child assessment. Most QRIS also included indicators of curriculum and teacher-child interaction, program environment, and family partnerships. However, the way that indicators are operationalized and the requirements to meet the quality indicators differ across systems. As a result, ECE programs from different states with the same rating might have demonstrated very different levels of quality.

Measures, Methods, and Limitations of the Validation Studies

To address the two main validation questions, the studies recruited ECE programs participating in the QRIS, as well as preschool-aged children and sometimes infants/toddlers attending those programs (ES Table 1). All state studies included center-based community child care programs and Head Start programs. Eight of the ten states included family child care programs (all but Massachusetts and Rhode Island), and eight included school-based pre-kindergarten programs (all but Massachusetts and Wisconsin).

ES Table 1. Program and Child Sample Sizes by Validation Study

	Centers	Family Child Care	Preschoolers	Infants/ Toddlers
Arizona	774	148	n/a	n/a
California	175	47	1,611	n/a
Delaware	139	42	1,123 children	
Maryland	257	98	n/a	n/a
Massachusetts	126	n/a	481	190
Minnesota	278	66	1,181	n/a
Oregon	153	159	n/a	n/a
Rhode Island	71	n/a	332	n/a
Washington	76	24	522	239
Wisconsin	122	35	887	n/a

Source: Individual state validation reports

The number of center-based programs in the study samples ranged from 71 in the smallest state (Rhode Island) to 774 in Arizona, where data were drawn from both observations conducted by the research team and data maintained by the QRIS. Family child care programs generally involved a smaller sample, ranging from 24 in Washington to 159 in Oregon. Massachusetts and Rhode Island did not include family child care in their validation studies. The size of the child samples ranged from 332 in Rhode Island to 1,611 in California.

Of the seven states that included assessment of children's skills,[5] only one (Washington) included children of all age groups (infants, toddlers, and preschoolers). Delaware and Massachusetts included toddlers and preschoolers (2- through 5-year-olds). The remainder included only children in the year prior to kindergarten (4-year-olds) or children who would likely enroll in kindergarten in one or two years (3- and 4-year-olds).

A challenge noted across the validation studies was insufficient samples of programs available at each QRIS level to compare each of the four or five levels in the system to every other level. Three states (Arizona, Minnesota, and Wisconsin) combined programs into a lower- and a higher-quality group for comparison

[5] Oregon's analyses examining QRIS ratings and child development will be included in a forthcoming report.

purposes. This combination across rating levels allowed for more statistical power to address the research questions. The methods used to combine rating levels address the question of whether QRIS ratings are distinguishing between levels of program quality, but cannot inform whether *every* level in a QRIS distinguishes levels of quality relative to every other level. In four of the state studies (California, Delaware, Washington, and Wisconsin), the lowest level or levels of the QRIS were not included in the validation analyses.

Of the seven validation studies that examined the association between QRIS rating and child development (California, Delaware, Massachusetts, Minnesota, Rhode Island, Washington, and Wisconsin), the family and child factors most commonly controlled for in the analyses were gender and children's fall assessment scores (five states). Four states controlled for family income, parent education, and child language; three controlled for race/ethnicity, time between fall and spring assessments, special need/disability status of children, subsidy receipt, and child age. Other variables included in some studies were the number of hours children were scheduled to attend care, absenteeism, and regional characteristics.

Several different tools were used in the studies to measure observed quality in classrooms and family child care programs at each QRIS level and to examine the differences in scores across levels or combinations of levels. The Classroom Assessment Scoring System (CLASS Pre-K, Pianta et al.,2008; CLASS Toddler, La Paro et al., 2012), Environment Rating Scales (ECERS-R, Harms et al., 2005; FCCERS-R, Harms et al., 2007; ITERS-R, Harms et al., 2007)[6] and Program Quality Assessment (Highscope, 2003) were all used in at least two validation studies included in this synthesis.

Evaluators conducted child assessments with preschoolers and/or toddlers in the fall and in the spring in multiple domains of development, including language and literacy [Peabody Picture Vocabulary Test (PPVT), Test of Preschool Early Literacy (TOPEL), and Woodcock Johnson III Tests of Achievement - Letter and Word], math (Woodcock Johnson III Tests of Achievement – Applied Problems), executive function [Head, Toes, Knees, Shoulders (HTKS) and Peg Tapping], general cognition (Bracken Basic Concepts), and physical development. Teacher reports of children's social-emotional development and approaches to learning [Social Competence and Behavior Evaluation (SCBE-30), Preschool Learning Behaviors Scale (PLBS), and the Devereux Early Childhood Assessment (DECA)] were also collected.[7]

The validation studies examined the associations between QRIS ratings and each developmental outcome. Two states (Massachusetts and Minnesota) used gain scores as the outcome (spring score minus fall score), and other states used spring score as the outcome and include the fall score as a covariate. Validation researchers typically used Hierarchical Linear Modeling (HLM, also known as multilevel regression models) for their analyses. HLM is necessary because multiple children provide scores for each classroom or program (children are nested in programs).

The studies were limited by contextual and methodological issues. The QRIS examined in the studies were in a period of change due to RTT-ELC grant activities. Program recruitment was challenging due to small sample sizes, and study recruitment rates were low among some types of programs. Recruitment rates for children were difficult to calculate because the strategy used to recruit families and children in some studies does not allow simple calculation of the denominator (the total number of families reached by recruitment materials). Studies had missing data for several key data points for programs and children.

[6] ECERS-R: Early Childhood Environment Rating Scale-Revised; FCCERS-R: Family Child Care Environment Rating Scale-Revised; ITERS-R: Infant Toddler Environment Rating Scale-Revised.

[7] Measures noted in this paragraph are those used in at least two state studies. References for each are included in the reference section.

Question 1 Results: To what extent are QRIS ratings associated with measures of observed quality?

All states (nine total) examining QRIS ratings and associations with observed quality found positive associations with at least one of the quality measures examined. ES Table 2 highlights the findings for each state. In most states, researchers aimed to include at least one measure not already assessed as part of the QRIS rating process.

Five of seven state studies found a significant association between QRIS level and CLASS Pre-K Instructional Support (IS). For other CLASS domains (Emotional Support - ES and Classroom Organization - CO) and CLASS Toddler, findings were mixed. Arizona found significantly higher scores on the CLASS Pre-K (Emotional Support and Classroom Organization) for higher-rated relative to lower-rated programs. Oregon found significant differences between higher and lower rating levels on CLASS Pre-K (Emotional Support, Classroom Organization, Instructional Support) and CLASS Toddler Engaged Support for Learning. Rhode Island also found evidence for the CLASS Pre-K and Class Toddler but used a different type of analysis than the other states.[8] Maryland and Minnesota did not find any significant associations with CLASS Pre-K and ratings.

Arizona, Maryland, Massachusetts, Minnesota, and Wisconsin found a significant association between QRIS level and ERS scores. In all five states, ECERS-R/ECERS-3 was significantly higher in higher-rated center programs than in lower-rated programs. Arizona and Maryland also found significant differences across rating level on the FCCERS-R. In Minnesota, there was no evidence for differences in FCCERS-R scores across rating levels. In Wisconsin, the FCCERS-R was collected but was analyzed jointly with the ECERS-R.

Delaware and California also found a significant association between QRIS level and observed measures of quality with the PQA. Both studies found significantly higher PQA scores (either overall mean or the adult-child interaction subscale) at the highest QRIS level than at lower levels. Massachusetts found significant differences in preschool classrooms on the CIS.

In seven of the nine studies examining ratings and observed quality (all but Delaware and Rhode Island), at least one measure of quality used in the validation study was also included in the QRIS rating calculation itself (at least at some levels of the QRIS). Overall, five of seven validation studies that used the CLASS found at least one significant association between CLASS and QRIS level; CLASS was included in the rating for three studies that found significant associations and in two studies that did not find significant associations. Five of five studies that used the ERS found at least one significant association between ERS and QRIS level; ERS was included in the rating for four of the studies. While the validation studies did use measures included in the QRIS rating, six of the nine studies examining ratings and observed quality found significant associations using at least one independent quality measure.

[8] Researchers treated observation scores as continuous variables and tested for the association between QRIS level and quality outcomes.

ES Table 2. Summary of Associations between QRIS Ratings and Observed Quality, by State and Observational Measure

	CLASS Pre-K			CLASS Toddler		ERS			Other Quality Measures	
State	**Instructional Support**	**Emotional Support**	**Classroom Org.**	**Emotional and Behavior Support**	**Engaged Support for Learning**	**ECERS-R**	**ITERS-R**	**FCCERS-R**	**PQA***	**CIS**
Arizona	✓	✓	✓			✓		✓		
California	✓	*ns*	*ns*						✓	
Delaware	✓	*ns*	*ns*	*ns*	*ns*				✓	*ns*
Maryland	*ns*	*ns*	*ns*			✓		✓		
Massachusetts						✓	✓			✓
Minnesota	*ns*	*ns*	*ns*			✓		*ns*		
Oregon	✓	✓	✓	*ns*	✓					
Rhode Island	✓	✓	✓	✓	✓					
Wisconsin						✓		[1]		

Source: Individual state validation reports

*PQA total score for DE, Adult-Child Interaction for CA. [1]The FCCERS-R was collected in Wisconsin but analyzed jointly with the ECERS-R.

Note: A check mark indicates at least one statistically significant association was found demonstrating higher observed quality at higher rating levels. "Ns" indicates that no statistically significant associations were found. A gray, blank cell indicates that the measure was not collected.

Question 2 Results: To what extent are QRIS ratings associated with measures of children's development?

Across the seven validation studies that examined child development, evidence for a significant link between QRIS rating level and child development was inconsistent (see ES Table 3).

The validation of the California QRIS assessed children in four QRIS levels, using Level 3 as the comparison category. The results indicated one significant finding in the expected direction: children in Level 5 programs scored higher on peg tapping (a measure of executive function) than children in Level 3 programs. There were no other significant differences in the expected direction (higher levels scoring higher than lower levels).

Two other validation studies found some evidence for differences in executive function by QRIS level, but with caveats. In Delaware, children in Level 5 programs had significantly higher scores on HTKS than children in Level 2 programs. The sum of points on the six essential standards in Delaware was also significantly associated with executive function. In Wisconsin, QRIS level did not predict HTKS, but total rating points did significantly predict HTKS scores.[9]

Additionally, in Wisconsin, children in Level 5 scored significantly higher on the PLBS (approaches to learning, persistence) than children in Level 2. The Minnesota study also found a significant difference in PLBS (only the persistence subscale was used), with children in higher-rated programs scoring higher than children in lower-rated programs. Minnesota also found significant gains in social competence for children in higher-rated programs than in lower-rated programs.

[9] Because there is more variability in points across programs than in ratings levels, using points to predict child outcomes can reveal associations between program quality and child outcomes when overall rating level did not.

The Massachusetts study found that children in programs rated Level 3 showed significantly greater gains in their PPVT (receptive language) scores over the course of the school year than those in Level 2 programs. In addition, children in Level 3 programs showed significantly greater gains on the attachment subscale of the DECA than children in Level 1.

Two significant language outcomes were found in Washington. Infants and toddlers in Level 4 scored significantly higher than infants and toddlers in Level 3 on expressive language, and preschoolers scored significantly higher on receptive language in Level 3 than in Level 2. Also in Washington, infants and toddlers scored significantly higher on fine motor skills in Level 3 than in Level 2.

In the Rhode Island study, the authors designated findings by their statistical significance (at $p<.10$) and substantive significance (with an effect size of at least .07, per criteria set by the What Works Clearinghouse). A significant but not substantive negative association was found with overall rating and expressive vocabulary. No other statistically significant and substantive findings were noted with overall rating; however, significant and substantive associations were found between multiple components of the rating scale for math and social competence.

Overall, three of six states (California, Delaware, and Wisconsin) found evidence of a significant association between QRIS rating (or overall points obtained) and executive function. Significant associations between QRIS rating (or rating components) and social-emotional development were found in four states: Massachusetts, Minnesota, Rhode Island, and Wisconsin.[10]

ES Table 3. Summary of Associations between QRIS Ratings and Child Development, by State and Developmental Domain

	Executive Function	Language/ Literacy	General Cognition	Physical Development	Social/ Emotional	Math
State	Peg tapping/ HTKS	PPVT/ TOPEL/ IDGI/WJ/ Story and Print/Mullen	Bracken/ Mullen	BMI/ Mullen fine/ gross motor	SCBE-30/ PLBS/ DECA/ CBCL	TEAM/ WJ
California	✓	*ns*				*ns*
Delaware	✓[1]	*ns*			*ns*	*ns*
Massachusetts		✓			✓	
Minnesota	*ns*	*ns*	*ns*	*ns*	✓	*ns*
Rhode Island	*ns*	—[3]			✓[4]	✓[5]
Washington	*ns*	✓	*ns*	✓	*ns*	*ns*
Wisconsin	✓[2]	*ns*	*ns*		✓	*ns*

Source: Individual state validation reports
Note: A check mark indicates a statistically significant positive association was found between rating level and children's development. A negative sign indicates a statistically significant negative association was found between rating level and children's development. "Ns" indicates that no statistically significant associations were found. A gray, blank cell indicates that the measure was not collected.

[1] The analysis in Delaware found a significant difference between level 5 and level 2 only; in addition, a significant association was noted with executive function and the sum of points on the six essential standards. [2] The analysis in Wisconsin found a significant association with total rating points, not rating level. [3] The analysis in Rhode Island found a significant negative association between rating and expressive vocabulary. [4] The analysis in Rhode Island found significant associations between social competence and multiple rating components (but not overall rating). [5] The analysis in Rhode Island found significant associations between math and multiple rating components (but not overall rating).

[10] The overall pattern was not consistent in Wisconsin, which raises the question of whether it was a spurious association.

Summary and Implications

Within each state, validation study findings have been used as a key input to inform improvement of the state QRIS, and the research teams worked closely with their state partners to discuss and interpret the results from individual state studies. The value of the synthesis comes from looking across the 10 state studies, keeping in mind the limitations of comparing systems with different implementation timing and characteristics. Rather than informing specific decisions that each state makes about QRIS design and implementation, the purpose of the synthesis is to provide a general assessment of the extent to which QRIS ratings serve as useful tools for early childhood systems.

Acknowledging the limitations of the validation studies, two key implications can be drawn from the study results.

First, QRIS ratings appear to be a helpful tool for state early childhood systems to differentiate programs at lower and higher levels of quality. Overall, QRIS ratings reflect differences in environments, interactions, and activities in ECE programs at different rating levels. Although statistically significant, the differences in observed quality scores between QRIS rating levels were generally small, and findings for family child care programs had mixed results. A review of the individual state studies indicates that it is important to consider the integrity of the rating by ensuring an appropriate number of quality indicators, as well as indicators that can ensure rigor of differentiation, particularly at the highest QRIS levels.

Second, the results documenting observed quality at medium and low levels across many QRIS programs highlight the need for continued investment and innovation in quality improvement supports for ECE programs. Research indicates that programs improve over time in QRIS, but few studies have documented the most effective ways to promote meaningful improvement that can be sustained and can support children's positive development (Karoly, 2014). Longitudinal studies to understand how programs improve and how teachers and caregivers perceive the quality improvement process will add value to the existing literature (see, for example, Elicker et al., 2017).

Next Steps

Several activities could build upon the studies and results described in this report to enhance quality measurement and QRIS:

- It will be important to build the literature on family child care programs in QRIS and understand how current quality measures are working in these settings. As enrollment of family child care providers in QRIS increases, efforts to document their quality will inform the field.
- Validation studies (specifically) and quality improvement studies (generally) need to include children with special needs, infants and toddlers, and children who speak languages other than English and Spanish. Understanding how program quality is associated with outcomes is limited by the exclusion of these important populations of children.
- Measures of children's experiences in early care and education—beyond traditional school-readiness skills—should be included. The forthcoming study in Oregon will include a measure of children's engagement in early learning settings and a measure of family-teacher relationship quality.[11] Other important measures of children's experiences in ECE could include the quality of relationships with staff and peers and children's continuity in high-quality settings.
- As described in this report, validation studies examine how ratings are associated with measures of observed quality. However, quality ratings incorporate different domains of quality that may not be assessed by current observational measures. Even when small differences are noted between levels on observational measures, the QRIS indicators may be capturing other aspects of quality that contribute to the experiences of children and families in ECE programs. For example, programs that meet indicators related to the work environment may

[11] Although not described in the synthesis, the Maryland validation study included a measure of child engagement but did not find significant differences by rating level (Swanson et al., 2017).

have more stable staff than other programs. A next step for validation studies and other studies of quality is to examine associations between ratings and indicators such as turnover, compensation, and other workforce supports. These may not be directly associated with observational measures or children's development, but may reflect important infrastructure elements for building and improving quality. Oregon, for example, included some of these outcomes in its validation study (Lipscomb et al., 2016).

- Finally, QRIS ratings are used for a variety of purposes. For example, ratings are used to target quality improvement supports, target scholarships for vulnerable children to access higher-quality programs, and provide information to parents making ECE choices. In a QRIS with a block structure, some QRIS levels may be set to encourage quality improvement, but not to discern meaningful differences in children's development at each level. Some quality levels may be set to engage programs in the system. Additionally, some quality levels and indicators may have clear connections to higher-quality practices that can support children's positive development. Clarifying the theory of change for each QRIS can help identify more accurate hypotheses about which quality levels and quality indicators should be differentiating observed quality and children's development. This may not, in fact, be an explicit goal at every level of the system (Schilder et al., 2015).

Overall, the validation studies described in this report provide helpful guidance to inform the next round of improvements to QRIS ratings across the country. Indeed, each of the states approached validation as a strategy to improve their rating process and tools. The studies indicated that the ratings are generally working to distinguish lower and higher quality, but that further work is needed to strengthen quality measurement. Limited positive associations were found between ratings and children's development. These findings can prompt discussions about how to improve quality measurement and support quality improvements that promote the development of young children in ECE programs.

Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies

QRIS Validation and Purpose of the Validation Synthesis

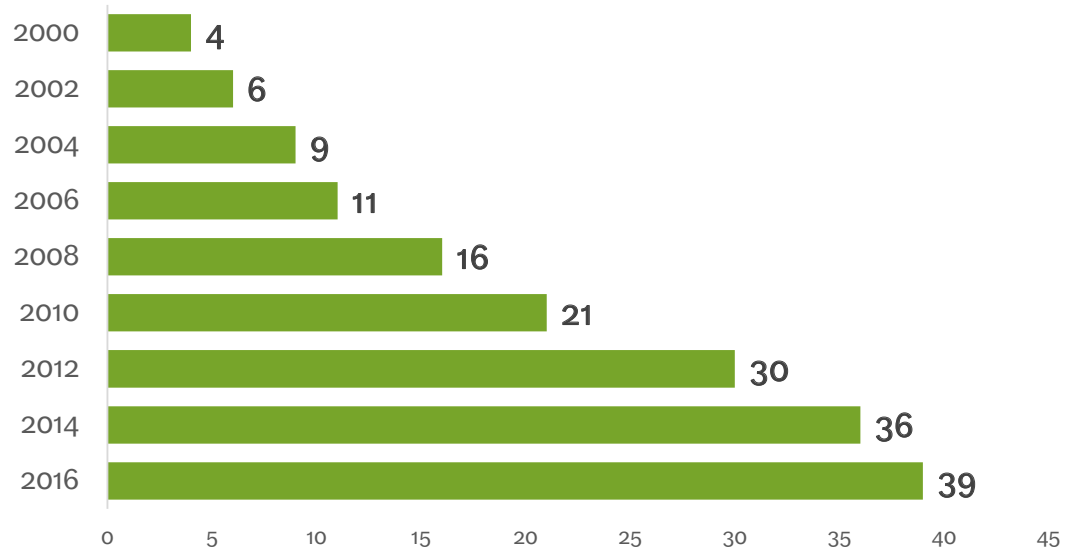
The purpose of this report is to compile and analyze findings from 10 validation studies examining ratings of early care and education (ECE) programs participating in state Quality Rating and Improvement Systems (QRIS). The availability of recent research results addressing similar questions in 10 different states offers a rare opportunity to synthesize findings across multiple contexts and discuss the implications for design, implementation, and future research on state ECE quality initiatives. The report is intended to update state administrators and other stakeholders about the effectiveness of current QRIS quality ratings in distinguishing meaningful levels of quality. The report also addresses issues of interest to researchers conducting evaluations of quality initiatives.

QRIS are initiatives implemented in 39 states[12] to promote improvement in the quality of ECE programs. Although systems vary in their specific features, QRIS typically include a process for measuring and rating ECE program quality, sharing ratings with parents and the public, and providing supports (including financial incentives) to help programs improve their quality. QRIS are implemented by state agencies in partnership with other statewide and community-based organizations. These organizations (including child care resource and referral agencies and universities) provide observations of ECE programs; coordinate the provision of technical assistance on best practices; and connect directors, teachers, and family child care providers to professional development. Investment in QRIS activities represents a significant portion of state spending on early care and education. In structure and scope, most QRIS are intended to be a centerpiece of the early childhood system and serve as an organizer of state or local efforts to improve the quality of ECE programs, particularly among those programs serving children who are vulnerable or need additional supports (Mathias, 2015; Zaslow & Tout, 2014).

The first QRIS in the United States became operational in 1997. Growth in state QRIS was relatively slow and steady in the early 2000s, when there was a 45 percent increase from 2006 to 2008 (Figure 1). Rapid growth from 2010 to 2014 (a 54% increase) was due in part to the federal Race to the Top – Early Learning Challenge (RTT-ELC). Grants were awarded to 20 states and required a QRIS (RTT-ELC, 2011).

[12] Note that California and Florida each have multiple QRIS operating at a county or local level.

Figure 1. Growth in State QRIS



Source: QRIScompendium.org , October 2017

Participation of ECE programs is voluntary in many QRIS. Recently, however, several QRIS have added mandatory participation for center-based and licensed family child care programs receiving funding, such as Child Care and Development Fund subsidies or state pre-kindergarten funding. Some states embed QRIS into licensing for ECE programs; in these states, licensing serves as the first level of the QRIS, and programs have the option to achieve rating levels above the first level. Across this mix of state systems, the proportion of eligible programs participating in the QRIS is a helpful metric to depict the reach of the system. A review of trends in QRIS participation reveals significant growth over time in the overall number and proportion of eligible programs participating. In 2010, more than half of the 22 QRIS implemented had fewer than 1 of 3 eligible center-based programs enrolled in the QRIS (Tout et al., 2010). In 2016, program participation rates in many QRIS included over half of the eligible programs in the state or local area: 22 of 41 QRIS reported that more than 50 percent of eligible center-based programs were participating, while 16 QRIS reported that more than half of eligible family child care programs were participating in the QRIS (Friese, Starr, & Hirilall, forthcoming).

With the recent expansion in QRIS participation, it is important to take stock of QRIS as a system framework for quality improvement. Evaluation can play an important role in the process of examining the components of QRIS—ratings, outreach strategies, quality improvement supports, grants and awards for programs, websites with ratings and information about selecting ECE programs—and their effectiveness in supporting desired outcomes.

Validation studies are one type of QRIS evaluation that examine a critical but relatively narrow set of questions about how well the quality measurement and rating process are working to differentiate meaningful levels of ECE program quality in a QRIS (Zellman & Fiene, 2012). Researchers in the Quality Initiatives Research and Evaluation Consortium (INQUIRE) have described validation as a collection of research activities that contributes to improvement of the QRIS. Validation does not result in a yes or no answer about effectiveness of the QRIS (e.g., Lahti et al., 2013; Tout & Starr, 2013; Zellman & Fiene, 2012).[13] Validation analyses can include a review of evidence for the quality indicators included in the QRIS, examination of scoring on the quality indicators and an assessment of how the overall ratings are associated with external measures of quality and patterns of children's development. In a "valid" QRIS, a higher-rated

[13] INQUIRE is designed to facilitate the identification of issues and the development and exchange of information and resources related to the research and evaluation of QRIS and other quality initiatives. INQUIRE is funded through the Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families (ACF) and managed through a contract with Child Trends.

program also has higher observed quality, technical assistance supports programs in improving their quality and ratings over time, and ratings reflect the dimensions of program quality that promote positive child development (Karoly, 2014). Validation studies analyzing how QRIS ratings predict measures of quality and children's development were required by states that received RTT-ELC grants.

While validation was a new activity for some of the RTT-ELC grantees, validation studies had been conducted as early as 1999. Karoly (2014) conducted a literature review of 14 validation studies conducted between 1999 and 2011 (not including the 10 studies in this synthesis) and summarized their questions, methods, and findings. Across the studies, the most commonly addressed research question asked about the association between QRIS ratings and other measures of quality. All validation studies used the Environment Rating Scales (ECERS-R, Harms et al., 2005; FCCERS-R, Harms et al., 2007; ITERS-R, Harms et al., 2007)[14] as a measure of observed global quality. A few also used the Caregiver Interaction Scale (Arnett, 1989) or the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) to measure process quality (including teacher-child interactions) and to provide a measure that was not embedded in the QRIS rating.

Results across most studies in Karoly's (2014) synthesis found a small positive correlation between ratings and quality scores, suggesting that the ratings were generally capturing meaningful, albeit small, differences in quality on the measures. A limitation of the studies was that the findings often did not include a measure of observed quality that was external to the QRIS rating.

Six of the validation studies Karoly reviewed examined program improvement in QRIS ratings over time. The studies each found evidence for program movement up the rating scale. However, the study designs did not allow conclusions to be drawn about whether the QRIS or quality supports caused the improvement (Karoly, 2014).

Seven validation studies in Karoly's (2014) synthesis addressed the association between QRIS ratings and child development. The studies varied in their methods. Although some studies found significant associations between ratings or components of ratings and child development measures, others did not. Taken together, there was weak, inconsistent evidence of a link between QRIS ratings and child development (Karoly, 2014).

The current paper builds on Karoly's (2014) synthesis by examining findings from 10 state QRIS validation studies conducted after 2013 and completed by August 2017: in Arizona (Epstein et al., 2017), California (Quick et al., 2016), Delaware (Karoly et al., 2016), Maryland (Swanson et al., 2017), Massachusetts (Wellesley Centers for Women & UMass Donahue Institute Applied Research & Program Evaluation, 2017), Minnesota (Tout et al., 2016), Oregon (Lipscomb et al., 2017), Rhode Island (Maxwell et al., 2016), Washington State (Soderberg et al., 2016), and Wisconsin (Magnuson & Lin, 2015; Magnuson & Lin, 2016).[15] Whereas the studies in the original review varied significantly in their sample sizes and methods, the 10 new studies included in this synthesis were remarkably similar in the overall designs and methods used.

This synthesis focuses on study results for two core validation questions: 1) To what extent are QRIS ratings associated with measures of observed quality? 2) To what extent are QRIS ratings associated with gains in children's development? These questions align with those required by RTT-ELC grants. This report focuses on the core validation questions and the consistency of findings across the states. Overall, the synthesis is intended to provide QRIS administrators, stakeholders, and researchers with a comprehensive look at findings across states to inform next steps in QRIS design, implementation, and evaluation.

The report begins with an overview of the approach to the synthesis and a review of the QRIS features for each state included. The report also describes validation methods, measures, and data collection strategies. Results are presented in two sections, one describing QRIS ratings and observed quality and one describing

[14] ECERS-R: Early Childhood Environment Rating Scale-Revised; FCCERS-R: Family Child Care Environment Rating Scale-Revised; ITERS-R: Infant Toddler Environment Rating Scale-Revised.

[15] Nine of the ten studies were conducted by state RTT-ELG grantees. Arizona is not an RTT-ELC grantee, but its validation study addressed the first validation question and was conducted during the same time period as the other state studies.

QRIS ratings and children's development. The report concludes by putting the findings in context with a summary of recommendations in the state reports, a summary of validation study findings, a review of limitations of the studies, and a discussion of implications. The report also contains three commentaries that offer perspectives on QRIS validation activities and findings from three vantage points: a national perspective, a state perspective, and a research perspective. Information about the authors of each commentary is included at the end of the report.

Approach to Synthesis of Validation Studies

This synthesis of QRIS validation studies looks across 10 validation reports, including nine from RTT-ELC states, to draw conclusions about consistency and patterns of findings. The process used to produce this report included collaboration across the research teams for the 10 states in the synthesis.[16] The group developed the plan for the synthesis and the outline for the report, and contributed analysis and writing. The collaborative process for the synthesis included several steps of document review, meetings with research teams, and review and discussion of findings across studies:

1. QRIS materials and validation reports were reviewed from 10 states: Arizona, California, Delaware, Maryland, Massachusetts, Minnesota, Oregon, Rhode Island, Washington, and Wisconsin. Details about the QRIS—including participation rates, rating structure, and quality standards—were entered in an Excel table to facilitate review and analysis. Similarly, details from validation reports—such as research methods used, research questions addressed, data collection tools used, and findings—were also organized for each state.
2. A team of validation researchers was invited to participate in regular conference calls to discuss production of the validation synthesis.
3. The team identified alignment of data collection tools and compared studies that used the same or similar measures of observed quality and child development.
4. Individual meetings with researchers from each state were conducted to gather any additional data and request new analyses to facilitate comparison of findings.
5. Group meetings were held to discuss findings, the framework for the synthesis, and implications of the results.
6. Team members reviewed and edited multiple drafts of the report.

The main goal of this report is not to assist any one state in understanding its specific QRIS and how it might be improved, but rather to consider what can be learned about QRIS ratings in general. Within each state, validation study findings have been used as a key input to inform improvement of the state QRIS, and the research teams worked closely with state partners to discuss and interpret the results from individual state studies. The value of the synthesis comes from looking across the 10 state studies, keeping in mind the limitations of comparing systems with different implementation timing and characteristics. Rather than informing specific decisions that each state makes about QRIS design and implementation, the purpose of the synthesis is to provide a general assessment of the extent to which QRIS ratings serve as useful tools for early childhood systems.

[16] Although timing and release of the state reports varied, the research teams began meeting in 2015 to discuss experiences with recruitment, measures selection, and analysis strategies.

Characteristics of QRIS and the State Systems Included in Synthesis

To provide context for the results, this section provides a description of general QRIS features and specific details about the state QRIS included in the synthesis. QRIS were developed initially to bridge the gap between licensing, which provides regulations to protect children's health and safety, and ECE program accreditation, which outlines a comprehensive set of program quality standards (Mitchell, 2005). QRIS pioneer states in the late 1990s found that few programs were able to meet accreditation standards, and the leap from licensing to accreditation was too challenging for many programs, especially those serving children receiving child care subsidies with limited resources to invest in improvement (Zellman & Perlman, 2008). Early state QRIS in Oklahoma, Colorado, and North Carolina developed QRIS models that created tiers of quality intended to help programs move toward accreditation, and made ratings available to parents and the public to incentivize quality improvement (Zellman & Perlman, 2008). Early QRIS also embraced the notion of QRIS as an organizing structure to deliver and coordinate quality improvement activities in the state or local early childhood system. This goal was challenging for states to accomplish because of the level of coordination required across partners and the lack of funding to support necessary implementation structures for collaboration (Mitchell, 2008; Zaslow & Tout, 2014). In 2011, the RTT-ELC grants provided funds to promote a new phase of systems building in early care and education and emphasized QRIS as a centerpiece of system activities (Stoney, 2012; Mathias, 2015).

The basic QRIS components developed by the pioneer states are still included in the current set of 39 state QRIS, including the 10 states profiled in this synthesis. These components include quality standards, a monitoring and rating process, improvement supports and financial assistance for participating programs, and outreach to providers and parents. Each QRIS develops these components according to state or local contexts and available resources. Although some QRIS share common features, each one is unique, which makes comparison and synthesis of validation findings challenging.

As noted, a validation study focuses on the quality standards and the rating process in a QRIS. The quality standards are operationalized in a QRIS as indicators that can be measured with tools (either state-developed and/or published measures) or documented through review of materials submitted by the program. With revisions made as part of state RTT-ELC grants, QRIS standards have incorporated a greater focus on elements of quality linked to children's school readiness, including the use of child assessment tools and curriculum planning processes aligned with the developmental needs of the children served. Standards also shifted to include processes intended to support continuous quality improvement (CQI), including the development of improvement plans and gathering input from families and staff (Mathias, 2015). Nearly all QRIS include quality indicators related to staff qualifications, the learning environment, program administration and management, and family engagement (Build Initiative and Child Trends, 2016).

From their inception, most QRIS have included observational measures that assess aspects of process quality, including features of the learning environment, activities, routines, and children's experiences in the program. This trend has persisted over time, with nearly all QRIS in 2016 including an observational tool in some capacity—either as part of the rating process or to support self-assessment. Other modes for monitoring progress on quality indicators include the use of web-based evidence portfolios, monitoring visits by QRIS staff, and self-assessment of quality criteria.

A QRIS rating is generated in different ways depending on the rating structure chosen for the QRIS. Some structures designate indicators at each QRIS level and require that all indicators are met to achieve the level and move to the next (block structure). Others assign points to indicators and award a rating based on the number of points achieved. Some QRIS use a hybrid of these methods and have a mix of points and required indicators. The rating structure and the indicators selected for a QRIS influence the distribution of ratings awarded to programs. One study compared simulated QRIS models (block, points, and hybrids) while holding constant the quality indicators included in each model. The findings indicated that QRIS using block structures tend to have ratings skewed toward lower rating levels, while point and hybrid structures have ratings at higher levels (Tout et al., 2014).

QRIS also offer different rating options to programs with different characteristics. For instance, QRIS may offer an accelerated rating process to programs that have met quality criteria in a process outside the QRIS (e.g., national accreditation, Head Start, state pre-kindergarten standards). The accelerated rating process may include a rating based on fewer quality indicators and/or streamlined documentation requirements that provide a "fast track" into the QRIS, often to a rating at the highest level.

State QRIS

When reviewing the findings in this synthesis, it is important to keep in mind the unique features of each state system, including structure, number of rating levels, rating process, and whether quality observation is included in the rating.[17] Table 1 provides a brief summary of QRIS characteristics for the 10 states in the synthesis. Additional details about each QRIS follow the table. Details about other QRIS characteristics are also described, including stage of implementation, quality standards, rating structure, and participation numbers.

Table 1. Characteristics of QRIS in Synthesis

State	Rating Structure	Number of Levels	Accelerated Pathways*	Observation Score Part of Rating
Arizona	Hybrid	5 (1-5)	yes	ERS and CLASS
California	Hybrid	5 (1-5)	no	ERS and CLASS (at levels 3-5)
Delaware	Hybrid	5 (starting with stars, 2, 3, 4, 5)	yes	ERS (at levels 3-5)
Maryland	Block	5 (1-5)	yes	ERS and CLASS (at level 5)
Massachusetts	Block	4 (1-4)	no	ERS (at levels 3-4)
Minnesota	Hybrid	4 (1-4)	yes	CLASS (at levels 3-4)
Oregon	Block	5 (1-5)	yes	CLASS or FCCERS-R (at level 5 only)
Rhode Island	Block	5 (1-5)	no	ERS (at levels 3-5)
Washington	Hybrid	5 (1-5)	yes	ERS and CLASS
Wisconsin	Hybrid	5 (1-5)	yes	ERS (at levels 4-5)

Source: Individual state validation reports.
*Accelerated or automatic pathways are options available for certain programs to receive credit for reaching standards through other systems, including Head Start or national accreditation; as a result, programs have fewer requirements for rating and may move more quickly through the rating process.

[17] The terms levels and stars are used interchangeably in this report. In some cases, the terms are combined (e.g., "the highest star level") when it helps clarify the statement or finding.

Arizona

Arizona's QRIS, Quality First, was piloted in 2009–2011 and fully implemented in 2011. It is a voluntary system available to all licensed center-based and family child care programs. Quality First is a hybrid system with five star levels. Nonaccredited programs start the rating process with an ERS assessment. If programs score lower than a 3.0 on the ERS, they are rated at 1 or 2 stars. If the program receives an average program ERS score of at least 3.0, a CLASS assessment is administered. Programs that meet the cutoff scores for both the ERS and CLASS are then on the Quality First Points Scale (QFPS), which includes indicators related to staff qualifications, staff retention, ratios and groups size, and curriculum and child assessment. Accredited and Head Start programs start with the CLASS and only receive the ERS if they do not reach the criteria for at least the 3-star level on the CLASS (see Table 2). If they do reach the 3-star level, they are next assessed on the QFPS elements.

Table 2. Minimum Scores on Observation Tools Needed to Achieve Star Level in Arizona's Quality First

	Level 1	Level 2	Level 3	Level 4	Level 5
CLASS Pre-K: Emotional Support	n/a	n/a	4.5	5.0	6.0
CLASS Pre-K: Instructional Support	n/a	n/a	2.0	2.5	3.0
CLASS Pre-K: Classroom Organization	n/a	n/a	4.5	5.0	6.0
ERS Average Program Score	1.0	2.0	3.0	4.0	5.0

Source: Individual state validation reports

California

The California QRIS began in 2012 and was implemented by 17 local consortia across 49 California counties. It is a voluntary program in which all licensed center-based and family child care programs are eligible to participate. California's QRIS is a five-level hybrid system of block requirements and earned points. Programs must have a license in good standing to receive a tier 1 rating. Ratings at tiers 2–5 are based on requirements within three indicator categories: child development and school readiness, teachers and teaching, and program and environment. Requirements at tiers 1, 3, and 4 are common to all consortia. At tiers 2 and 5, local consortia can change requirements. Independent ERS and CLASS assessments are embedded into the ratings at tiers 3–5. For both observational tools, an independent observation is required at tier 3. The tier 3 observations are used to inform the professional development and quality improvement plans, but no minimum score is required. Programs must achieve minimum point thresholds to earn ratings at tiers 4 and 5, as seen in Table 3.

Table 3. California's QRIS Minimum Point Thresholds

	Level 3	Level 4	Level 5
CLASS Pre-K: Emotional Support	No minimum	5.0	5.5
CLASS Pre-K: Instructional Support	No minimum	3.0	3.5
CLASS Pre-K: Classroom Organization	No minimum	5.0	5.5
CLASS Toddler: Emotional & Behavioral Support	No minimum	5.0	5.5
CLASS Toddler: Engaged Support for Learning	No minimum	3.5	4.0
ERS	No minimum	5.0	5.5 or accreditation

Source: Individual state validation reports

The California system does *not* have an automatic or accelerated path to rating (e.g., accredited programs receiving the highest rating without going through the rating process). However, accreditation may be substituted for the ERS minimum at tier 5.

Delaware

Delaware Stars for Early Success, Delaware's QRIS, was implemented statewide as a block system in 2008 and changed to a hybrid system in 2012 using points-based standards and thresholds for the ERS. In 2014–2016, further modifications were phased in to add more required elements, among other changes. Delaware Stars is a voluntary system open to all licensed early care and education providers, including those operating within family care settings, private and public centers, and public schools. Delaware Stars has five tiers or star levels of quality.

To participate at the first star ("Starting with Stars"), programs must have a license in good standing. To move to the second star level, programs participate in required orientation sessions introducing the QRIS system, meet with a TA provider, and complete a Quality Improvement Plan. At star levels 3–5, programs earn points in four domains: family and community partnerships, qualifications and professional development, management and administration, and learning environment and curriculum. Programs must also earn minimum scores on the ERS (3.4 for level 3, 4.4 for level 4, 5.4 for level 5) and, as of 2015–2016, achieve several "essential" quality standards (previously part of the points-based component of the rating scale). Delaware Stars has an alternative ratings pathway for several types of programs. Programs for preschool-age children with disabilities enter at a Star 3 and move to higher ratings levels after demonstrating that they meet the associated ERS minimum score and satisfy the required number of points-based standards as determined by a portfolio review. Head Start programs enter at star level 4 and advance to star level 5 based on successfully meeting the ERS minimum score for that level. Programs accredited by NAEYC automatically receive ratings at star level 5.

Maryland

Maryland EXCELS went statewide in 2013, following a pilot and field test. Participation is mandatory for programs receiving child care subsidies. The QRIS serves licensed center-based programs and Head Start, licensed family child care programs, school-based programs, and school-age programs. It is a block system with five levels. Standards beyond licensing include staff qualifications and professional development, accreditation and rating scales (ERS and CLASS, for program improvement purposes only), developmentally appropriate learning and practice (includes curriculum and assessment), and administrative policies and practices.

Massachusetts

The Massachusetts QRIS was implemented at scale in 2011. It is mandatory for programs participating in child subsidies or grants programs, and voluntary for other programs. The QRIS serves licensed center-based and family child care programs, Head Start/Early Head Start, school-based programs, legally license-exempt center-based programs, and school-age programs. It is a block system with four levels. Programs are rated on five categories of standards: curriculum and learning (curriculum, assessment, diversity, teacher-child interactions); safe, healthy indoor and outdoor environments; workforce development and professional qualifications (for administrators and program staff); family and community engagement; and leadership, management, and administration. ERS scores are included at levels 2 (minimum overall ERS score 3.5), 3 (minimum 4.5), and 4 (minimum 5.5).

Minnesota

Minnesota's QRIS, Parent Aware, began its pilot phase in select counties in 2007, and was rolled out statewide between 2012 and 2015. Parent Aware is a voluntary program that serves licensed child care centers and family child care programs, Head Start and Early Head Start, and school-based pre-

kindergarten and Early Childhood Special Education programs. Parent Aware is a hybrid system with four star levels. Licensed, nonaccredited center and family child care programs must meet all requirements at the 1-star level to move to the 2-star level, and all requirements at the 2-star level before earning points to reach 3 and 4 stars. This "full rating" process includes indicators in four quality domains: physical health and well-being, teaching and relationships, assessment of child progress, and teacher training and education. The Classroom Assessment Scoring System (CLASS) is part of the rating for centers reaching 3 and 4 stars. Points are earned for scores of 4.0 or above for Emotional Support, 2.5 or above for Instructional Support, and 2.6 or above for Classroom Organization. All classrooms must earn at least a 2.0 on Instructional Support for a program to earn 3-stars and 2.5 to earn 4-stars. Accredited child care centers and family child care providers, Head Start and Early Head Start, and school-based pre-kindergarten and Early Childhood Special Education programs are eligible for the Accelerated Pathway to Rating and receive a 4-star rating after submitting documentation on curriculum and assessment indicators.

Oregon

Field-testing of Oregon's QRIS began in selected areas of the state in early 2013 and went statewide shortly after. It is a voluntary system for all types of regulated providers. Oregon has mostly a block structure in which programs must pass most of the standards for the 3-, 4-, or 5-star level to achieve a rating at that level. Level 1 of Oregon's QRIS represents programs that are licensed but have not voluntarily participated in the rating process. The validation study used an additional measurement process to select a subset of level 1 programs that were not likely to achieve a QRIS rating if they had applied for one. Level 2 indicates that the program has made a formal commitment to quality improvement by attending a QRIS training. Some of these Level 2 programs have submitted a portfolio but did not earn a rating of 3 or higher; the validation study only included level 2 programs that had applied for and not received a rating of 3 or higher. Programs are only required to submit materials specifically related to the star level for which they are applying. Standards are clustered into five domains: learning and development, personnel qualifications, family partnerships, health and safety, and administration and business practices. Observations of quality are only included at the 5-star level; programs must receive an average score of at least 5.0 on the relevant observational tool (CLASS infant, toddler, PreK, or FCCERS-R). Accredited and Head Start programs only needed to submit documentation on requirements not included in NAEYC accreditation or Head Start/Early Head Start standards, and are then rated based on the standards they meet. The QRIS ratings also rely on data from licensing and the Oregon Registry Online (a statewide information system containing workforce data) to assess qualifications of the workforce.

Rhode Island

Rhode Island's QRIS, BrightStars, was launched in 2009 and revised in 2013. BrightStars began as a voluntary system, but was changed in April 2014 to a system that required BrightStars participation for all programs participating in the Department of Human Services' Child Care Assistance Program (CCAP). BrightStars serves licensed child care centers and preschools (including public school and Head Start programs), family child care programs, and school-age programs. BrightStars uses a block rating structure to create a program rating that ranges from 1 star to 5 stars.

The 2013 BrightStars framework for child care centers and preschools includes 23 criteria grouped into 10 standards across six domains. These domains include health, safety, and nutrition; enrollment and staffing; staff qualifications and ongoing professional development; administration; early learning and development; and family engagement. Observational measures of classroom quality are conducted as part of the rating. Programs must have an average ECERS-R and ITERS-R score of 3.0 (no classroom below 2.5) to reach level 3, an average score of 4.0 (no classroom below 3.0) to reach level 4, and an average score of 5.0 (no classroom below 3.0) to reach level 5. Although Rhode Island does not have an "automatic" rating process, programs can use NAEYC accreditation or compliance with Head Start Performance Standards as evidence of meeting particular standards at certain rating levels.

Washington

The Washington state QRIS, called Early Achievers, was implemented statewide in 2012. The program is voluntary, with two exceptions: participation is mandatory for state pre-K programs and for programs that accept state funds or child care subsidies. Programs eligible to participate in the system include licensed center-based and family child care facilities, Head Start/Early Head Start programs, and state pre-kindergarten programs. Early Achievers' hybrid rating structure includes block foundational requirements at levels 1 and 2, and points earned at levels 3–5.

Programs are rated on four categories: child outcomes; facility curriculum, learning environment, and interactions; professional development and training; and family engagement and partnership. To achieve a level 1 rating, a program must have a license in good standing. To move to level 2, the program director/owner completes professional training series and an ERS self-assessment of the program. Ratings at levels 3–5 require ERS and CLASS observations by an independent assessor. Minimum scores on both the ERS and the CLASS (3.5 for ERS, 2.0 for Instructional Support, 3.5 for Emotional Support and Classroom Organization) are required to achieve a level 3 rating and points beyond the minimum accumulate toward higher ratings. Head Start and Early Childhood Education and Assistance Program (ECEAP) sites automatically enter Early Achievers at level 3 and are assessed only on the ERS and CLASS tools.

Wisconsin

Wisconsin's Quality Rating and Improvement System, YoungStar, was launched statewide in 2010 and is a voluntary system open to all licensed center-based and family child care programs. Programs serving children who receive CCDF subsidies are required to participate in YoungStar. Programs achieve star ratings within a points-based structure, in which there are minimum standards for higher rating levels. The rating system is comprised of five quality levels based on four indicator categories: education and training qualifications, learning environment and curriculum, professional and business practices, and child health and well-being practices.

Minimum scores are required on ERS observations for programs applying for ratings at the 4- and 5-star rating levels (averages must be 4.0 and 5.0 respectively). The YoungStar process includes automatic ratings for Head Start programs, which receive 5-star ratings; and accredited programs, which receive 4 or 5 stars depending upon the accrediting agency. In addition, programs that receive CCDF funding but do not want further involvement with YoungStar may receive a 2-star rating without any formal rating process, as long as they are meeting licensing requirements (a 1-star program is out of compliance with licensing regulations).

Stage of Implementation

Table 4 shows when each QRIS was started; when it went through significant transitions in areas served, rating structure, and/or revision of standards; and when validation data were collected. Validation data in Massachusetts, Minnesota, Oregon, Rhode Island Washington, and Wisconsin data were collected within 2 or 3 years of statewide implementation or a revision of the QRIS. Data were collected in even closer proximity to a system transition in Delaware and California.

Table 4. Stage of QRIS Implementation by State

QRIS	Year QRIS Started	Transition Type and Year (structure, area served)	Validation Data Collected
Arizona	2009	2011 went statewide	2016-2017
California	2011	2013 adopted in 17 consortia (pilot)	2013-2015
Delaware	2008	2012 changed from block to hybrid; 2015 added essential standards at higher star levels	2014-2015
Maryland	2011	2013 went statewide	2014-2016
Massachusetts	2010	2011 went statewide	2014-2015
Minnesota	2007	2011 went statewide and changed to hybrid	2013-2015
Oregon	2013	2014 went statewide	2013-2015
Rhode Island	2008	2009 went statewide; 2013 revision	2015-2016
Washington	2008	2012 went statewide	2014-2015
Wisconsin	2010	Implemented statewide	2013-2014

Source: Individual state validation reports

Program Participation and Density

The number of early care and education programs participating in a QRIS (participation) and the proportion of eligible programs enrolled (density) are two metrics that reflect the reach of a QRIS (Friese, Starr, & Hirilall, forthcoming). Program participation and density are dependent on stage of implementation and are important contextual factors. Information on participation and density is provided in Table 5. Some states (CA, DE, OR, WA) automatically enter licensed programs into the QRIS, and as a result have higher participation rates than other states. The number of family child care programs in a state can also affect participation density because family child care programs tend to participate at a lower rate than centers.

Participation is particularly relevant to validation studies. Low QRIS participation limits the number of programs that can be included in a validation study, generalizability, and the extent to which differences can be examined across the levels of the QRIS. A successful validation study depends in part on having a sufficient number of programs available to sample across each of the different rating levels in the QRIS.

Table 5. Participation and Density across QRIS in the Synthesis

QRIS	Number of Participating Programs (date)	Density (proportion of eligible providers participating in the QRIS)
Arizona[a]	962 (2016)	32%
California[b]	2,232 (2015)	7.4% (percentage in 2015 of licensed sites in the 16 participating counties)
Delaware[b]	454 (2014)	38%
Maryland[c]	3803 (2016)	53% (from 2015 APR)
Massachusetts[d]	4492 (from 2015 APR)	54%
Minnesota[b]	2,247 (2015)	18%
Oregon[e]	1,181 (from 2015 APR)	28%
Rhode Island[e]	739 (from 2015 APR)	83%
Washington[c]	2,303 (at initial recruitment into study)	43% (from 2014 APR)
Wisconsin[e]	4,339 (from 2014 APR)	79%

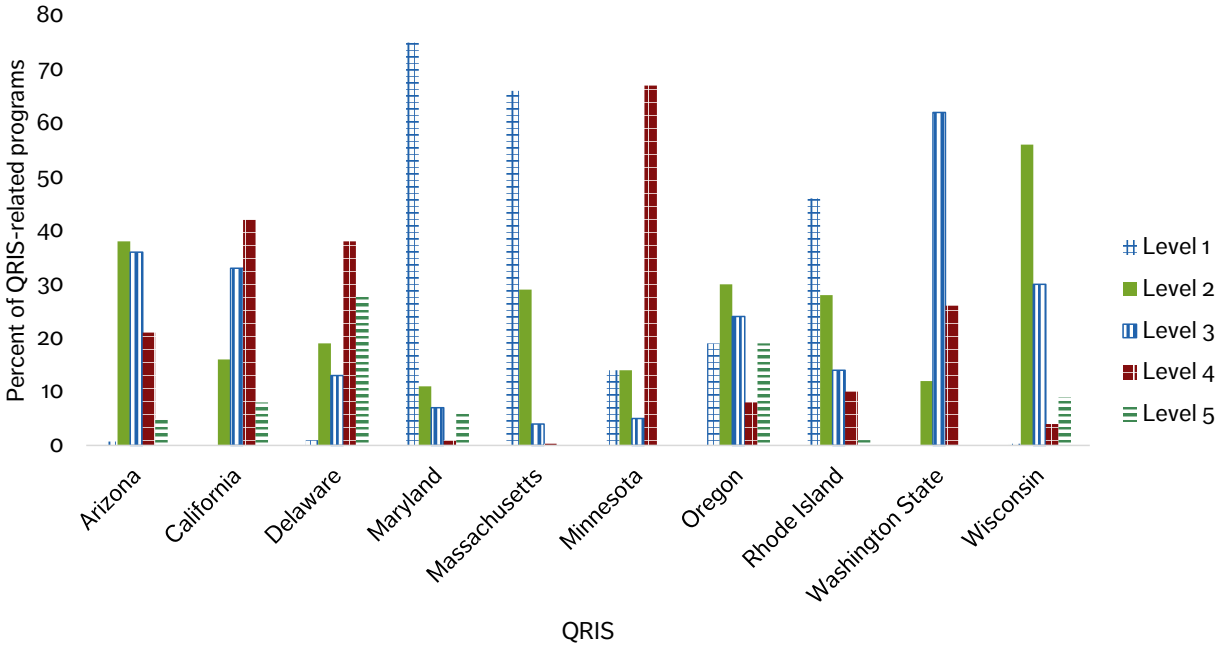
[a] *Source*: Participation data are from the state validation report; density data were provided via personal communication
[b] *Source*: State validation report
[c] *Source*: Participation data are from the state validation report; density data are from the Annual Performance Report, Race to the Top – Early Learning Challenge
[d] *Source*: Participation data are from the Annual Performance Report, Race to the Top – Early Learning Challenge; density data are from the QRIS Compendium (http://qriscompendium.org/)
[e] *Source*: Annual Performance Report, Race to the Top – Early Learning Challenge

Distribution of QRIS Ratings

The distribution of QRIS ratings across programs provides a useful context for examining the results of validation studies. The distribution is influenced by the rating structure, whether participation is mandatory for some or all programs, the stage of QRIS implementation, the difficulty of the indicators included, and the quality level of the programs enrolled in the QRIS. Each study in the synthesis provided some information on how the QRIS ratings are distributed (Figure 2).

Five of the QRIS (Maryland, Massachusetts, Minnesota, Washington, and Wisconsin) had uneven distributions, with one level representing over half of the program ratings. Each of the QRIS were still in relatively early stages of implementation, and, as discussed, the distributions posed a problem for some studies and their ability to recruit sufficient programs at each level.

Figure 2. QRIS Ratings Distributions by State



Source: Individual state validation reports

* MA and MN have four levels, while the other QRIS have five. This figure is meant to provide an approximate distribution of programs across levels.

Quality Standards and Indicators

QRIS vary widely in the quality standards and indicators used, the scores or QRIS levels assigned to each indicator, and the weight given to indicators in the overall rating.[18] Some indicators are found in very few QRIS across the country, while others are used in nearly all QRIS. To facilitate comparison across the 10 QRIS in the synthesis, the indicators were grouped into categories (see Table 6). The categories were created based on those used in the QRIS Compendium (QRIScompendium.org). In some instances, categories were combined or separated to more clearly capture the indicators used by the states in the present report. The teacher-child interaction category reflects requirements related to observational assessment—for example, using a tool like the ERS to support quality improvement in teacher-child interactions. Nine QRIS include indicators in program administration and management, curriculum, and family and community partnerships. All 10 states include indicators of child health and wellness, teacher-child interaction, program environment, child assessment, and professional qualifications.

[18] As noted, quality standards are the aspects of quality addressed in a QRIS and indicators are the way in which those standards are operationalized and measured.

Table 6. Categories of QRIS Indicators Included in Each State QRIS

	Child Health & Wellness	Curriculum	Teacher-Child Interactions	Program Administration & Management	Program Environment	Family & Community Partnerships	Assessment of Child Progress	Professional Qualifications: Director	Professional Qualifications: Staff
Arizona	X	X	X	X	X	X	X	X	X
California	X		X	X	X		X	X	X
Delaware	X	X	X	X	X	X	X	X	X
Maryland	X	X	X	X	X	X	X	X	X
Massachusetts	X	X	X	X	X	X	X	X	X
Minnesota	X	X	X		X	X	X	X	X
Oregon	X	X	X	X	X	X	X	X	X
Rhode Island	X	X	X	X	X	X	X	X	X
Washington	X	X	X	X	X	X	X	X	X
Wisconsin	X	X	X	X	X	X	X	X	X
Total	**10**	**9**	**10**	**9**	**10**	**9**	**10**	**10**	**10**

Source: Individual state validation reports

Rating Structure

As noted, QRIS rating structures are designed as points, blocks, or a hybrid that uses both points and blocks. The rating structure and placement of quality indicators can affect the distribution of programs in a QRIS. In general, QRIS with points and hybrid structures tend to have a greater number of programs at the higher levels, most likely because there are multiple options to earn points even if some indicators are more challenging for programs to meet. Block structures, by contrast, tend to produce ratings at lower levels because one or more areas of weakness prevent a program from attaining a higher rating (Tout et al, 2014).

Six of the QRIS included in the synthesis (Arizona, California, Delaware, Minnesota, Washington, and Wisconsin) use a hybrid structure, while four (Maryland, Massachusetts, Oregon, and Rhode Island) use block structures. The QRIS in Massachusetts and Minnesota have four levels, while the other states' QRIS have five.

Summary of QRIS Similarities and Differences

The QRIS in the 10 states designate four or five levels of quality ratings using a block structure or a hybrid of blocks and points. The indicators of quality used in each QRIS tap common domains of quality, including supports for child health and wellness, staff qualifications, and child assessment. Most QRIS also include indicators of curriculum and teacher-child interaction, program environment, and family partnerships. However, the way that indicators are operationalized and the requirements to meet the quality indicators differ across systems. As a result, ECE programs from different states with the same rating might have demonstrated very different levels of quality. In addition, the systems vary in their stage of implementation and penetration into the ECE system in each state. The similarities and differences in the QRIS included in the synthesis provide an important backdrop for examining the validation study findings.

Methods Used in Validation Studies

Nine of the ten validation studies included in the synthesis were conducted in state ECE systems at a time of rapid change due to implementation of the Race to the Top – Early Learning Challenge grants. The studies accommodated many dynamic factors, including participation rates in the QRIS, distribution of programs across the rating levels, and (in some cases) changes or clarifications to the QRIS indicators that make up the rating. Researchers also encountered challenges faced in other studies of ECE quality, including program recruitment into the study, data collection on such important factors as children's demographic characteristics and attendance that are needed in statistical analyses, and possible selection bias of children with different characteristics using differently rated programs. This section describes the methods used in the validation studies and similarities and differences in how the studies addressed particular needs and challenges.

Variation in Data Collection Strategies

Types of programs included. All state studies included center-based community child care programs and Head Start programs. Eight of the ten states included family child care programs (all but MA and RI), and eight included school-based pre-kindergarten programs (all but MA and WI). Decisions about which types of programs to include or exclude were based on practical considerations such as budget, timeline, and number of programs participating in the QRIS, which would affect recruitment. It is important to remember that the results may not be generalizable to program types that were excluded from the study or that did not participate in the QRIS more generally.

In addition to deciding which types of programs to include, each study had to decide how to handle variation of program types in the analyses. All states that included Head Start and/or school-based pre-kindergarten programs combined them with community-based child care for analyses. Different states handled program type (center-based and family child care programs) differently. For observational quality measures, four of the seven states that included both center-based and family child care programs (Arizona, Maryland,

Minnesota, and Oregon) presented data separately by program type (Arizona and Oregon also presented data with program types combined); Wisconsin analyzed data with program types combined; Delaware used program type as a covariate; and California analyzed center data only due to the low number of family child care programs in its sample (descriptive statistics were provided). For child development analyses, four of the five states that included both center-based and family child care programs (CA, MN, WA, and WI) combined data from centers and family child care programs. Wisconsin did so after conducting a robustness check to confirm that the pattern of results was the same for both types of providers. Delaware used program type as a covariate in the child outcome data, as they did for the observational quality data.

Validation sample sizes. To address the two main validation questions, the studies recruited ECE programs participating in the QRIS, as well as preschool-aged children and sometimes infants/toddlers attending those programs (see Table 7). The number of center-based programs in the study samples ranged from 71 in the smallest state (Rhode Island) to 774 in Arizona, where data were drawn from both observations conducted by the research team and data maintained by the QRIS. Family child care programs generally involved a smaller sample, ranging from 24 in Washington to 159 in Oregon. Massachusetts and Rhode Island did not include family child care in their validation studies. The size of the child samples ranged from 332 in Rhode Island to 1,611 in California.

Table 7. Program and Child Sample Sizes by Validation Study

	Centers	Family Child Care	Preschoolers	Infants/ Toddlers
Arizona	774	148	n/a	n/a
California	175	47	1,611	n/a
Delaware	139	42	1,123 children	
Maryland	257	98	n/a	n/a
Massachusetts	126	n/a	481	190
Minnesota	278	66	1,181	n/a
Oregon	153	159	n/a	n/a
Rhode Island	71	n/a	332	n/a
Washington	76	24	522	239
Wisconsin	122	35	887	n/a

Source: Individual state validation reports

Ages and numbers of classrooms observed. States varied in their decisions about which classrooms (and how many) to include in the observational portion of the study. Eight states (all except MD and WI) included both preschool and toddler classrooms; two (MA and WA) included infant classrooms as well; and two included only preschool classrooms (MD and WI). This decision was likely based on practical considerations, such as budget and timeline, as well as the interests of the state QRIS leaders. Different observations tools, along with additional training and monitoring, are required for classrooms serving children of different ages, so including a broad age range has implications for the study resources. The number of classrooms observed at each age level also varied. For example, in Oregon and Wisconsin, up to four classrooms per program were observed to ensure the best possible representation of program quality. In most states, however, in programs with more than one classroom, one or two classrooms were randomly chosen to participate in the study.

Ages of children assessed. Of the seven states that included assessment of children's skills for the present synthesis,[19] only one (WA) included children of all age groups (infants, toddlers, and preschoolers). Delaware and Massachusetts included toddlers and preschoolers (2- through 5-year-olds). The remainder included

[19] Oregon's analyses examining QRIS ratings and child development will be included in a forthcoming report.

only children in the year prior to kindergarten (i.e., 4-year-olds) or children who would likely enroll in kindergarten in one or two years (i.e., 3- and 4-year-olds). Measuring children's development is challenging, particularly with children younger than age three. Additionally, conducting assessment of infants and toddlers requires extra training, time, and cost. Evaluation teams may have decided that their limited resources were best deployed by focusing on older children, where assessments tend to be more reliable.

Child assessment languages. Early childhood programs increasingly serve multilingual children. Four (DE, WA, WI, and CA) of the six studies that conducted child assessments assessed children in Spanish as well as English, if Spanish was their home language. In Minnesota, children who did not pass the English language screener received an abbreviated battery of assessments that was not language-dependent. English and Spanish assessments cannot be combined in the analyses and most states had too few Spanish-language assessments to analyze them separately. Accordingly, all but one state (CA) that assessed children in Spanish excluded those assessments from their analyses. Additionally, Washington reported the pre-/post-test means for children assessed in Spanish, but did not analyze the association between rating and child development of the children assessed in Spanish.

Study recruitment rates. Study recruitment rates represent the percentage of programs successfully recruited into the study divided by the total number of programs researchers attempted to contact and recruit. Four states (AZ, DE, MN, and RI) had challenges with program recruitment, resulting in low study recruitment rates (less than 40%). Becoming rated was a time-intensive, challenging process for many programs. It may be that program directors did not want to ask their staff to participate in these optional validation studies, given the amount of time and energy they had recently devoted to becoming rated. Three states (MD, OR, and WI) had recruitment rates over 60 percent. California and Washington recruitment rates were around 45 percent (see Table 8). Massachusetts' recruitment rate was unavailable. Only one state (RI) reported the parent consent rate, so it is not possible to know how well the children in these studies represent the classrooms they attend.

Table 8. Program Recruitment Rates by Validation Study

	Recruitment Rate
Arizona	32%
California	46%
Delaware	25%
Maryland	79%
Massachusetts	Not available
Minnesota	32%
Oregon	73%
Rhode Island	Levels 1-2 25%, Levels 3-5 39%
Washington	44%
Wisconsin	63%

Source: Individual state validation reports

Variation in Analysis Strategy

In addition to data collection, states varied in their approach to data analysis. Although they each addressed a similar set of questions from a similar set of data, they varied in how they addressed issues of statistical power and between-group comparisons, representativeness of the data and weighting, missing data, multiple tests, and change over time.

Statistical power and between-group comparisons. None of the studies had enough participating programs in each rating group to compare every level to every other level (whether due to low QRIS enrollment, low recruitment response rates, and/or a decision based on research costs). Only four states (AZ, CA, DE, and MN) presented a formal power analysis or even referenced having done one. The Arizona and Delaware studies explicitly discussed how their sample sizes were determined and the implications for analyses. Other states conducted power analyses when designing their studies or their analysis plans, but did not present them in their final reports. Small sample sizes from some or all groups were noted as a limitation in most state reports.

A sufficiently high level of program participation in QRIS is important for validation studies, which need adequate sample sizes, preferably across each of the rating levels. Even with sufficient participation in the QRIS, researchers need a good response rate to ensure that study samples are large enough to have the statistical power to detect differences in program quality and children's development across QRIS levels. In many validation studies, there were insufficient samples of programs available at each QRIS level to compare each of the 4 or 5 levels in the system to every other level. Moreover, comparisons across all levels of a system require a much larger sample size than comparisons that combine across levels, and for some state validation projects the costs of such a large sample size were prohibitive. Researchers addressed this in various ways.

In four state validation studies, the lowest level or levels of the QRIS were not included in the validation analyses (see Table 9). The exclusion of the lowest level(s) may be due to low numbers of programs rated at that level, or to the nature of the rating. For example, in Delaware, programs do not start earning points on QRIS indicators until level 3. To participate in level 1, a program only needs a license in good standing. At level 2, they attend an orientation session, meet with a TA provider, and set a program improvement plan. Washington could not include level 5 in its validation analyses due to a low number of programs rated at this level (n = 1) at the time of study recruitment.

Table 9. QRIS Levels Included and Excluded from Validation Analyses by State

	Number of Levels Included in Validation Analyses	Levels Excluded from Validation Analyses
Arizona	5	none
California	3	levels 1 and 2
Delaware	4	level 1 (starting with stars)
Maryland	5	none
Massachusetts	4	none
Minnesota	4	none
Oregon	5	none
Rhode Island	5 (no comparison of levels)	none
Washington State	3	levels 1 and 5
Wisconsin	4	level 1

Source: Individual state validation reports

Three states combined programs into a lower- and a higher-quality group for comparison purposes. This combination across rating levels allowed for more statistical power to address the research questions. The methods used to combine rating levels address the question of whether QRIS ratings distinguish between

levels of program quality, but cannot inform whether *every* level in a QRIS distinguishes levels of quality relative to every other level.

Minnesota did the analyses by combining 1- and 2-star programs into a lower-quality group, and 3- and 4-stars into a higher-quality group. Oregon compared programs at levels 1 and 2[20] to programs at levels 3–5. Wisconsin compared lower-quality (2 star) to higher-quality (3-5 star) programs but also compared all star levels to one another. Arizona created three quality groups (programs at levels 1 and 2, those at level 3, and those at levels 4 and 5). Studies in California and Delaware compared across each rating level. Rhode Island did not have enough programs at the higher star levels, so instead treated the star rating as a continuous variable and analyzed whether higher star ratings were associated with higher quality. Washington did not combine rating levels in its analyses.

Representativeness of the sample and weighting. The Delaware study was the only validation study to test whether the programs and children included in the study were similar to those not included. Programs in the study were compared to those not in the study using administrative data, and information about children in the study was compared to other children using Census data. Furthermore, the Delaware researchers created a set of sampling weights to adjust for ways in which programs and children in the study differed systematically from the larger population and applied them to all analyses, ensuring that results accurately reflect the population. No other state used this approach.

Handling missing data. Missing data, or information that the project intended to collect but did not, is inevitable in research studies. This can result from teachers or children electing not to respond to individual questions, scheduling problems that result in some information not being collected, or simple human error resulting in some responses not being recorded correctly. Five of the validation studies discussed missing data and how they were handled: California, Maryland, Minnesota, Washington, and Wisconsin. Maryland determined that the characteristics of programs with data missing across data collection cycles did not bias the study results. The other three states used a statistical technique called imputation to fill in missing values without biasing the results. Oregon indicated that there was no missing data in their study linking rating to observed quality.

Adjusting for multiple tests. Most statistical analyses provide an estimate of the probability that the result came about by chance. When that probability is less than 5 percent ($p < .05$), the result is often referred to as *statistically significant*. When many separate statistical analyses are being conducted, some statisticians argue that the level at which a difference is considered statistically significant should be adjusted to make it more conservative. Others argue that such adjustments are unnecessarily restrictive and cause researchers to miss important associations (see Gelman, Hill, & Yajima, 2011 for a discussion). One of the seven studies (DE) elected to statistically adjust for multiple comparisons.

Change over time. Because parents select ECE settings for their children, children with different characteristics are enrolled in different types of care. For instance, families with more resources may send their children to higher-quality care because they can pay higher fees. Those same children may enter the setting with more advanced skills because they have had other advantages, such as a language- and literacy-rich home environment (Lamb, 1998; Neuman & Dickinson, 2010). To account for any differences among children that existed before the studies began, all seven studies examining children's development included child-level assessments at the start of the school year.

Pre-test assessments can be included in the analyses in various ways. Four studies elected to estimate statistical models that predict spring assessment scores, *controlling for* fall assessments scores. Two states, Massachusetts and Minnesota, calculated *gain scores* by subtracting the fall values from the spring values and estimating statistical models that predict those differences. In addition, Massachusetts created groups

[20] Oregon selected a subset of level 1 programs that were unlikely to achieve a QRIS rating based on personnel education and training data. The state also included only level 2 programs that applied for a rating of 3 or higher but did not achieve it. Therefore, the state had confidence that its sample of level 1 and 2 programs was "lower quality" relative to the pool of programs that were not fully rated.

equivalent at fall testing across rating levels and controlled for fall scores when examining gain scores. Wisconsin used both approaches and found that the findings were the same.

These approaches make differing assumptions about how measurement error should be handled. The gain score approach includes an explicit measure of growth and eliminates unobserved confounding variables, thereby reducing bias due to unobserved factors. However, it also results in lower variance in the outcome variable, resulting in less power to detect associations.

The approach that involves controlling for fall scores results in greater variance in the outcome variable than the gain score approach, and therefore provides more power to detect associations. In addition, this approach can be used when different assessments were used in the fall and spring. However, the approach does not include an explicit measure of development, so it can result in biased estimation if unobserved background factors have a different impact on fall and spring assessment scores.

Common Measures

A variety of tools were used in the validation studies to measure observed quality and child development, and the battery of tools used varied by state. In several cases, validation researchers used the same outcome measures for observed quality and child development as researchers in other states, allowing for an examination of results across studies.

Evaluators in all states except Washington conducted observations of quality in programs and examined the associations between global quality, teacher-child interaction, and/or other aspects of quality and QRIS ratings. Some common measures were used in two or more states and findings can be compared. Table 10 shows the measures of quality used in each study (see Appendix for more details on all tools).

Table 10. Quality Measures Used in QRIS Validation Studies by State

QRIS	ECERS-R	ECERS-E[a]	ECERS-3	FCCERS-R[b]	CLASS Pre-K	CLASS Toddler	CLASS Combined	PQA[c]	STARE[d]	CIS[e]
Arizona	X		X	X	X	X				
California					X	X		X		
Delaware					X	X		X		X
Maryland	X			X	X	X			X	
Massachusetts	X									X
Minnesota	X	X		X	X					
Oregon					X	X	X			
Rhode Island					X	X				
Wisconsin	X			X						
Used by two or more	**X**			**X**	**X**	**X**		**X**		**X**

Source: Individual state validation reports
[a] Extension to the Early Childhood Environment Rating Scale (Sylva, Siraj-Blatchford, & Taggart, 2011)
[b] Family Child Care Environment Rating Scale
[c] Program Quality Assessment
[d] Scale for Teachers' Assessment of Routines Engagement
[e] Caregiver Interaction Scale

The Early Childhood Environment Rating Scales Revised Edition (ECERS-R), Family Child Care Environment Rating Scale Revised Edition (FCCERS-R), CLASS Pre-K, CLASS Toddler (LaParo, Hamre, & Pianta, 2012), Program Quality Assessment (PQA), and Caregiver Interaction Scale (CIS) were all used in at least two of the validation studies synthesized here.

The studies in California, Delaware, Massachusetts, Minnesota, Rhode Island, Washington, and Wisconsin all examined the association between QRIS ratings and child development. Evaluators conducted child assessments with preschoolers and/or toddlers in the fall and spring in multiple domains of development. Some common measures were used in two or more states and findings can be compared. Tables 11–15 show the measures of child development used in each study.

Table 11. Child Assessments Used in Validation Studies by State: Language and Literacy

QRIS	PPVT[a]	TOPEL[b]	IDGI[c] Picture Naming	WJ-III[d] Letter-Word	WJ-III Picture Vocabulary	Story and Print Concepts[e]	Mullen Expressive Language[f]	Mullen Receptive Language
California				X		X		
Delaware	X			X				
Massachusetts	X			X				
Minnesota		X	X					
Rhode Island				X	X			
Washington	X			X			X	X
Wisconsin		X		X				
Used by two or more	**X**	**X**		**X**				

Source: Individual state validation reports
[a] Peabody Picture Vocabulary Test
[b] Test of Preschool Early Literacy
[c] Individual Growth and Development Indicators (Early Childhood Research Institute on Measuring Growth and Development, 1998)
[d] Woodcock Johnson III Tests of Achievement
[e] Story and Print Concepts (Zill & Resnick, 2000)
[f] From Mullen Scales of Early Learning (Mullen, 1995)

Table 12. Child Assessments Used in Validation Studies by State: Math Skills

QRIS	Tools for Early Assessment in Math (TEAM[a])	WJ-III Applied Problems
California		X
Delaware		X
Massachusetts		X
Minnesota		X
Rhode Island		X
Washington	X	
Wisconsin		X
Used by two or more		X

Source: Individual state validation reports
[a] Clements & Sarama, 2011

Table 13. Child Assessments Used in Validation Studies by State: Executive Function and General Cognition

QRIS	Peg Tapping	Head, Toes, Knees, and Shoulders	Bracken	Mullen Scales of Early Learning
California	X			
Delaware		X		
Minnesota	X		X	
Rhode Island	X			
Washington		X		X
Wisconsin		X	X	
Used by two or more	**X**	**X**	**X**	

Source: Individual state validation reports

Table 14. Child Assessments Used in Validation Studies by State: Physical Development

QRIS	Body Mass Index (BMI)	Mullen Fine Motor	Mullen Gross Motor
California			
Delaware			
Minnesota	X		
Washington		X	X
Wisconsin			
Used by two or more			

Source: Individual state validation reports.

Table 15. Child Assessments Used in Validation Studies by State: Social/emotional Development

QRIS	SCBE-30[a]	PLBS[b]	DECA[c]	CBCL[d]
California				
Delaware			X	
Massachusetts		X	X	
Minnesota	X	X		
Rhode Island	X	X		
Washington				X
Wisconsin	X	X		
Used by two or more	**X**	**X**	**X**	

Source: Individual state validation reports
[a] Social Competence and Behavior Evaluation
[b] Preschool Learning Behaviors Scale
[c] The Devereux Early Childhood Assessment
[d] Child Behavior Checklist (Achenbach & Edelbrock, 1983)

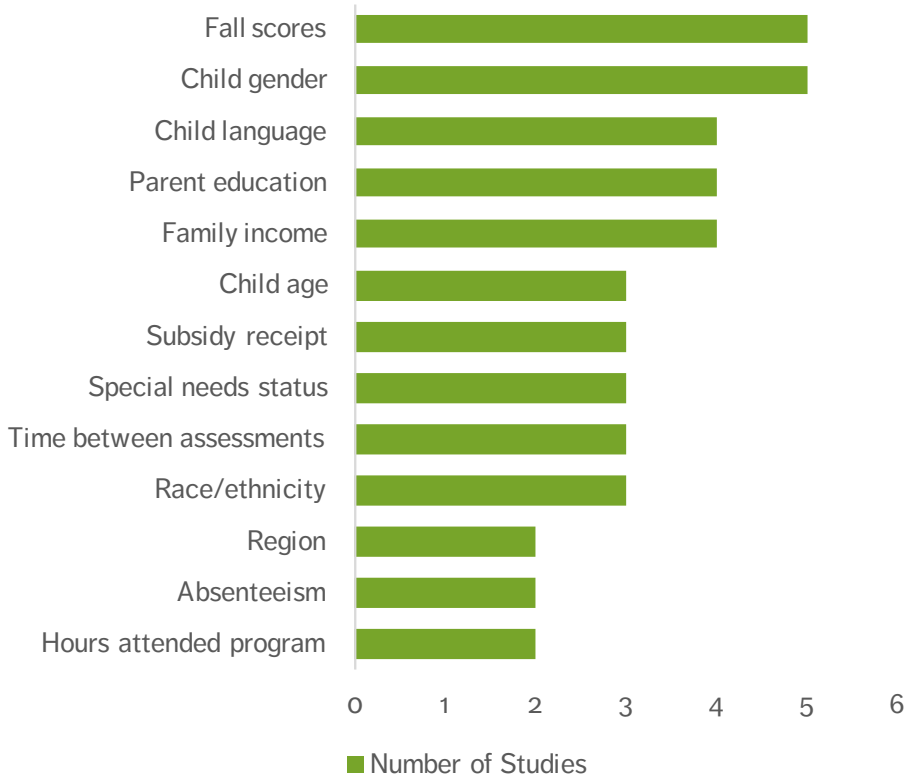
In addition, Washington used Lens on Science (LENS) and the Early Writing Assessment (EWA). Several child assessment tools were used in multiple validation studies to measure language/literacy (PPVT [Dunn & Dunn, 2007], TOPEL [Lonigan et al., 2007], WJ-III Letter-Word), executive function and cognition (peg tapping [Diamond & Taylor, 1996], Head, Toes, Knees, and Shoulders, Bracken [Bracken, 2007]), social/emotional development (SCBE-30 [LaFreniere & Dumas, 1996], PLBS [McDermott, Leigh, & Perry, 2002]), and math skills (WJ-III Applied Problems).

Control Variables

When examining the association between QRIS ratings and child development, it is important to consider other factors (such as family income and parent education) that are likely to affect a child's development. Many family or child characteristics have the potential to affect children's developmental skills and therefore must be accounted for in the analyses. Researchers statistically account for the effects of family or child variables by adding them as covariates to the analysis models.

Of the seven validation studies that examined the association between QRIS rating and child development, the family and child factors most commonly controlled for in the analyses were gender and children's fall assessment scores (five states). Four states controlled for family income, parent education, and child language; three controlled for race/ethnicity, time between fall and spring assessments, special need/disability status of children, subsidy receipt, and child age. Other variables included in some studies were hours children are scheduled to attend care, absenteeism, and regional characteristics (see Figure 3).

Figure 3. Control Variables Used in the Analyses Examining QRIS Rating and Child Development in Seven Validation Studies



Source: Individual state validation reports

In synthesizing results across states, it is important to note other factors that could have affected the findings in each study.

The next two sections of the synthesis provide information about findings from the state studies. The results are presented by the two research questions: ratings and observed quality and ratings and children's development.

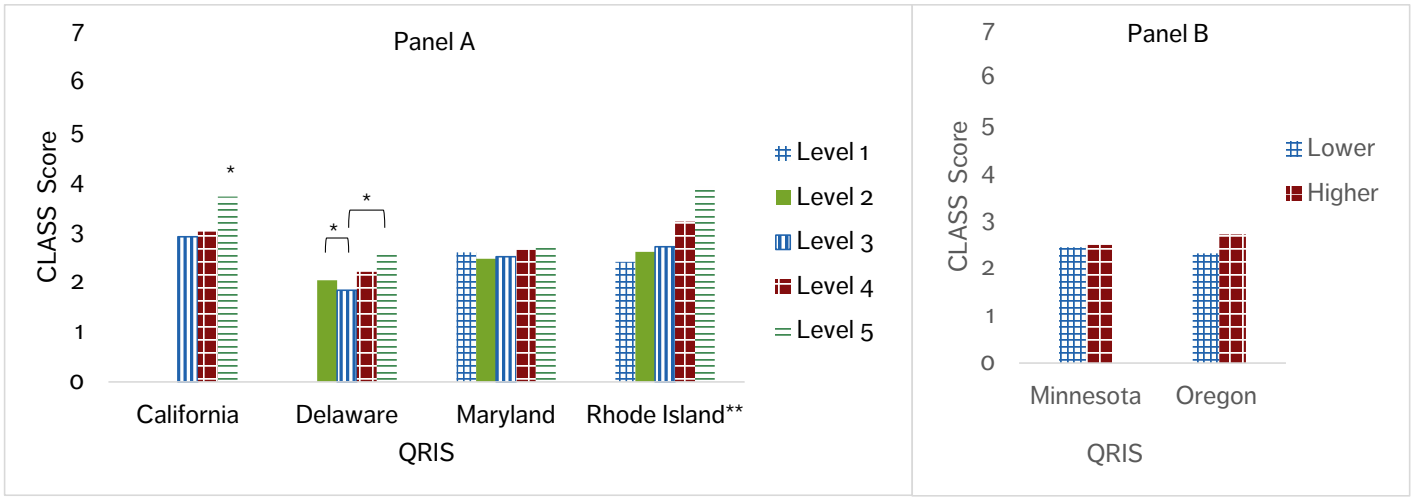
QRIS Ratings and Observed Quality

The purpose of QRIS ratings is to provide a useful, simple metric of tiered quality that can be used as an accountability measure in ECE systems (e.g., to promote and reward quality) and to support parent decision making. Establishing that QRIS ratings are associated with measures of observed quality provides some confidence that the ratings are distinguishing quality in a meaningful way. As noted, several different tools were used in the studies to measure observed quality in classrooms and programs at each QRIS level, and to examine differences in scores across levels or combinations of levels. The Classroom Assessment Scoring System (Pianta et al., 2008; CLASS Pre-K and CLASS Toddler) Environment Rating Scales (Harms et al. 2005; 2007; ECERS-R and FCCERS-R) and Program Quality Assessment (Highscope, 2003) were all used in at least two of the validation studies included in this synthesis.

Observed Quality across States

The CLASS-PreK was used by studies in California, Delaware, Maryland, Minnesota, Oregon, and Rhode Island. Arizona used both the CLASS Pre-K and CLASS Toddler but combined scores from the two instruments in reporting results (Arizona's data are described in the text but not included in the figures). Scores on the CLASS for each domain by QRIS level (or comparison group: higher/lower) for each of the other states are presented in Figures 4-6.

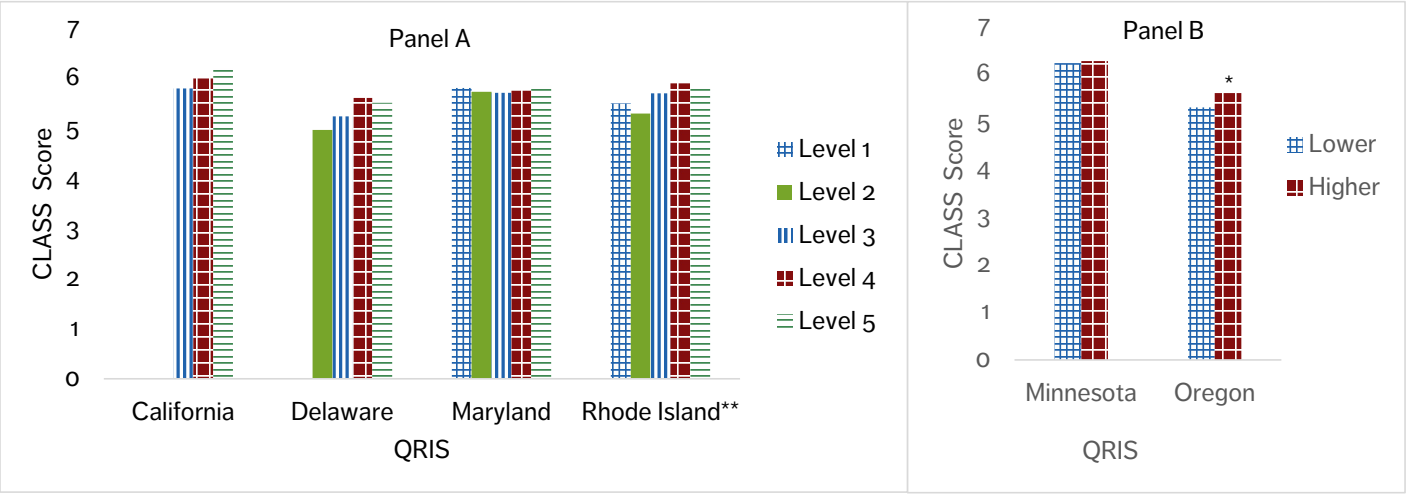
Figure 4. CLASS Pre-K Instructional Support Scores by QRIS Level



Source: Individual state validation reports
*Indicates significant difference between level and at least one other level
**Indicates general linear trend of higher ratings having higher scores

Three states found significant differences in CLASS Instructional Support scores across QRIS levels. In California, Level 5 programs scored higher than Level 3 and Level 4 programs. In Delaware, Level 5 was greater than Level 3 and Level 4, but Level 2 was greater than Level 3 (this is not unexpected because programs do not start earning points through ERS scores across domains until they try for Level 3). In Rhode Island, sample sizes were too small to conduct direct comparisons among QRIS levels. However, regression analyses indicated a significant positive association between QRIS level and CLASS Instructional Support (a less stringent approach than direct comparisons).

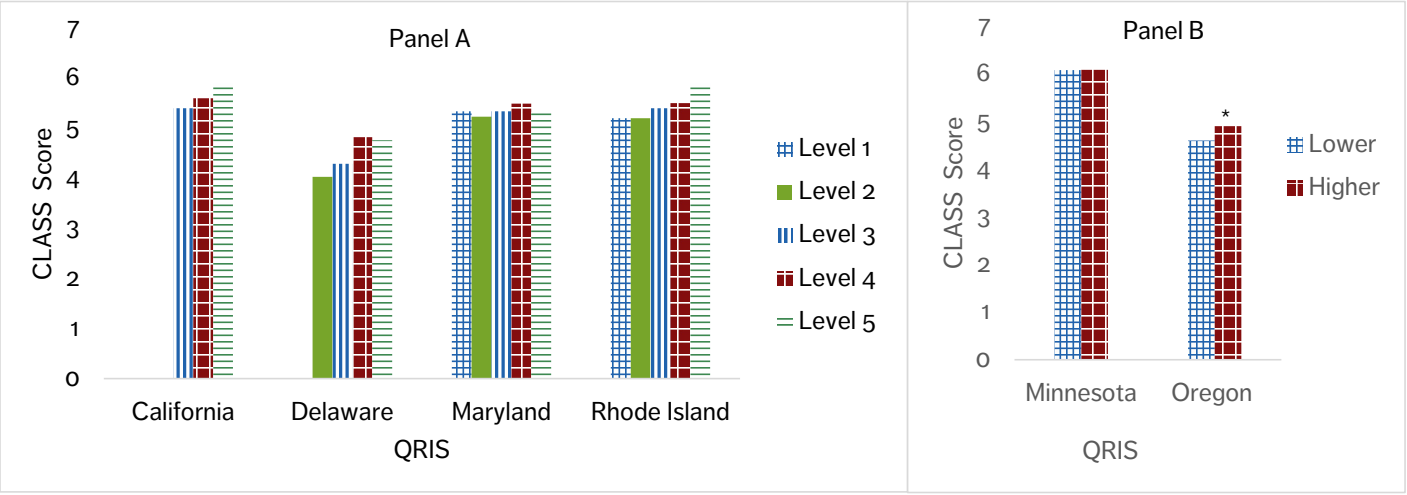
Figure 5. CLASS Pre-K Emotional Support Scores by QRIS Level



Source: Individual state validation reports
*Indicates significant difference between highest level and at least one lower level
**Indicates significant general linear trend of higher ratings having higher scores

In Oregon, programs with higher ratings scored significantly higher than lower-rated programs on Emotional Support. In Rhode Island, regression analyses indicated a significant positive association between QRIS level and CLASS Emotional Support. No other studies had significant differences in Emotional Support across quality levels.

Figure 6. CLASS Pre-K Classroom Organization Scores by QRIS Level

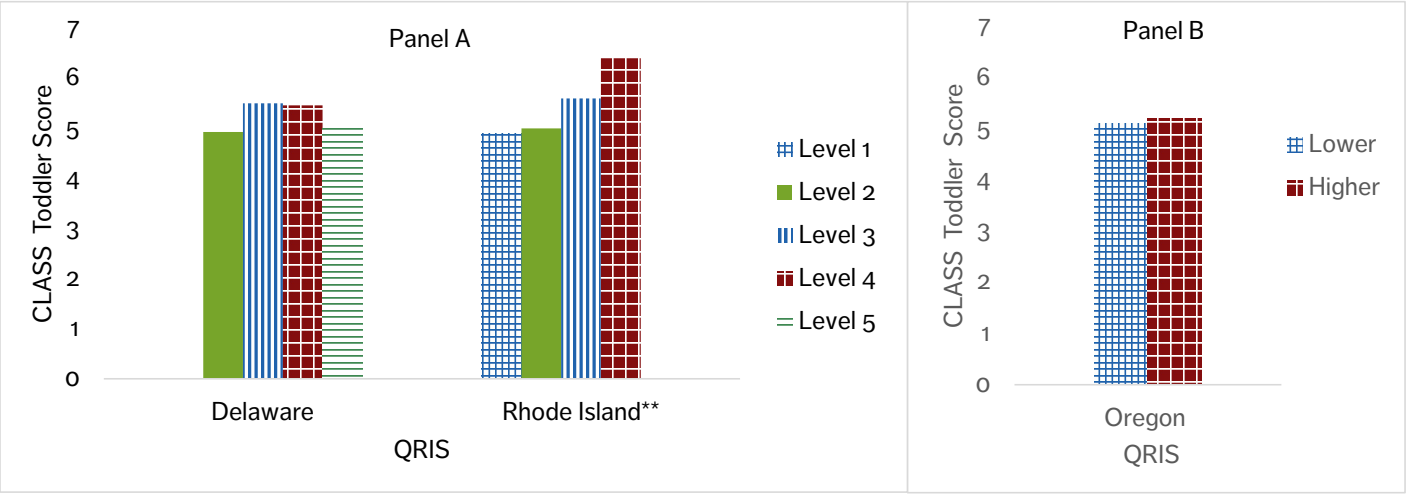


Source: Individual state validation reports
*Indicates significant difference between highest level and at least one lower level
**Indicates significant general linear trend of higher ratings having higher scores

In Oregon, programs with higher ratings scored significantly higher than lower-rated programs on Classroom Organization. In Rhode Island, regression analyses indicated a significant positive association between QRIS level and CLASS Classroom Organization. No other studies had significant differences in Classroom Organization across quality levels.

The CLASS Toddler was used in Delaware, Oregon, and Rhode Island. Scores across QRIS levels are presented in Figures 7 and 8. California also used the CLASS Toddler, although small sample sizes did not allow for statistical comparison across QRIS level.

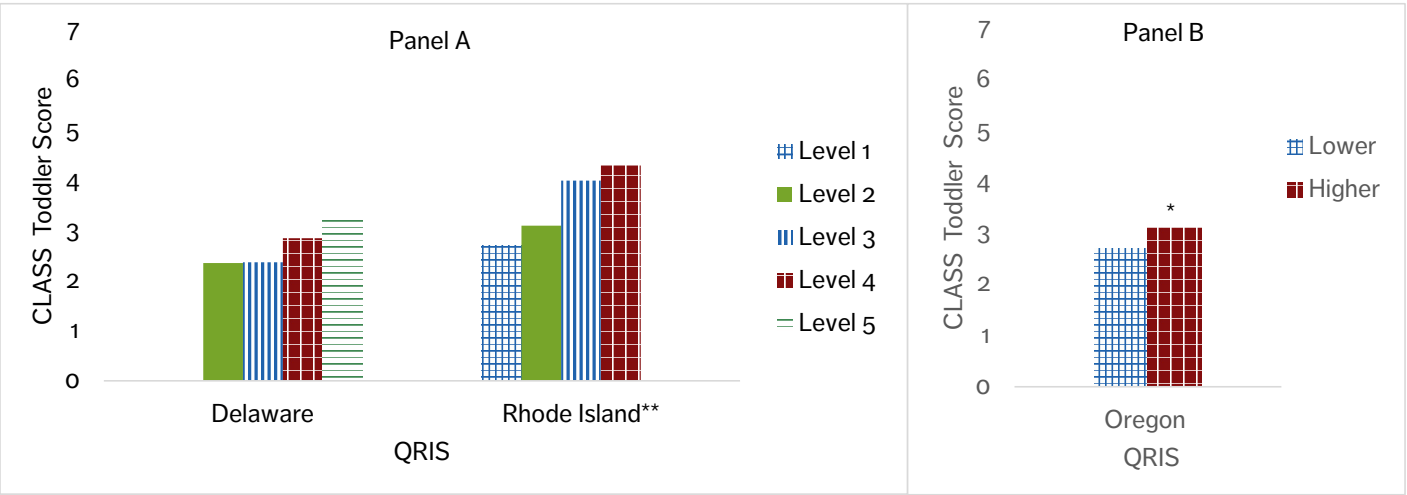
Figure 7. CLASS Toddler Emotional/Behavioral Support Scores



Source: Individual state validation reports
**Indicates significant general linear trend of higher ratings having higher scores

In Rhode Island, CLASS Toddler Emotional and Behavioral Support (ESB) scores were significantly associated with QRIS level. There were no significant differences in ESB scores across quality levels in Delaware or Oregon.

Figure 8. CLASS Toddler Engaged Support for Learning Scores



Source: Individual state validation reports
*Indicates a significant difference between high and low
**Indicates significant general linear trend of higher ratings having higher scores

Programs with high ratings scored significantly higher on the CLASS Toddler Engaged Support for Learning than lower-rated programs in Oregon. In Rhode Island, CLASS Toddler Engaged Support for Learning scores were significantly associated with QRIS level.

Oregon also used the Combined CLASS (Joseph et al., 2010) in family child care programs with mixed age groups (family child care programs were also included in the PreK and Toddler CLASS when children's ages aligned with those tools). Programs with high ratings scored significantly higher than lower-rated programs on all Combined CLASS subscales (Emotional Support, Classroom Organization, and Instructional Support).

In Arizona, which reported CLASS Pre-K and Toddler results together, higher-rated programs (those at

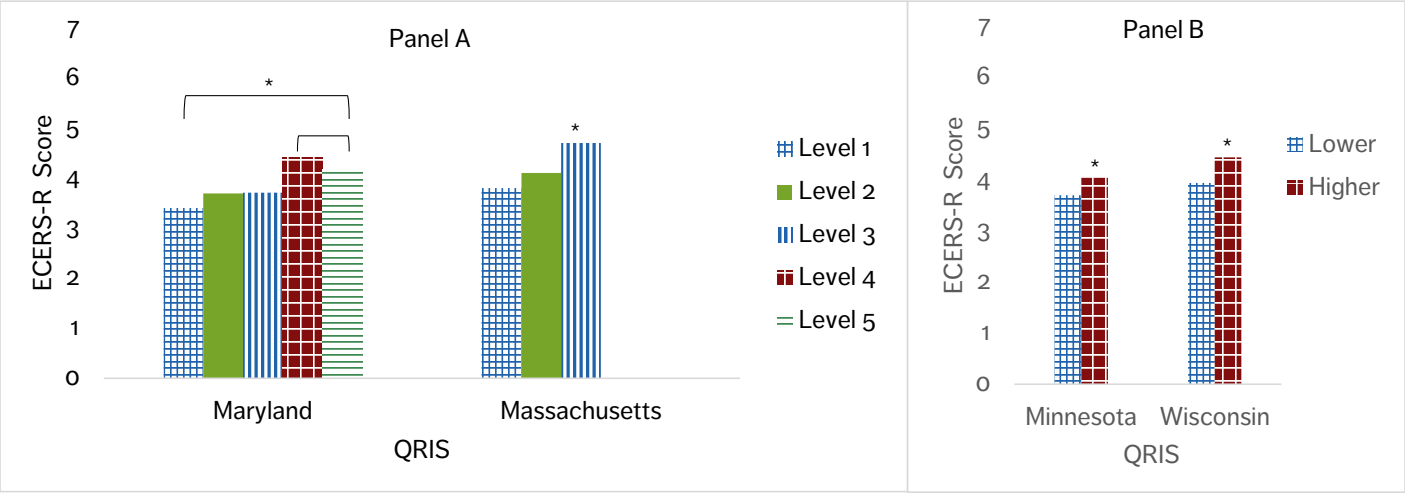
Levels 4 or 5) scored higher on Emotional Support and on Classroom Organization than programs at Level 3, which in turn scored higher than lower-rated programs (Levels 1 or 2). For Instructional Support, scores were higher for programs at Level 4–5 than Level 1–2, and both of those levels were higher than Level 3. The report authors provided possible explanations for this unexpected finding, including data collection discrepancies (the evaluators collected the data for programs at Levels 1 and 2 while the QRIS collected CLASS data for programs at higher rating levels) or the possibility that the CLASS Instructional Support scale may not be sensitive to reliably detect differences in low- and medium-quality classrooms.

The ECERS-R was used in the validation studies in Massachusetts, Maryland, Minnesota, and Wisconsin. ECERS-R scores by QRIS level are presented in Figure 9. In Massachusetts, programs at Level 3 had higher ECERS-R scores than those at Levels 1 and 2 (Level 4 programs were not included in the analysis due to small sample size). In Maryland, programs at Levels 4 and 5 had higher ECERS-R scores than those at Level 1. Both Minnesota and Wisconsin found a significant difference in ECERS-R scores between programs with low and high ratings. Massachusetts also collected ITERS-R data and found that programs rated at Levels 2 and 3 both had higher ITERS-R scores than Level 1-rated programs.

For the Arizona study, the research team collected ECERS-3 and used Quality First administrative data for ECERS-R and ITERS-R scores, as they are both collected as part of the QRIS rating. ERS scores drawn from these three instruments were analyzed by program rating level. In center-based programs, ERS scores were significantly higher for programs rated at levels 4–5 compared to level 3-rated programs, which in turn were higher than scores for programs rated at levels 1–2.

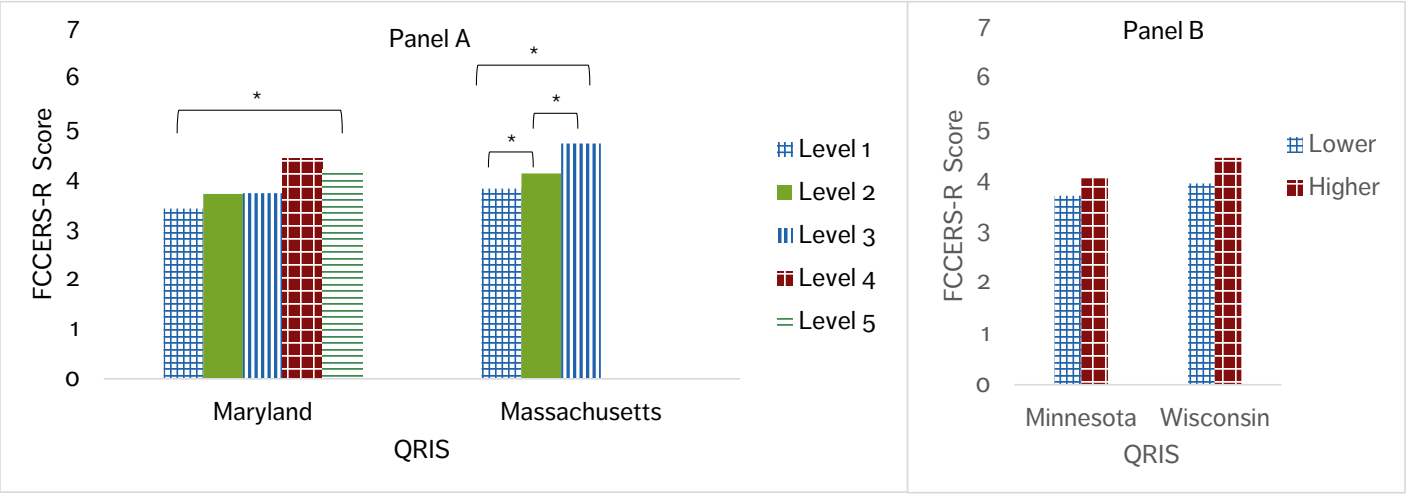
Arizona, Maryland, Minnesota, and Wisconsin used the FCCERS-R in family child care settings. These data are presented in Figure 10. In Arizona, higher-rated programs (those rated 4 or 5 stars) had significantly higher FCCERS-R scores than medium-rated (3-star) programs, which in turn had higher scores than lower-rated programs (1 or 2 stars). In Maryland, Level-5 programs had higher FCCERS-R scores than Level-1 programs (FCCERS-R was collected in very few Level-4 programs). Minnesota found no significant differences between family child care programs with low and high ratings. Wisconsin did not have large enough sample sizes to report findings.

Figure 9. ECERS-R Scores by QRIS Level



Source: Individual state validation reports
* Indicates significant difference

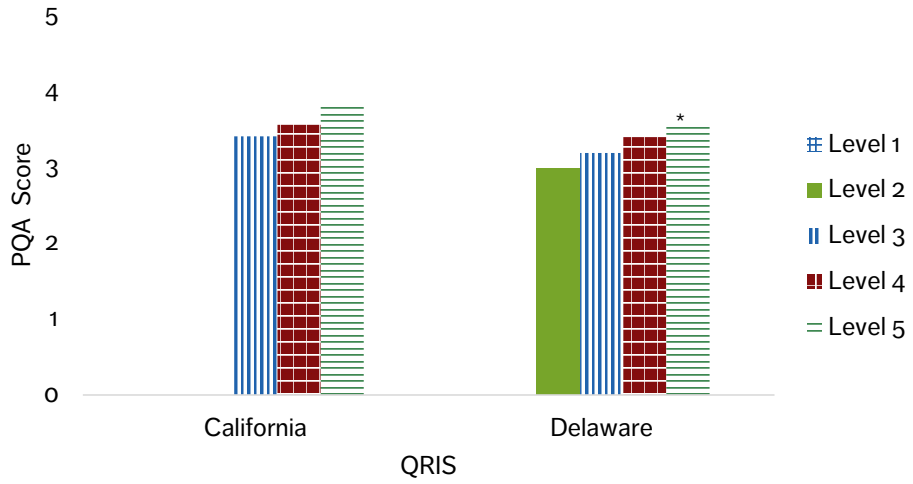
Figure 10. FCCERS-R Scores by QRIS Level



Source: Individual state validation reports
* Indicates significant difference

The Preschool Program Quality Assessment (PQA) was used by validation studies in California and Delaware. PQA total scores by QRIS levels are presented in Figure 11.

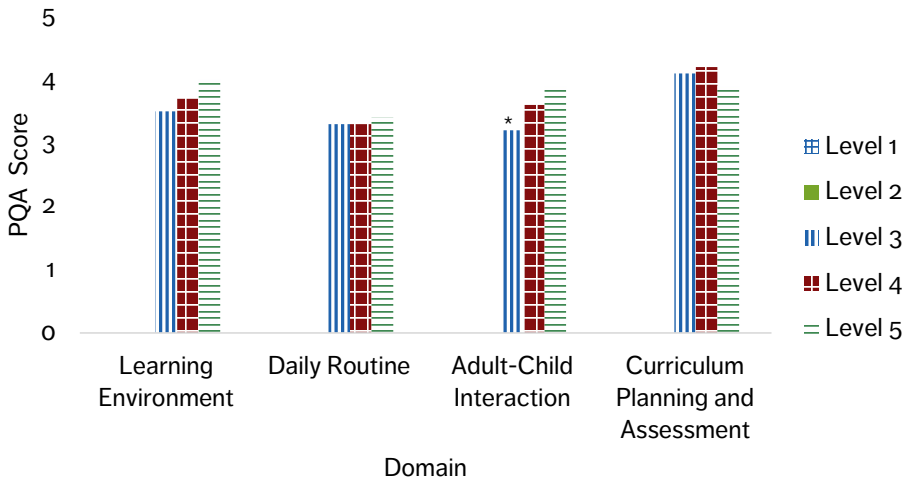
Figure 11. PQA Mean Score by QRIS Level



Source: Individual state validation reports
*PQA at level 5 is significantly higher than scores at levels 2, 3, and 4

The PQA total score was associated with QRIS level in Delaware, but not in California. However, when California assessed the association between QRIS level and PQA by domain, researchers found significant differences between tier 3 and tiers 4 and 5 for adult-child interaction (see Figure 12).

Figure 12. California PQA Mean Scores by Domain and QRIS Level



Source: Individual state validation reports
*PQA Adult-Child Interaction means were significantly lower at level 3 than levels 4 or 5

The Caregiver Interaction Scale (CIS) was used in Delaware and Massachusetts. In Massachusetts, CIS scores in preschool classrooms were significantly higher in level 3 programs than in level 2 programs. CIS scores did not differ across levels in toddler classrooms. Delaware found no significant differences across QRIS level in CIS scores.

Summary of QRIS Ratings and Observed Quality across States

Table 16 summarizes findings on the associations between QRIS ratings and observed measures of quality.

Table 16. Summary of Associations between QRIS Ratings and Observed Quality, by State and Observational Measure

	CLASS Pre-K			CLASS Toddler		ERS			Other Quality Measures	
State	**Instructional Support**	**Emotional Support**	**Classroom Org.**	**Emotional and Behavior Support**	**Engaged Support for Learning**	**ECERS-R**	**ITERS-R**	**FCCERS-R**	**PQA***	**CIS**
Arizona	✓	✓	✓			✓		✓		
California	✓	*ns*	*ns*						✓	
Delaware	✓	*ns*	*ns*	*ns*	*ns*				✓	*ns*
Maryland	*ns*	*ns*	*ns*			✓		✓		
Massachusetts						✓	✓			✓
Minnesota	*ns*	*ns*	*ns*			✓		*ns*		
Oregon	✓	✓	✓	*ns*	✓					
Rhode Island	✓	✓	✓	✓	✓					
Wisconsin						✓		[1]		

Source: Individual state validation reports
*PQA total score for DE, Adult-Child Interaction for CA. [1] The FCCERS-R was collected in Wisconsin but analyzed jointly with the ECERS-R.
Note: A check mark indicates at least one statistically significant association was found demonstrating higher observed quality at higher rating levels. "Ns" indicates that no statistically significant associations were found. A gray, blank cell indicates that the measure was not collected.

Five of seven state studies found a significant association between QRIS level and CLASS Pre-K Instructional Support. For other CLASS domains (Emotional Support and Classroom Organization) and CLASS Toddler, findings were mixed. Arizona found significantly higher scores on the CLASS Pre-K (Emotional Support and Classroom Organization) for higher-rated relative to lower-rated programs. Oregon found significant differences between higher and lower rating levels on CLASS Pre-K (Emotional Support, Classroom Organization, Instructional Support) and CLASS Toddler Engaged Support for Learning. Rhode Island also found evidence for the CLASS Pre-K and Class Toddler but used a different type of analysis than the other states.[21] Maryland and Minnesota did not find any significant associations with CLASS Pre-K and ratings.

Arizona, Maryland, Massachusetts, Minnesota, and Wisconsin found a significant association between QRIS level and ERS scores. In all five states, ECERS-R was significantly higher in higher-rated center programs than in lower-rated programs. Arizona and Maryland also found significant differences across rating level on the FCCERS-R. In Minnesota, there was no evidence for differences in FCCERS-R scores across rating levels. The FCCERS-R was collected in Wisconsin but analyzed jointly with the ECERS-R.

Delaware and California also found a significant association between QRIS level and observed measures of quality with the PQA. Both studies found significantly higher PQA scores (either overall mean or the Adult-Child Interaction subscale) at the highest QRIS level than at lower levels. Massachusetts found significant differences in preschool classrooms on the CIS.

Although the validation studies found some significant differences in observed quality across rating levels, there was also overlap in the range of scores across programs at different rating levels. That is, two programs could have the same ERS or CLASS scores, but have different QRIS ratings. The range of program scores on measures of observed quality at different QRIS levels reflects the fact that many components go into a QRIS rating. Observed quality measures are not the only sources used to derive a rating. A program may be able to achieve a higher rating by scoring well on other aspects of the QRIS, even if scores on the observed quality measures are in the medium range.

In seven of the nine studies examining ratings and observed quality (all but Delaware and Rhode Island), at least one measure of quality used in the validation study was also included in the QRIS rating calculation itself (at least at some levels of the QRIS; see Table 1). Overall, five of seven validation studies that used the CLASS found at least one significant association between CLASS and QRIS level; CLASS was included in the rating for three studies that found significant associations and in two studies that did not find significant associations. Five of five studies that used the ERS found at least one significant association between ERS and QRIS level; ERS was included in the rating for four of the studies.

While the validation studies used measures included in the QRIS rating, six of the nine studies examining ratings and observed quality found significant associations using at least one independent quality measure.

QRIS Ratings and Children's Development

Decades of early childhood research documents the role of ECE program quality in supporting young children's development, although associations are modest (Burchinal et al., 2009). QRIS administrators and stakeholders have invested in QRIS in part because of the potential to improve program quality and ultimately enhance support for children's development. Examining the ratings produced by QRIS and their association with measures of children's development provides information about whether the ratings identify aspects of quality linked to children's developmental gains. The studies addressed this question by looking at gains over a relatively short time (about 6 months in the year before kindergarten for preschool children) and using standardized measures that account for age-related changes on the skills assessed.

[21] Researchers treated observation scores as continuous variables and tested for the association between QRIS level and quality outcomes.

Several different standardized tools were used in the studies to measure child development. Assessments included measures of language and literacy, math skills, executive function, physical development, and social/emotional development. The PPVT, TOPEL, WJ-III Letter-Word, peg tapping, Head, Toes, Knees, and Shoulders, Bracken, SCBE-30, PLBS, and WJ-III Applied Problems were all used in at least two of the validation studies synthesized here. In some cases, states computed the scores on a measure in different ways (e.g., standard score vs. raw score). The figures in this section group states together by comparable measures used.

The findings on children's development are presented in two parts. First, the average gains children make on developmental measures are described. These gains are calculated for all children, regardless of the QRIS rating of the program they attend. These descriptive analyses are included to provide context and information about the children in the state samples. Next, the results of analyses examining children's development by QRIS ratings are presented.

Child Development Gains from Fall to Spring

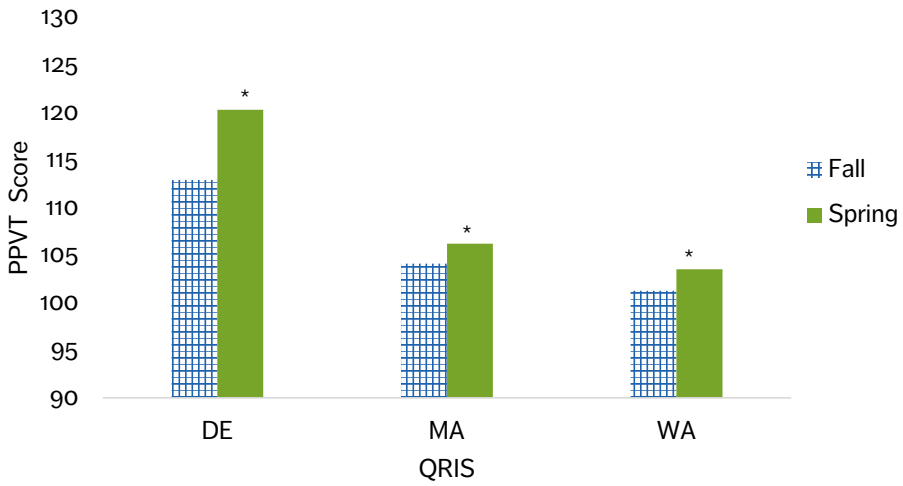
Many validation studies first examine any changes in child assessment scores across the school year (differences between fall and spring scores). Significant increases in mean scores from fall to spring indicate that children are improving in domains such as language/literacy, math skills, executive function, and social/emotional development when participating in QRIS-rated programs. California, Minnesota, Rhode Island, Washington, and Wisconsin all reported fall and spring average child assessment scores in some format. Wisconsin reported scores by star level only, and the remaining states presented overall fall/spring scores (CA presented scores both ways). Minnesota, Rhode Island, and Washington also reported significance tests for gains across the school year. Massachusetts collected child assessment data in the fall and spring but did not report significance tests for the gain across the school year. Data were obtained from researchers to facilitate comparison across states.

Most of the studies conducted simple paired t-tests (with no covariates) to determine whether spring scores were significantly different from fall scores. The exception was the Massachusetts study, which used Hierarchical Linear Models to determine differences between fall and spring scores while controlling for English language learner status, subsidy receipt (as a proxy for socioeconomic status), and special needs diagnosis. For social skills measures, the studies also controlled for child age because scores were not standardized. Most studies used standardized scores when available. The type of score used for each measure in each state is noted in the explanation of each finding in the sections that follow.

Language and literacy measures

Woodcock-Johnson letter-word (Woodcock, McGrew, & Mather, 2001), TOPEL phonological awareness, and the PPVT were used in multiple validation studies. Fall and spring scores for these measures by state are presented in figures 13 through 15. These measures are presented using standardized scores, such that scores for each age are distributed around a mean of 100 and a standard deviation of 15. Therefore, significant gains from fall to spring represent gains beyond what would be expected by age alone. One exception is that Delaware did not use standardized scores for the PPVT. They used growth scale values because they were interested in absolute growth rather than growth relative to age-normed growth.

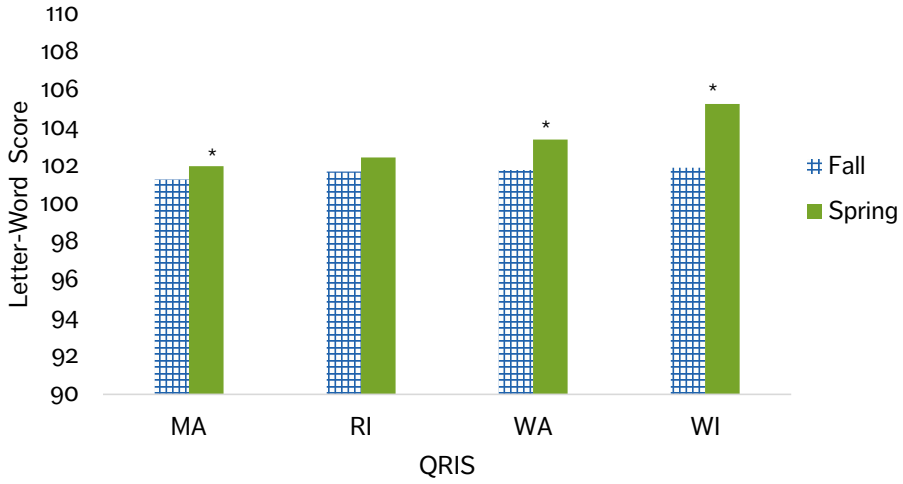
Figure 13. Fall and Spring PPVT Scores



Source: Individual state validation reports

Delaware, Massachusetts, and Washington all found significant growth from fall to spring on the PPVT.

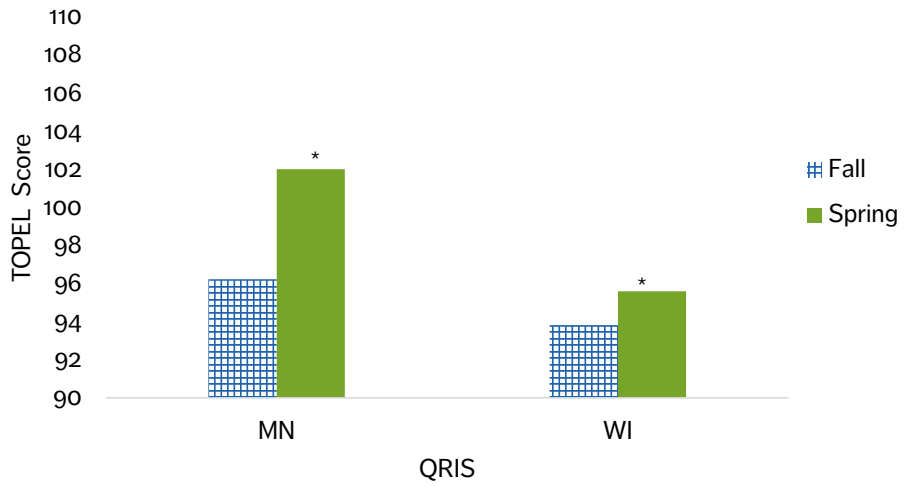
Figure 14. Fall and Spring Woodcock-Johnson: Letter-Word Scores



Source: Individual state validation reports

Gains in Letter-Word standardized scores in Massachusetts, Washington, and Wisconsin were significant, but not in Rhode Island. Significant gains in standard scores indicate that scores increased more than would be expected based on the amount of time that passed between assessments. Positive but nonsignificant gains, such as those in Rhode Island, indicate that children learned the amount that would be expected. California and Delaware reported significant fall to spring gains with age-equivalent Letter-Word scores.

Figure 15. Fall and Spring TOPEL Phonological Awareness Scores



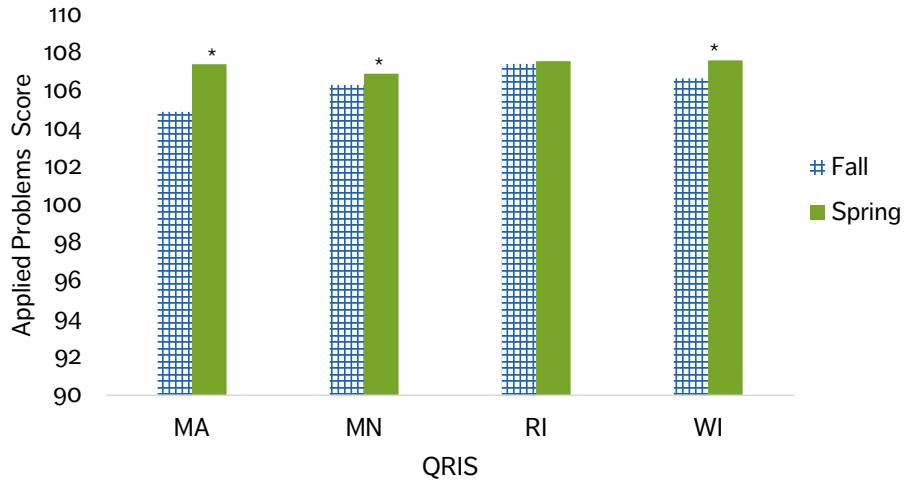
Source: Individual state validation reports

Both Minnesota and Wisconsin found significant gains in TOPEL Phonological Awareness.

Math skills

The Woodcock-Johnson Applied Problems subscale (Woodcock, McGrew, & Mather, 2001) was used by six validation studies to assess early math skills. Fall and spring standardized scores by state are presented in Figure 16. Massachusetts, Minnesota, and Wisconsin found significant gains, while Rhode Island did not.

Figure 16. Fall and Spring Woodcock-Johnson Applied Problems Scores



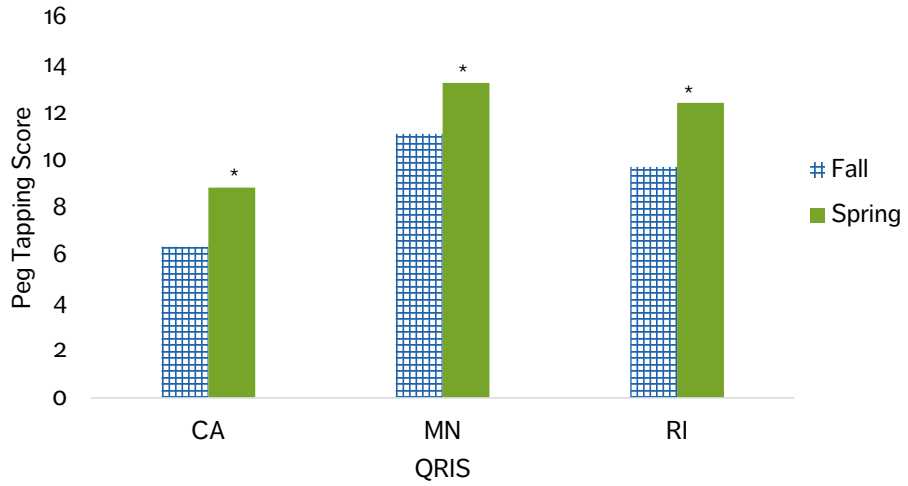
Source: Individual state validation reports

California and Delaware reported significant fall to spring gains with age-equivalent Applied Problem scores.

Executive function measures

Peg (or pencil) tapping and Head, Toes, Knees, and Shoulders (HTKS) are both measures of executive function. Three states used peg tapping and three states used HTKS. Fall and spring raw scores on these two measures are presented in Figures 17 and 18.

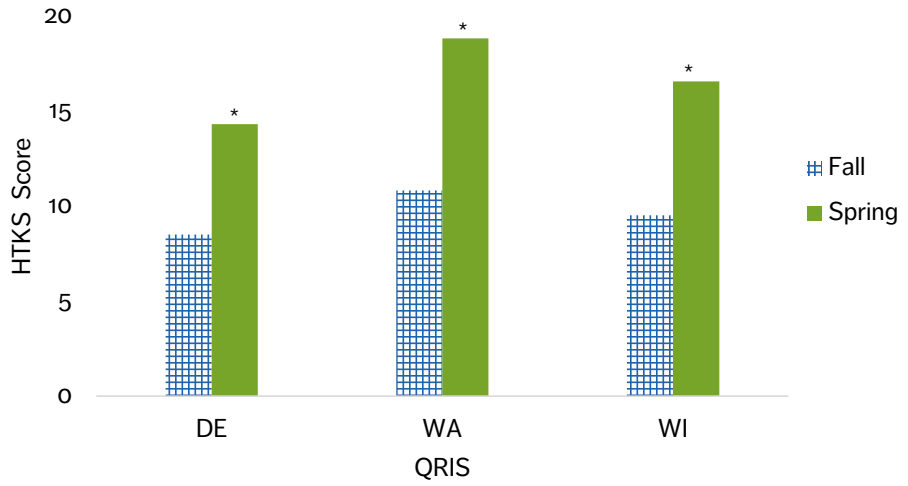
Figure 17. Fall and Spring Peg Tapping Scores



Source: Individual state validation reports

Fall to spring increases were significant in all three states.

Figure 18. Fall and Spring Head, Toes, Knees, and Shoulders Scores



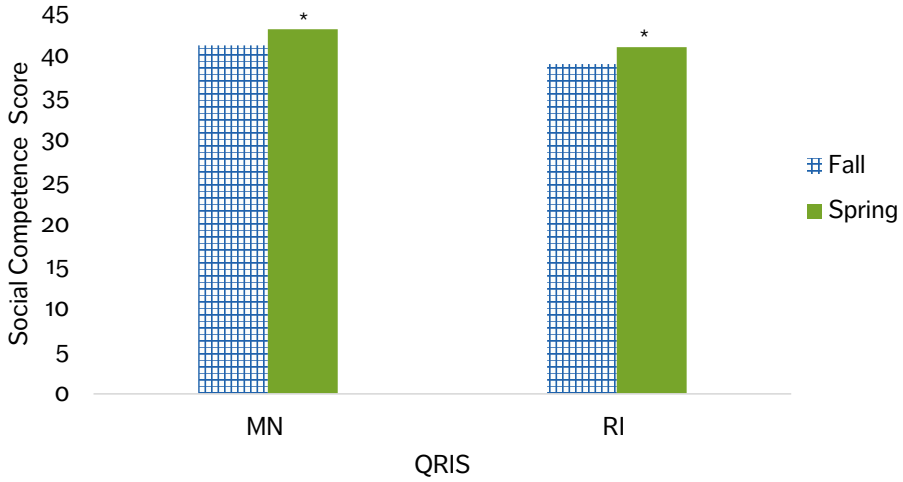
Source: Individual state validation reports

All states that reported significance tests for gains in executive function from fall to spring found significant differences. Children attending QRIS-rated programs improve in executive function over the school year. The executive function measures are raw scores, meaning they do not account for age. Therefore, significant gains indicate that scores increased, but not necessarily more than what would be expected with the passage of time alone.

Social and emotional development measures

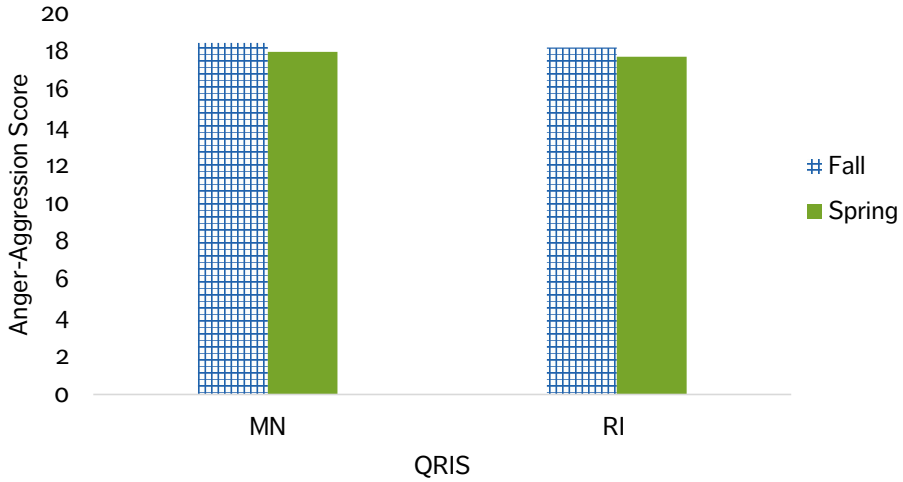
Minnesota, Rhode Island, and Wisconsin used the SCBE-30 as a measure of social competence, anger-aggression, and anxiety-withdrawal. Fall and spring scores for Minnesota and Rhode Island are presented in figures 19 through 21. Wisconsin used raw scores that showed the same pattern of results.

Figure 19. Fall and Spring SCBE-30 Scores: Social Competence



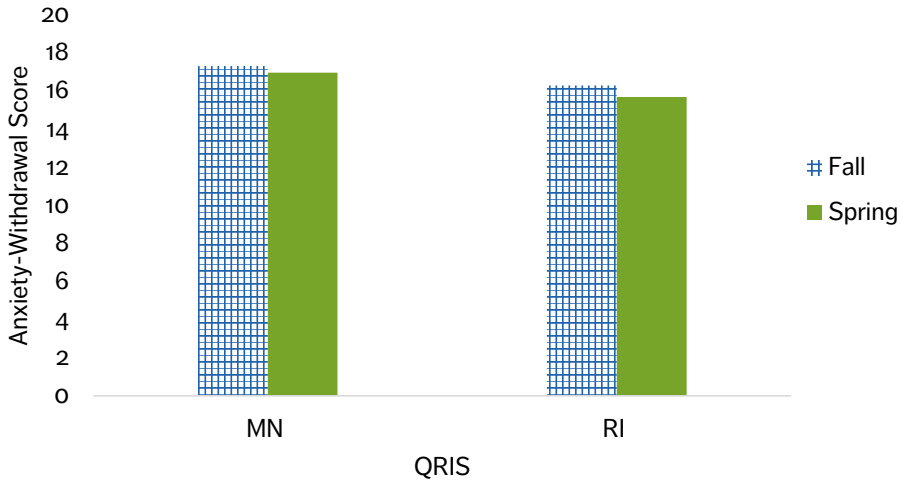
Source: Individual state validation reports

Figure 20. Fall and Spring SCBE-30 Scores: Anger-Aggression



Source: Individual state validation reports

Figure 21. Fall and Spring SCBE-30 Scores: Anxiety – Withdrawal



Source: Individual state validation reports

Social competence scores were significantly higher in the spring than in the fall in all three states (MN, RI, and WI). Children's scores on the anger-aggression and anxiety-withdrawal scales did not change over the year in any of the three states.

Gains in the Preschool Learning Behavior Scale (PLBS) persistence subscale were significant in both Minnesota and Wisconsin. Gains on PLBS Learning Behaviors were significant in Rhode Island. Massachusetts also showed significant gains on the Devereux Early Childhood Assessment (DECA; LeBuffe & Naglieri, 2012).

In summary, with some exceptions, children participating in ECE programs included in the state validation studies made significant gains in the year before kindergarten on developmental skills, including executive function, language and literacy, early math skills, social competence, and persistence. Children did not reduce their anger-aggression or anxiety withdrawal over the year, but starting levels were already quite low.

QRIS Ratings and Child Development

The analyses presented in the previous section described the average fall to spring changes in developmental skills across children in all programs, regardless of their quality level. A next step is to examine whether developmental gains differ by QRIS rating. Analyses to examine this question require considerable complexity to account for a variety of factors that could bias results.

As noted in the section on methodology, many other factors besides experiences in the program affect how children perform on standardized assessments. Children may be influenced by the circumstances of assessment administration (including the tasks included in the assessments, characteristics of the person administering the assessment, and the features of the environment). Research teams plan for and train the assessment team to diminish the influence of these contextual factors related to the assessments. In addition, several family characteristics such as income and parent education, as well as child characteristics such as race/ethnicity and home language, have the potential to affect a child's performance on assessments. Family characteristics are also likely to affect parent choice of ECE setting, with more affluent children generally experiencing higher quality than more disadvantaged children. It is important for validation researchers to consider these potential confounds in their studies. Researchers can statistically control for many family and child characteristics in their analyses, thereby reducing the chances that any associations are due to confounding factors.

One statistical method to address confounding variables is propensity score matching. With this method, children in differently rated programs are "matched" on variables such as gender, ethnicity, family income, etc. The idea behind propensity score matching is that it creates groups of children who are similar on family and child characteristics so that their differences in participation in various types (or levels of quality) of programs can be isolated. That is, the children are theoretically similar except for their participation in different programs (Tout et al., 2016).

Validation researchers typically use Hierarchical Linear Modeling (HLM, also known as multilevel regression models) to examine the association between QRIS ratings and child development. HLM is necessary because multiple children provide scores for each classroom or program (children are nested in programs).

The validation studies examined the associations between QRIS ratings and each developmental outcome. Two states (MA and MN) used gain scores as the outcome (spring score minus fall score) and other states used spring score as the outcome and include the fall score as a covariate.

Multiple models were run in each study. Results are summarized first for the full sample of children in each study and then by income level where possible.

Associations between Ratings and Child Development for All Children

Overall, inconsistent associations between ratings and children's development were noted across states and developmental domains. This section presents the findings by state. Table 17 provides a summary by state and domain.

The validation of the California QRIS assessed children in four QRIS levels, using Level 3 as the comparison category. The results indicated one statistically significant finding: Children in Level 5 programs scored higher on peg tapping (a measure of executive function) than children in Level 3 programs. There were no other statistically significant differences.

Two other validation studies found some evidence for differences in executive function by QRIS level, but with caveats. In Delaware, children in Level 5 programs had significantly higher scores on Head, Toes, Knees, and Shoulders (HTKS) than children in Level 2 programs. The sum of points on the six essential standards in Delaware was also significantly associated with executive function. In Wisconsin, QRIS level did not predict HTKS, but total rating points did significantly predict HTKS scores.[22]

Additionally, in Wisconsin, children in Level 5 scored significantly higher on the Preschool Learning Behavior Scale (PLBS; approaches to learning, persistence) than children in Level 2. The Minnesota study also found a significant difference in PLBS (only the persistence subscale was used) with children in higher-rated programs scoring higher than children in lower-rated programs. Minnesota also found significant gains in social competence for children in higher-rated programs than in lower-rated programs.

The Massachusetts study found that children in programs rated Level 3 showed significantly greater gains in their PPVT (receptive language) scores over the course of the school year than those in Level 2 programs. In addition, children in Level 3 programs showed significantly greater gains on the attachment subscale of the DECA than children in Level 1 programs.

Two significant language outcomes were found in Washington. Infants and toddlers in Level 4 scored significantly higher than infants and toddlers in Level 3 on expressive language, and preschoolers scored significantly higher on receptive language in Level 3 than in Level 2. Also in Washington, infants and toddlers scored significantly higher on fine motor skills in Level 3 than in Level 2.

In the Rhode Island study, the authors designated findings by their statistical significance (at $p<.10$) and substantive significance (with an effect size of at least .07, per criteria set by the What Works Clearinghouse). A significant but not substantive negative association was found with overall rating and expressive vocabulary. No statistically significant and substantive findings were noted with overall rating; however, significant and substantive associations were found between multiple components of the rating scale for math and social competence.

Associations between Ratings and Child Development by Income Level

The Race to the Top – Early Learning Challenge encouraged ECE systems to prioritize support for vulnerable children and emphasized the importance of serving children with high needs (including children in families with low incomes) in high-quality programs (US Department of Education, 2016). As a result, validation studies also emphasized low-income children during study recruitment and in analysis. Studies from Delaware, Minnesota, Rhode Island, and Wisconsin conducted child development analyses with subgroups of low-income children.

Two studies, Minnesota and Rhode Island, found select (albeit inconsistent) significant associations between ratings and children's development by income. In the Minnesota validation study, there was a significant interaction between income group and rating such that low-income children attending higher-

[22] Because there is more variability in points across programs than in ratings levels, using points to predict child outcomes can reveal associations between program quality and child outcomes when overall rating level did not.

rated programs made greater gains in an early literacy skill (print knowledge) and social competence than higher-income children in higher-rated programs. In Rhode Island, there was no association between star rating and spring math scores for children from lower-income families, but there was a significant positive association for children from higher-income families. For social competence and learning behaviors, the pattern was reversed. Among children from lower-income families, a higher star rating was significantly linked to stronger social competence and learning behaviors; among children from higher-income families, there was no association between star rating and spring social competence or learning behaviors.

Two studies, Delaware and Wisconsin, reported no differences in associations with ratings and child development for low-income children relative to higher-income children.

Summary of QRIS Ratings and Child Development across States

Findings on the statistically significant associations between QRIS ratings and child development are summarized in Table 17. Note that in some states, significant associations were noted with rating components but not overall rating. These are designated in the table.

Table 17. Summary of Associations between QRIS Ratings and Child Development, by State and Developmental Domain

	Executive Function	Language/ Literacy	General Cognition	Physical Development	Social/ Emotional	Math
State	Peg tapping/ HTKS	PPVT/ TOPEL/ IDGI/ WJ/ Story and Print/ Mullen	Bracken/ Mullen	BMI/ Mullen fine/ gross motor	SCBE-30/ PLBS/ DECA/ CBCL	TEAM/ WJ
California	✓	*ns*				*ns*
Delaware	✓[1]	*ns*			*ns*	*ns*
Massachusetts		✓			✓	
Minnesota	*ns*	*ns*	*ns*	*ns*	✓	*ns*
Rhode Island	*ns*	—[3]			✓[4]	✓[5]
Washington	*ns*	✓	*ns*	✓	*ns*	*ns*
Wisconsin	✓[2]	*ns*	*ns*		✓	*ns*

Source: Individual state validation reports
Note: A check mark indicates a statistically significant positive association was found between rating level and children's development. A negative sign indicates a statistically significant negative association was found between rating level and children's development. "Ns" indicates that no statistically significant associations were found. A gray, blank cell indicates that the measure was not collected.

[1] The analysis in Delaware found a significant difference between level 5 and level 2 only; in addition, a significant association was noted with executive function and the sum of points on the six essential standards. [2] The analysis in Wisconsin found a significant association with total rating points, not rating level. [3] The analysis in Rhode Island found a significant negative association between rating and expressive vocabulary. [4] The analysis in Rhode Island found significant associations between social competence and multiple rating components (but not overall rating). [5] The analysis in Rhode Island found significant associations between math and multiple rating components (but not overall rating).

Overall, three of six states (California, Delaware, and Wisconsin) found evidence for a significant association between QRIS rating (or overall points obtained) and executive function. Significant associations between QRIS rating (or rating components) and social-emotional development were found in four of six states: Massachusetts, Minnesota, Rhode Island, and Wisconsin.[23]

[23] The overall pattern was not consistent in Wisconsin, which raises the question of whether it was a spurious association.

Putting the Validation Findings in Context

The availability of findings from 10 recent validation studies of ratings used in Quality Rating and Improvement Systems offers a unique opportunity to examine the consistency of findings that emerge across state studies with different characteristics and unique system features. Two strategies are used in this section to reflect upon the findings and their implications for early care and education system initiatives. First, we summarize common themes from a review of the recommendations included in the 10 individual state reports. This summary highlights the options that individual research teams put forth for consideration by their state partners based on the validation study findings. Second, the key conclusions of the cross-state synthesis are shared with implications for next steps in QRIS implementation and evaluation. This discussion is framed by noting the limitations of the research strategies used in the state studies and current challenges in research on quality of early care and education. Together, the discussions of within- and cross-state findings put validation study results in context and show the different ways they can inform ongoing decision making about ECE quality improvement efforts.

A Summary of Recommendations Emerging from Individual Validation Studies

As noted, the validation studies were conducted during a period of rapid change in state QRIS, in part due to new funding from the Race to the Top – Early Learning Challenge grants and the system changes that accompanied them. For example, states were navigating changes from voluntary to mandatory participation for some program types, expansion from a pilot to statewide implementation, and newly revised quality indicators. Most of the research teams reported that they developed partnerships with states to design studies that could meet the requirements of the Early Learning Challenge grants while also providing information to inform ongoing QRIS implementation. The state validation teams developed a range of recommendations to accompany the reports for state partners. Although each report was tailored to the needs of the particular state, it is informative to look across recommendations to identify common themes. Four themes emerged.

First, the most common recommendations addressed some aspect of the ratings, including the number of levels, the number of criteria, and the scoring process. For example, four state reports (Delaware, Massachusetts, Oregon, and Wisconsin) included recommendations to reduce or streamline the number of quality indicators that comprise the rating. The recommendations typically emerged from findings indicating that certain quality indicators were not differentiating quality across rating levels or associated with children's development outcomes. In contrast, three state reports (Maryland, Washington, and Wisconsin) recommended adding indicators to the rating scale (e.g., an observation tool such as the CLASS to focus on instructional quality) and/or reducing the number of levels of the QRIS. The California and Washington reports included recommendations about changing the scoring method or criteria by, for example, modifying cut points that differentiate levels.

A second set of recommendations focused on developing stronger quality criteria and verification processes. The Delaware, Massachusetts, Oregon, and Wisconsin reports recommended strengthening quality standards and/or including a verification process at higher levels of the QRIS. Two state reports, California and Massachusetts, suggested enhancements to the verification process by incorporating better verification requirements (e.g., outline what is needed for documentation requirements).

Third, findings on the overall quality levels observed in ECE programs prompted recommendations related to quality improvement. Four state reports (California, Delaware, Massachusetts, and Minnesota) included recommendations about strengthening or revising the quality improvement supports, focusing in particular on coaching. Recommendations addressed the need for more sustained, ongoing coaching throughout the year for programs (California), allowing for more flexibility in technical assistance supports that differentiate by program type and size (Massachusetts), or for programs serving more low-income children (Minnesota).

A final common recommendation that emerged was to continue developing high-quality data that can be used for progress monitoring. Six state reports (California, Delaware, Maryland, Oregon, Rhode Island, and Washington) included recommendations about better use of data. Both the Delaware and Rhode Island reports recommended engaging in better or additional data collection efforts that can be used for quality improvement. For example, the Rhode Island report recommended collecting data on all quality indicators at all rating levels to understand whether programs meet indicators at levels above the one for which they are applying (and revising indicators based on what is learned). Furthermore, the Delaware, Oregon, and Rhode Island reports encourage the use of administrative data collected on programs to support QRIS refinement and quality improvement. The Delaware report, for example, recommended using administrative data for monitoring purposes, as well as linking data within the QRIS to better answer questions about areas of need in supporting quality improvement practices. The California, Maryland, and Washington reports emphasized the importance of engaging in ongoing validation and evaluation studies as the QRIS systems continue to mature and change.

Across the states, recommendations conveyed both general and specific options for improvements to the QRIS. Tracking whether and how the QRIS change in response to validation studies and report recommendations will be an important activity for future research. For example, states may want to investigate the implications of any changes made for the validity of their ratings.

Conclusions and Implications of Findings in the Cross-State Synthesis

With background on how the validation studies were used to generate state-specific recommendations, this section turns to conclusions and implications that can be developed by looking across the 10 studies. As noted, the studies included in this synthesis asked two questions (required by the Race to the Top – Early Learning Challenge grant) to understand the extent to which ratings are distinguishing quality in a meaningful way.

1. To what extent are QRIS ratings associated with measures of observed quality in ECE programs?
2. To what extent are QRIS ratings associated with children's developmental gains in skills important for school readiness?

Each question is addressed in sequence.

Summary of Findings: Observed Quality

Each of the nine state studies (all but WA) addressing the question about ratings and observed quality in center-based programs found evidence of a significant association. Overall, QRIS ratings reflect differences in environments, interactions, and activities in ECE programs at different rating levels. Although statistically significant, the differences in observed quality scores between QRIS rating levels were generally small.

The findings for family child care program QRIS ratings and observed quality were more mixed. Seven (Arizona, California, Delaware, Maryland, Minnesota, Oregon, and Wisconsin) of the ten studies included family child care programs in the analysis of ratings and quality. Three studies (Arizona, Maryland, and Oregon) found significant positive associations with observed quality and ratings. Three studies (California, Delaware, and Wisconsin) did not have an adequate sample of family child care programs to draw conclusions, and one (Minnesota) found no differences in observed quality among family child care programs at different QRIS rating levels.

Because of limitations in the number of programs at each level of the QRIS and in the study samples, the findings about observed quality indicate that ratings generally distinguish lower and higher quality but do not support the conclusion that each level of a QRIS reflects a meaningful difference in quality from other levels.

Findings with observed quality were noted for different types of measures, including those with a focus on global quality (the Environment Rating Scales and the Preschool Quality Assessment) and those with a focus on teacher/caregiver-child interactions (the Classroom Assessment Scoring System and the Caregiver Interaction Scale).

Overall levels of quality in the state studies were in the medium-quality range rather than the-high quality range.

Summary of Findings: Child Developmental Gains

Seven state studies examined QRIS validity by looking at children's developmental gains. These studies yielded inconsistent evidence of small positive associations. Some selective positive associations were found in some states, but not across all developmental domains examined, nor across all measures within a domain.

Three of six studies found some evidence that QRIS ratings were associated with measures of executive function, and four found some selective associations between ratings and measures of social-emotional development. Two studies found evidence for an association between ratings and receptive language, and one found evidence for a link between ratings and math.

Within the studies, the patterns of associations were not consistent for all comparisons across higher and lower star levels or for differing measures within a developmental domain. All significant associations were small. For measures of language and math skills, the studies found more evidence of a lack of associations than evidence of positive associations. Overall, QRIS ratings were not strongly associated with patterns of children's growth across the range of developmental skills assessed in the validation studies.

Limitations of Validation Studies

When reviewing findings, it is important to keep in mind limitations of the validation studies. In most states, very few programs had been rated at the highest level of the QRIS. Overall, the programs included in the "high-quality" group (which sometimes were collapsed across three levels) for analysis had scores on measures of observed classroom quality that were in the medium range of quality (according to measure developers) and not above the thresholds identified by some research (e.g., Burchinal, Zaslow, & Tarullo, 2016) as those that have associations with children's learning. This limited range of quality scores may make it difficult to detect associations with children's development.

Additionally, important subgroups of programs and children were not examined in many state studies. Family child care programs were not included in all studies; in some studies, conclusions could not be drawn about family child care quality and ratings because of small samples. Also, few studies included measures for infants and toddlers, children with special needs, or children speaking languages other than English and Spanish. The studies reported here focused primarily on preschoolers in the year before kindergarten, which provides an important yet narrow understanding of children's development in QRIS-rated programs.

The studies were also limited by the contextual and methodological issues discussed in earlier sections of the paper. Program recruitment was challenging due to small sample sizes, and study recruitment rates were low among some types of programs. Recruitment rates for children were difficult to calculate because the strategy used to recruit families and children in some studies does not allow simple calculation of the denominator (the total number of families reached by recruitment materials). Studies had missing data for several key data points for programs and children.

Limitations of Quality Measures and Their Associations with Children's Development

The findings highlight limitations of current measures and approaches and raise general issues about quality measurement in ECE programs. For example, recent studies that include measures of observed quality in preschool classrooms demonstrate a relatively narrow range of scores on tools such as the CLASS and ERS—similar to the range documented in the validation studies (Aikens et al, 2016). Additional research may be needed to clarify the practices that comprise quality, and to develop or refine observational tools so that they capture the variation in practices that better differentiate lower- and higher-quality programs. Programs typically score in the mid-range of the tools, which indicates that programs have room to improve. The scoring patterns could also indicate, however, that further calibration of the tools is needed to identify practices between the medium and high ranges that are missed with current scoring guidelines, and could help differentiate quality in programs.

In addition, observed quality in center-based settings varies by classroom, and use of center averages on observational measures may mask significant variation in quality that exists across classrooms (Karoly, Zellman, & Perlman, 2013). This variation in quality across classrooms may affect both the QRIS ratings and the outcomes in the validation studies. QRIS measurement protocols typically use averages across one-third or one-half of classrooms serving each age group, and often include requirements that scores are not below certain thresholds. In programs with small numbers of same-age classrooms, QRIS classroom selection protocols may result in most classrooms being rated. Karoly and colleagues (2013) found that observing quality in more classrooms will promote better accuracy in classifying programs at different rating levels. They also note that QRIS face trade-offs in spending more resources on measurement of more classrooms in addition to investments in quality improvement efforts.

Potential Challenges of Current QRIS Structures for Validation Studies

The structure of current QRIS provides additional challenges for validation research. First, QRIS incorporate many quality components into one overall rating. Some components are directly aligned with children's learning and development; others are associated with activities to support teachers' and caregivers' professional development and families' needs, but not children's pre-academic skills directly (Burchinal, Tarullo, & Zalsow, 2016; Tout et al., 2010; Zaslow & Tout, 2014). Decades of research show that even the components most likely to reflect children's interactions and experiences in their early learning settings (e.g., ERS, CLASS, use of assessment tools and curriculum, ratio, teacher/director education) tend to have small associations with gains in children's outcomes (Burchinal et al., 2016). Thus, it is not surprising that the associations between overall QRIS ratings (that include many components) and children's gains are not consistent in the studies included in the synthesis.

Similarly, QRIS typically do not include indicators that document specific practices related to supporting children's language and math skills (Smith et al., 2012). Recent research demonstrates that linkages between certain academic skills such as language and math are stronger with quality measures aimed at documenting developmentally appropriate interactions and practices that support the acquisition of these skills (Burchinal et al., 2016).

Alternatively, some aspects of quality more strongly associated with children's development are harder to measure and are not currently captured in QRIS ratings or in the measures of observed quality that QRIS use. For example, a tool was recently developed to assess the extent to which teachers use the results of assessments to individualize their practices to support children's needs (Monahan et al., 2016). Understanding more about children's individual experiences in ECE settings, and whether they support children's unique needs, are important enhancements for quality measures.

In addition, the distribution of ratings in QRIS and pathways for programs to advance in ratings are also important to consider in the review of validation findings. For example, in some QRIS, a Level 1 or 2 may reflect the programs at the lowest level of quality. In other states, such as Oregon and Wisconsin, lower

levels may include programs with automatic ratings that choose not to apply for a higher rating level. In states like Minnesota, this rating may include programs at higher quality levels that first apply for a level 1 to access resources before applying for a higher rating level. The Oregon validation study used methods to select a sample of Level 1 programs that would not have achieved a higher rating level if they had applied for one. In Wisconsin, analyses revealed that rated Level 2 programs looked similar to unrated Level 2 programs (i.e., those that had not tried to attain a higher level). In Minnesota, family child care providers were encouraged to start the rating process at lower levels so they could access supports and engage in a process of improvement. Scores on the FCCERS-R in Minnesota did not differ by quality level. These examples demonstrate that variation at lower levels may or may not be an issue of consequence for the QRIS. It is worth investing in research to understand how quality is distributed at the lower levels, and enacting methods (such as those used in Oregon) to understand the sample being selected at the lower quality levels in states where variation may pose challenges to analyses.

Implications of Validation Findings

Acknowledging the limitations of validation studies, key implications can be drawn from study results.

QRIS ratings appear to be a useful tool for state early childhood systems to differentiate programs at lower and higher levels of quality. A review of individual state studies indicates that it is important to consider the integrity of the rating by ensuring an appropriate number of quality indicators, as well as indicators that can ensure rigor of differentiation (particularly at the highest QRIS levels).

The results documenting observed quality at medium and low levels across many QRIS programs highlight the need for continued investment and innovation in quality improvement supports for ECE programs. Research indicates that programs improve over time in QRIS, but few studies have documented the most effective ways to promote meaningful improvement that can be sustained and can support children's positive development (Karoly, 2014). Longitudinal studies to understand how programs improve and how teachers and caregivers perceive the quality improvement process will add value to the existing literature (see, for example, Elicker et al., 2017).

Next Steps

Several activities could build upon the studies and results described in this report to enhance quality measurement and QRIS:

- As noted, it will be important to build the literature on family child care programs in QRIS and understand how current quality measures are working in these settings. As enrollment of family child care providers in QRIS increases, efforts to document their quality will inform the field.
- Validation studies (specifically) and quality improvement studies (generally) must include children with special needs, infants and toddlers, and children speaking languages other than English and Spanish. Understanding how program quality is associated with outcomes is limited by the exclusion of these important populations of children.
- Measures of children's experiences in early care and education beyond traditional school-readiness skills should be included. The forthcoming study in Oregon will include a measure of children's engagement in their early learning settings and a measure of family-teacher relationship quality.[24] Other important measures of children's experiences in early care and education could include the quality of relationships with staff and peers and children's continuity in high-quality settings.
- As described in this report, validation studies examine how ratings are associated with measures of observed quality. Yet quality ratings incorporate different domains of quality that may not be assessed

[24] Though not described in the synthesis, the Maryland validation study included a measure of child engagement but did not find significant differences by rating level (Swanson et al., 2017).

by current observational measures. Even when small differences are noted between levels on the observational measures, the QRIS indicators may be capturing other aspects of quality that contribute to the experiences children and families have in ECE programs. For example, programs that meet indicators related to the work environment may have more stable staff than other programs. A next step for validation studies and other studies of quality is to examine associations between ratings and indicators such as turnover, compensation, and other workforce supports that may not be directly associated with observational measures or children's development but may reflect important infrastructure elements for building and improving quality. Oregon, for example, included some of these outcomes in its validation study (Lipscomb et al., 2016).

- Finally, QRIS ratings are used for a variety of purposes. For example, ratings are used to target quality improvement supports, target scholarships for vulnerable children to access higher-quality programs, and provide information to parents making ECE choices. In a block QRIS design, some QRIS levels may be set to encourage quality improvement, but not to discern meaningful differences in children's development at each level. Some quality levels may be set to engage programs in the system. And some quality levels and indicators may have clear connections to higher-quality practices that can support children's positive development. Clarifying the theory of change for each QRIS can help identify more accurate hypotheses about which quality levels and quality indicators should be differentiating observed quality and children's development. This may not, in fact, be an explicit goal at every level of the system (Schilder et al., 2015).

Overall, the validation studies described in this report provide helpful guidance to inform the next round of improvements to QRIS ratings across the country. Indeed, each state approached validation as a strategy to improve its rating process and tools. The studies indicated that the ratings are generally working to distinguish lower and higher quality, but that further work is needed to strengthen quality measurement. Limited positive associations were found between ratings and children's development. These findings can prompt discussions about how to improve quality measurement and support quality improvements that promote the development of young children in ECE programs.

Commentary 1. A National Perspective

Author: Elizabeth Shuey, Society for Research in Child Development Fellow

The validation studies included in this synthesis suggest good news for the investments made in Quality Rating and Improvement Systems (QRIS) over the past two decades: QRIS ratings are associated with measures of observed quality. In other words, states are successfully differentiating quality of early care and education (ECE) programs using the indicators included in their QRIS. The realization of this goal is significant both from a measurement perspective—developing a valid measure of something as broad and difficult to capture as ECE quality is no small feat—and also from a policy perspective. Through QRIS, most states now have infrastructure in place that can enable engagement with ECE programs, tracking of program quality, and targeting of ongoing investments in quality improvement.

Nonetheless, as this synthesis shows, the association between QRIS ratings and gains in child development is less clear. The findings related to child development raise important questions about the meaning and measurement of ECE program quality, as well as the specific goals of and expectations for QRIS. In this commentary, I highlight a few key considerations around measurement of program quality and child development before providing suggestions for ongoing improvement of QRIS themselves and recommendations for ongoing research.

Measuring ECE Program Quality and Child Development

First, measuring ECE program quality is a continual challenge for research and practice alike. A few measures of observed quality, notably the CLASS and Environment Rating Scales (ERS), dominate the research literature on ECE program quality, as well as the protocols used for monitoring ECE programs in QRIS, state pre-Kindergarten programs, and Head Start programs. Although they are widely used, these measures tend to show limited variability across programs, meaning that the range of scores is relatively small; therefore, very small differences in scores can be important. With this in mind, it is not surprising that differences in observed quality scores across QRIS levels in this synthesis were generally small.

Second, measures of observed quality are collected at the classroom level rather than at the program level. For most policy purposes, program-level quality, which is the focus of QRIS, is of greater interest than classroom-level quality; however, measures of observed quality are designed for classroom-level use. Findings from individual studies included in this synthesis (e.g., Delaware, Wisconsin; see Karoly et al., 2016 and Magnuson & Lin, 2015) confirm that there is a wide range of classroom-level quality within programs at different levels of QRIS. Once again, small differences in average observed quality scores at various QRIS levels in the validation studies reviewed is to be expected given the wide range of classroom quality present within ECE programs.

Perhaps more importantly, the variability in classroom quality within programs has implications for children's development. Although on average, children are in classrooms observed to be of higher quality when they attend programs at higher levels of a QRIS, some children in programs at lower levels of a QRIS also experience higher-quality classrooms. This situation makes linking differences in child development to QRIS ratings extraordinarily challenging.

Measuring young children's development in meaningful ways also can be challenging, given the rapid pace of growth during these early years, as well as expected day-to-day variability in children's behaviors, interests, and reactions to assessors. Moreover, young children naturally make impressive gains from fall to spring of an academic year and, although we know quality ECE programs can support these gains (see Phillips et al., 2017), growth is to be expected for all children. Thus, we set a high bar when we expect experiences with programs at different levels of a QRIS to be associated with different levels of growth for children, particularly if we are not addressing the classroom-level quality children experience within their programs or measuring children's development with great precision.

Nonetheless, as described in this synthesis, children attending programs that participate in QRIS made meaningful gains, on average, across the academic year regardless of their programs' ratings. Although the validation studies do not include comparison groups of children who were *not* attending programs that participate in QRIS, the gains in standardized test scores (e.g., Woodcock-Johnson) suggest that simply participating in a QRIS program may be beneficial for children, in ways that are consistent with past literature on the importance of ECE programs for supporting children's development. Regardless, given the substantial measurement challenges involved in making links between QRIS levels and children's development, these associations should not be a primary focus of assessing the meaning or value of QRIS. Instead, the findings described in the synthesis serve to demonstrate the validity of these QRIS as measures of ECE program quality and are an excellent starting point for considering next steps in QRIS development and research.

Improving QRIS and Ongoing Research

Although the QRIS included in this validation synthesis successfully differentiated lower and higher levels of observed quality in ECE programs, the levels of quality at the highest QRIS ratings were not necessarily characteristic of optimal quality settings. Research findings suggest that ECE program quality is linked to children's outcomes only once a threshold of quality is met (Burchinal, Zaslow, & Tarullo, 2016). In other words, ECE programs must provide a relatively high level of quality before we can expect them to be associated with child development. This makes sense, given the natural growth young children experience in safe settings.

Thus, states should consider strategies to promote ongoing quality improvement, even among programs at the highest QRIS levels. Achieving a five-star rating or a national accreditation should not be seen as a final step in quality improvement, but rather as a marker of successful progress toward and a commitment to providing high-quality ECE. Although resources to bolster quality at the lower end of QRIS are essential, understanding the needs of higher-quality programs can help states identify efficient strategies to continue quality improvement across the QRIS spectrum. Targeted and tailored resources can allow programs to receive training and technical assistance for which they are ready and that will be meaningful for ongoing quality improvement regardless of the quality starting point.

A next step for research is to build upon the work identifying thresholds in quality, to help states understand the types of resources that may be most appropriate for programs at different levels of QRIS. Similarly, research can help identify when it may be appropriate to use child outcome data as a validation tool for QRIS, and when other metrics will be most relevant. The recently launched research project on Variation in Implementation of Quality Interventions (VIQI), funded by the Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families (ACF), promises to offer important insights to the quality improvement strategies that are most successful for ECE programs at different levels of quality (see https://www.acf.hhs.gov/opre/research/project/variations-in-implementation-of-quality-interventions-examining-the-quality-child-outcomes-relationship).

In addition to highlighting the value of ongoing quality improvement, the synthesis also notes the large number of standards that programs must address to participate in QRIS. Many of the individual validation reports suggest streamlining these standards by identifying a smaller set of indicators that are most meaningful to distinguish programs across levels of a QRIS. In considering this recommendation, states should identify the ways in which QRIS indicators translate into practice at the program level. For example, does requiring programs to complete an annual self-assessment contribute to quality improvement planning, or is the self-assessment simply an administrative burden to the program? QRIS indicators should be aspirational—promoting structures and practices that are most likely to support high-quality ECE—while striving to keep the effort programs must expend to demonstrate they meet each indicator to a minimum.

Further, more attention is needed to understand the ways in which parents use and understand quality ratings. The multi-dimensional nature of QRIS means that resulting ratings are complex and very different programs—such as a Head Start, a community-based child care center, and a family child care program—

could have the same ratings while providing very different services and experiences for a family. Consumer education is a major focus of the 2014 Child Care Development Block Grant (CCDBG) Reauthorization, meaning that states have an obligation to provide information to parents as part of receiving federal funds for child care subsidies. Yet we know from the 2012 National Survey of Early Care and Education (NSECE) that many families (37%) consider only one provider when searching for ECE (NSECE Project Team, 2014). As QRIS continue to expand, ongoing research is needed to identify the ways in which quality ratings can be most meaningful to families searching for care. Such research can also inform discussions of school choice that are pivotal in current discussions of K–12 schooling; we must better understand how parents make choices for their children in the context of diverse family needs and priorities, constrained sets of options, and limited information about the quality of these options.

This synthesis also suggests the importance of and continuing need for researchers, policymakers, and program administrators to work together toward ongoing quality improvement at a systems level. As a starting point, state teams should carefully consider goals and expectations for QRIS. By creating a logic model that identifies the investments in QRIS and the specific, anticipated results of these investments, states can determine key opportunities for measurement and evaluation of their systems. In considering the anticipated outcomes of investing in QRIS and developing logic models, states will need to attend to all of the measurement challenges noted in this commentary, as well as the many other challenges identified throughout the synthesis. Notably, helping children become "kindergarten-ready" may be a worthy goal, but much greater specificity is needed to identify developmentally appropriate indicators for children's outcomes that can reasonably be expected to change as a result of the quality dimensions assessed in QRIS.

In addition, logic models can help frame messaging around the importance and meaning of QRIS, facilitating communication with diverse audiences from policymakers to practitioners and parents. The synthesis highlights some of the difficulties in interpreting and using research findings when study questions are narrowly framed: validating QRIS is a worthy goal, but validation results are most meaningful in combination with process or implementation studies that can speak to the question, "what do we do now?" By rooting QRIS research and evaluation efforts in a logic model co-constructed by researchers and system stakeholders, future studies can play a role in continuous quality improvement of QRIS.

Researchers need to take a prominent role in state teams addressing ongoing quality improvement of QRIS. Although traditional approaches to evaluation assume a static, "final" intervention as the focus for research, QRIS are not discrete interventions. We should not assume that these complex systems will achieve a finalized, fixed form, and rather we should hope that QRIS are flexible to address shifting policy contexts and to meet emerging demands in the field of ECE. The need to have adaptive QRIS creates a tension with evaluation efforts, and researchers must take up this challenge. By engaging with policymakers and QRIS stakeholders in ongoing ways, researchers have the potential to develop innovative strategies that can provide data to inform continuous system improvement.

Finally, the strengths of QRIS identified through this synthesis provide a solid foundation for states to continue efforts to define quality and to engage ECE providers and families in conversations about quality. QRIS offer the structure to successfully move forward with these goals. This synthesis of findings is a valuable tool—not only for the 10 states that have recently released QRIS validation reports, but also for other states considering best practices in QRIS development and research.

References

Burchinal, M., Zaslow, M., & Tarullo, L. (2016). Quality thresholds, features, and doses in early care and education: Secondary data analyses of child outcomes. *Monographs of the Society for Research in Child Development, 81*, 1–128.

Karoly, L. A., Schwartz, H. L., Setodji, C. M., & Haas, A. C. (2016). Evaluation of Delaware Stars for Early Success: Final report. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR1426.html

Magnuson, K., & Lin, Y. (2015). Validation of the QRIS YoungStar rating scale, report 1. Madison, WI: School of Social Work and Institute for Research on Poverty. Retrieved from https://dcf.wisconsin.gov/files/youngstar/pdf/validationreport1.pdf

National Survey of Early Care and Education Project Team. (2014). Household search for and perceptions of early care and education: Initial findings from the National Survey of Early Care and Education (NSECE) (OPRE Report No. 2014-55a). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Phillips, D.A., Lipsey, M.W., Dodge, K.A., Haskins, R., Bassok, D., Burchinal, M., et al. (2017). Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects. Washington, DC: Brookings.

Commentary 2. A State Perspective

Authors: Katherine McGurk and Erin Gernetzke, Wisconsin Department of Children and Families, Division of Early Care and Education

Wisconsin's Department of Children and Families (DCF) created our Quality Rating and Improvement System, YoungStar, with rating criteria that were designed to evaluate overall program quality, and that reflected input from key stakeholders and experts in the field of early education. We were eager to work with Dr. Katherine Magnuson and the University of Wisconsin – Madison Institute for Research on Poverty (IRP) to determine if YoungStar ratings predict independently observed classroom quality and children's assessed school readiness. Results from the YoungStar validation study (Wisconsin Early Child Care Study [WECCS]), compelled us to determine if/how YoungStar can be improved in future years.

The WECCS concluded with mixed evidence regarding the validity of YoungStar. The first part of the validation study confirmed that as a measure of quality, the YoungStar rating system achieved validity. We were thrilled to see results supporting YoungStar's ability to successfully address a foundational piece of quality care. YoungStar achieved its original intent to improve broad dimensions of the quality of early education environments for the children of Wisconsin and the rating system can distinguish between, and predict, varying levels of overall program quality.

The second part of the validation study considered moving beyond the overall program quality and looked specifically at child outcomes. Results showed that, on average, children in YoungStar programs were meeting developmental expectations and learning an important range of skills during the time period of the study. However, analyses of the data did not support the conclusion that participation in more highly rated YoungStar programs predicted children's school readiness. Although disappointed, Wisconsin leadership was assured that these results are consistent with other states' validation studies and with studies that find that observational measures of general child care quality are not consistent and robust predictors of child outcomes. After many discussions, we agreed that the lack of results were related to two key take-aways:

- YoungStar rating criteria is based on evaluating overall program quality and was not developed as a means to specifically predict child outcomes.
- The broad dimensions of overall program quality assessed by YoungStar are necessary as a foundation of high-quality care. However, these broad measures do not intentionally or specifically address the more academic and cognitive skills that were measured in the study.

Finally, when interpreting the results as a whole, we worked with the University of Wisconsin's Institute for Research on Poverty (IRP) to identify the limitations of the study. Although multiple assessments, highly trained assessors, and a large sample of YoungStar participating programs were included, the study was not designed to test for differences in observed quality or child outcomes between programs at the higher end of the rating scale. Thus, the results should be largely interpreted as differences between 2 and 3 Star programs. Additionally, most 4 and 5 Star programs participating in the study were rated through YoungStar's automated accreditation rating option. Finally, the average ERS scores of 2, 3, 4, and 5 Star programs participating in the validation study were all in the "Moderate Quality" range, and one should not expect to see large differences in quality outcomes between programs that score in the same range.

In light of the findings, the study's limitations, and a desire to continuously improve our QRIS, DCF again partnered with the IRP to further investigate current YoungStar rating data, and uncover information that would help us refine our rating scale into one that better predicts and supports children's outcomes. We wanted to explore the impact that other assessment tools, accreditation, and teacher qualifications have

on improving child outcomes. We also wondered if we should consider focusing our rating criteria around measures that we believe are correlated with improvements in school readiness skills and behaviors, such as:

- Curriculum alignment with Wisconsin Model Early Learning Standards
- Intentional planning in a program's curriculum
- High-quality practices for family involvement
- Practices for enhancing the professional development of key program staff

DCF and the IRP undertook a collaborative data and research sharing process, and merged their policy and research expertise into two broad, concrete recommendations for improving the rating criteria. These recommendations were:

1. Consider introducing new, required Learning Environment and Curriculum indicators that will do a better job at differentiating between programs of varying quality at the lower end of the quality scale. This recommendation is based on the fact that there are not items within the rating scale that are differentiating learning environment quality between 2 and 3 Star programs. Because children's learning and behavior is most likely to be affected by dimensions of their learning environments, it is important to adequately measure the learning environment at each star level.
2. Consider whether including cut points from subscales of the ERS—or using a different observation tool—would better assess the learning environment at the higher end of the quality scale, and further differentiate program quality between star levels. As part of this work, we surveyed what some other states are using as quality indicators and found significant overlap with what Wisconsin is already doing. However, some other states specifically include cut points for the more instructional subscales of the ERS, include higher average ERS scores which must be attained in order to move to the highest rating level, and/or have moved to using ERS-3 tools (ECERS-3, ITERS-3) in lieu of the CLASS or older versions of the ERS.

With these recommendations in mind, DCF convened an advisory group, made up of key early education stakeholders, to solicit feedback and input. We intentionally included a diverse group of stakeholders, including providers, leaders of technical assistance (TA) organizations, researchers, higher education staff, community leaders, and legislators (from both parties).These representatives met together to:

- Determine if YoungStar should adopt a new focus on improving child outcomes.
- Provide feedback around the evidence-based recommendations to enhance YoungStar, keeping in mind the intent of YoungStar, any financial and practical implications, findings from other states' validation studies, and the additional rating criteria analysis completed by IRP.
- Bring forward additional changes that were not included in the analysis, and discuss how these recommendations would make YoungStar ratings a stronger predictor of quality.
- Identify resources and supports that are necessary to achieve and maintain high-quality child care programming.

After several Advisory Group discussions, solicitation of feedback from each stakeholder's membership/constituency, and review of DCF's current resources, we have decided on two changes that will be made to YoungStar rating criteria in 2019.

1. YoungStar will lower its 3 Star educational threshold for Lead Teachers/Group Leaders and family child care providers. This change will reduce cost and other structural barriers to higher education completion

and support family child care programs (as they tend to have lower ratings). Stakeholders found it challenging to balance the benefits of higher education with the benefits of removing barriers that prevent programs from improving practices. However, after this change was coupled with an increased focus on learning environment practices, most stakeholders were supportive.

2. YoungStar will require programs to demonstrate developmentally appropriate materials, interactions, and learning centers to achieve a 3 Star rating. This change was unanimously supported by stakeholders, validates programs that are already demonstrating quality classroom practices, and will likely increase TA requests for targeted classroom support.

The combination of these two changes creates a pathway to a 3 Star rating, allowing programs that have staff with slightly less formal education to instead demonstrate their quality practices within classrooms/programs. This allows for 2 Star programs with high-quality learning environments and stronger application of important early education practices to increase in rating faster. DCF plans to provide targeted supports to programs during the transition to this new QRIS structure, including an increase in access to supports (specific trainings and material kits), clear and timely communication, and specific inclusion of tribal considerations.

DCF will continue to investigate the feasibility and impact of the larger potential changes to rating criteria, including changing or adding observation tools (ERS-3, CLASS), adding staff retention strategies (for example, requiring salary parity with 4K – Wisconsin's Four-Year Old Kindergarten program, paid planning time), and making changes to the overall YoungStar structure (for example, reducing the number of quality levels and reducing optional quality criteria). This investigation would be conducted through a continued partnership with IRP as additional funding is identified. A pilot study, which tests one or more of the above changes, may also be considered as funding is identified.

Commentary 3. A Research Perspective

Author: Martha Zaslow, Society for Research in Child Development

This report reflects major steps forward in terms of research examining the validity of state Quality Rating and Improvement Systems (QRIS). The core of the report is the coordinated examination across ten states of how each state's QRIS quality ratings relate to observed quality and to child outcomes. Because the research teams from ten states have collaborated in producing this report, there is a unique opportunity to look across the state studies for both consistency and variation in the validity findings when QRIS are embedded in differing state contexts and use somewhat different rating approaches.

Important progress reflected in the descriptive data and analytic approaches. Before commenting on the report's contributions in the two core areas (looking at QRIS in relation to observed quality and child outcomes), however, it is important to note the indications of progress reflected in the descriptive data presented at the start of the report regarding the maturity of state QRIS, at least among the ten participating states. These ten states show more advanced stages of implementation of QRIS (beyond piloting or local implementation to statewide implementation for most of the states) and high rates of participation by eligible providers in multiple instances (for example, four of the ten states that collaborated in this report had more than 50 percent "density" or participation by eligible providers). Another key marker of overall progress that is noteworthy is the number of states for which participation in the QRIS is mandatory for programs participating in the child care subsidy system (half of the ten states collaborating in this report had QRIS for which such participation was mandatory). While QRIS are still evolving, the validation results summarized in this report clearly reflect QRIS at a more advanced stage of maturation than summarized earlier by Karoly (2014).

There is also progress in terms of the way in which these states focused their validation studies. It is not so long ago that there was uncertainty about what the focus of QRIS validation studies should be and how they should be conducted. The INQUIRE group has made major contributions to the conceptualization and operationalization of QRIS validation research (see especially Zellman & Fiene, 2012, a paper that grew out of and reflects substantial discussion among INQUIRE participants). In addition, as noted in the report, the Race to the Top – Early Learning Challenge Grant requirements have also fostered greater agreement on validation study approaches. Undoubtedly reflecting both sources of influence, rather than a scattering of differing approaches, the ten state research teams that have collaborated in this report have converged on specific research questions and conducted analyses with greater cross-state similarities than in the first round of QRIS validation studies. Something as apparently straightforward as the fact that this set of studies looks at *growth* in child outcomes in light of quality ratings of programs, rather than at a single measurement of child outcomes, represents growth in the field. The state studies in this report may have used differing statistical techniques to look at change over time, but there is clearly a consensus that it is growth—not level of child outcomes at a single point in time—that is critical to consider, whereas level was considered in a number of earlier studies.

The evolution in this area of research is also evident in terms of *which* observational measures of quality are either included in or examined in relation to the QRIS ratings. While early studies focused nearly exclusively on the relation of quality ratings to global measures of quality (such as the ECERS-R, that consider furnishings and types of daily activities in addition to the nature of teacher/caregiver-child interactions), it is exciting to see the greater prevalence here of observational measures that zero in on teacher/caregiver-child interactions ("interaction-specific measures of quality") in validation studies (such as the CLASS measures). Research indicates stronger prediction to child outcomes from interaction-specific, rather than global, measures of quality (although see below for discussion of a further wave of quality measurement focusing on interactions in specific domains, such as stimulation specifically for language development (Burchinal, Zaslow, & Tarullo, 2016).

Given the history of QRIS, this report reflects an exciting emerging consensus as to what the critical validity questions are, the analytic "ground rules" for examining these questions, and measures of quality to consider in such studies.

The core questions for validation studies: Where are we now and what are possible next steps with the data that have been collected in state studies?

Ratings and observed quality. Looking across the results from the ten state validation studies, the findings point to consistent (albeit modest) associations between quality ratings and observed quality. Even though the ten state QRIS are configured in differing ways and include somewhat different measures, we see evidence that summary ratings of quality are related to independent observations of quality, especially in center-based programs. This finding is critical to maintaining the buy-in of programs to being rated, their willingness to have their ratings shared, and their sense that investing in quality will result in legitimate markers of program improvement.

The report raises key questions about what the entry level QRIS ratings in different states represent. For example, in some states, programs that will eventually receive a higher rating are waiting at the entry level for certain requirements to be documented. As a result, the entry level rating actually reflects a range in terms of what observed quality would be likely to show. The dispersion of quality ratings documented in the report also suggests that time since establishment of the QRIS and density may be important to consider. It may take a while in terms of investment of resources in quality improvement for programs to progress toward higher levels and for a state to have a full distribution of ratings, including a substantial number at the highest levels.

Valuable next steps that states might be able to take with the data already collected would be to consider whether the patterns of associations of quality ratings and observed quality appear to differ when all entry-level programs are included (or only entry-level programs with fully completed ratings); when time since state-level implementation of the QRIS and density of participation are taken into account; and when state samples do or do not show a full dispersion in terms of quality ratings, including programs at the highest rating levels.

Ratings and child outcomes. Turning to the second core question of the relation of quality ratings and child outcomes, the report addresses mixed findings. On the one hand, there are encouraging findings here, with some statistically significant findings showing stronger growth in development among children in programs with higher QRIS ratings. But on the other hand, the pattern of significant associations is sparse relative to the pattern of associations considering quality ratings in relation to observed quality. There is one finding that goes counter to predictions, and all associations are modest in strength. Overall, the words used in the report to describe the pattern of findings seem appropriate. The report neither under- or overstates a sparse pattern of modest associations, but notes that the findings generally fall in the predicted direction of greater developmental gains in programs with higher-quality ratings.

What more might the collaborating states do to understand a weak pattern of findings that nevertheless generally aligns with predictions? Taking further steps would require addressing a fundamental and unresolved issue in the field: that of whether we should be calculating a single, overall quality summary score in QRIS, or several different summary scores focusing on different facets of quality. Earlier work (Zaslow & Tout, 2014) asked the question of whether there are outcomes of QRIS that should be considered in addition to child outcomes. These outcomes might include *professionalization of the workforce* and *positive engagement with and by families* from a range of backgrounds. An outcome of QRIS might also include movement toward a *better-integrated early care and education system* in which the same rubric for considering quality is used across different types of early care and education (home-based care, community based child care, Head Start, pre-k) and in which programs of all types receive resources and supports to participate in ongoing quality improvement.

Perhaps weak and inconsistent associations of quality ratings and child outcomes reflect an aggregate quality rating in QRIS that includes too much—that has contributing indicators that are included not because they are seen as strong contributors to child outcomes, but rather because they are key contributors to the other implicit QRIS outcomes. Psychometric work in which mock QRIS were developed using national datasets, limiting the QRIS ratings to just a few quality indicators clearly linked with child outcomes in previous research, shows a fairly consistent pattern of associations between the summary rating of quality and child outcomes (Burchinal et al., 2016; Burchinal, Tarullo, & Zaslow, 2016).

We are left with a number of key questions for the state teams that collaborated in this report, and for the field more broadly: Could a collaborative effort in multiple states work toward agreement upon those indicators in QRIS that should be hypothesized to be related to child outcomes (as opposed to other implicit QRIS outcomes such as family engagement)? Could our tests of associations between quality summary ratings and children's development include only those quality indicators that are known to predict children's outcomes? Or would such an approach be *post-hoc*, risk being a "fishing expedition," and not be something valuable to pursue at this point? This is not a small issue and would need to be fully discussed. A resolution might potentially involve viewing QRIS not as contributing to a single overall summary rating, but instead to multiple quality composite scores, each conceptualized as predicting a different outcome area.

High-priority next steps suggested by this report that would require further data collection. What high-priority next steps does this important report point to that are not bounded by the availability of existing data? While this report clearly reflects a progression from focus on observational measures of quality that are global to measures that focus specifically on the quality of interactions, a further step would involve inclusion of measures of the quality of domain-specific interaction, or measures that assess the quality of interactions supporting development in such specific areas as language development or math skills.

In the validation research in one state (in findings not included among those summarized for the present report), Minnesota found the inclusion of such domain-specific measures to be extremely helpful (Tout et al., 2016). Such measures differentiated Head Start and public school pre-k programs from other programs assigned four stars in the state QRIS, whereas the ECERS-R and CLASS did not show differences across these types of programs. Research including all "generations" of quality measures (global, interaction-specific, and domain-specific) points to the domain-specific measures as the strongest predictors of child outcome (Burchinal, Zaslow, & Tarullo, 2016). When we build observed measures of quality into our QRIS at the higher levels, we may need the more precise information from domain-specific measures to guide quality improvement efforts by programs. We may also need such measures in order to distinguish among programs with greater precision in validation research predicting to child outcomes.

The present study also suggests that an exciting potential next step for states would be to follow individual programs participating in QRIS over time. A key question for states is whether the resources that are being provided for quality improvement are sufficient to help programs progress. States also need to understand whether programs with differing initial characteristics show greater readiness to progress and change. A longitudinal study tracking programs over time, with measures both of initial characteristics and the nature and intensity of quality improvement investments, could help states understand program needs and target resources for quality improvement more effectively. Over time, if programs in such a longitudinal sample progress in quality ratings, once there is sufficient sample size showing improvement in ratings, it might also be possible to ask whether child gain scores differ in the same programs at timepoints when they had received lower and higher overall quality ratings.

References

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

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

Tout, K., Cleveland, J., Li, W., Starr, R., Soli, M., & Bultinck, E. (2016). The Parent Aware evaluation: Initial validation report. Minneapolis, MN: Child Trends.

Zaslow, M. & Tout, K. (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.

References

Achenbach, T., & Edelbrock, C. S. (1983). *Child Behavior Checklist*. In *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: T.M. Achenbach.

Aikens, N., Bush, C., Gleason, P., Malone, L., & Tarullo, L. (2016). *Tracking quality in Head Start classrooms: FACES 2006 to FACES 2014 technical report*, OPRE Report 2016-95. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology 10*(4): 541–552.

Bracken, B. A. (2007). *Bracken School Readiness Assessment-Third Edition: Examiner's Manual*. San Antonio, TX: Pearson.

Burchinal, P., Kainz, K., Cai, K., Tout, K., Zaslow, M., Martinez-Beck, I., & Rathgeb, C. (2009). *Early care and education quality and child outcomes,* Research-to-Policy, Research-to-Practice Brief OPRE #2009-15. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Burchinal, M., Zaslow, M., & Tarullo, L. (2016). Quality thresholds, features, and dosage in early care and education: Secondary data analyses of child outcomes [Commentary by E. Votruba-Drzal & P. Miller]. *Monographs of the Society for Research in Child Development, 81*(2), 1–128. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/mono.v81.2/issuetoc

Clements, D. H., & Sarama, J. (2011). *Tools for early assessment in math (TEAM).* Columbus, OH: McGraw-Hill Education.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do." *Developmental Psychobiology, 29,* 315–334.

Dunn, L. M., & Dunn, D. M. (2007). Peabody Picture Vocabulary Test, Fourth Edition. Minneapolis, MN: NCS Pearson, Inc.

Early Childhood Research Institute on Measuring Growth and Development. (1998). *Research and development of individual growth and development indicators for children between birth to age eight* (Technical report 4). Minneapolis, MN: Center for Early Education and Development.

Elicker, J., Gold, Z., Lane, S., Mishra, A. & Christ, S. (March, 2017). Indiana Paths to QUALITY. Factors predicting quality improvement: Provider characteristics and coaching. Presentation at the Annual Meeting of the Child Care and Early Education Policy Research Consortium. Washington, DC.

Epstein, D., Hegseth, D., Friese, S., Miranda, B., Gebhart, T., Partika, A. & Tout, K. (2017). Quality First: Arizona's Early Learning Quality Improvement and Rating System Implementation and Validation Study. Chapel Hill, NC: Child Trends.

Friese, S., Starr, R., & Hirilall, A. (forthcoming). Understanding and Measuring Participation in Quality Rating and Improvement Systems. OPRE Research Brief. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211.

Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early Childhood Environment Rating Scale, Revised Edition.* New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (2007). *Family Child Care Environment Rating Scale, Revised Edition.* New York: Teachers College Press.

HighScope Educational Research Foundation. (2003). *Preschool Program Quality Assessment, Second Edition (PQA) Administration Manual.* Ypsilanti, MI: HighScope Press.

Joseph, G. E., Feldman, E., Phillips, J. J., & Jackson, E. (2010). The combined CLASS: Assessing the adult-child interactions in mixed age family childcare. A procedure manual. University of Washington.

Karoly, L. A. (2014). Validation studies for early learning and care Quality Rating and Improvement Systems: A review of the literature. Santa Monica, CA: RAND Corporation.

Karoly, L. A., Schwartz, H. L., Setodji, C. M., & Haas, A. C. (2016). Evaluation of Delaware Stars for Early Success: Final report. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR1426.html

Karoly, L. A., Zellman, G. L., & Perlman, M. (2013). *Understanding variation in classroom quality within early childhood centers: Evidence from Colorado's Quality Rating and Improvement System*. *Early Childhood Research Quarterly, 28*(4), 645-657.

LaFreniere, P. J., & Dumas, J. E. (1996). Social competence and behavior evaluation in children ages 3 to 6 year: the short form (SCBE-30). *Psychological Assessment, 8*(4), 369-377.

Lahti, M., Sabol, T., Starr, R., Langill, C., & Tout, K. (2013). *Validation of Quality Rating and Improvement Systems (QRIS): Examples from four states,* Research-to-Policy, Research-to-Practice Brief OPRE 2013-036. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Lamb, M. (1998). Nonparental child care: Context, quality, correlates, and consequences. In W. Damon, I. E. Sigel, & K. A. Renninger (Eds.), *Handbook of child psychology: Child psychology in practice* (5th ed., Vol. 4, pp. 73-133). New York: Wiley.

LaParo, K., Hamre, B., & Pianta, R. (2012*). Classroom Assessment Scoring System: Toddler*. Baltimore, MD: Brookes Publishing.

LeBuffe, P. A., & Naglieri, J. A. (2012). *The Devereux Early Childhood Assessment for Preschoolers, Second Edition (DECA-P2) Assessment, Technical Manual, and User's Guide*. Lewisville, NC: Kaplan.

Lipscomb, S. T., Weber, R. B., Green, B. L., & Patterson, L. B. (2017). Oregon's Quality Rating and Improvement System (QRIS) validation study one: Associations with observed program quality. Retrieved from https://www.pdx.edu/ccf/sites/www.pdx.edu.ccf/files/QRISStudy1report_FINAL_Appendices_Nov17_2016.pdf

Lonigan, C. J., Wagner, R. K., Torgeson, J. K., & Rashotte, C.A. (2007). *Test of Preschool Early Literacy (TOPEL)*. Austin, TX: PRO-ED, Inc.

Magnuson, K., & Lin, Y. (2015). Validation of the QRIS YoungStar rating scale, report 1. Madison, WI: School of Social Work and Institute for Research on Poverty. Retrieved from https://dcf.wisconsin.gov/files/youngstar/pdf/validationreport1.pdf

Magnuson, K., & Lin, Y. (2016). Validation of the QRIS YoungStar rating scale, report 2. Madison, WI: School of Social Work and Institute for Research on Poverty. Retrieved from https://dcf.wisconsin.gov/files/youngstar/pdf/validationreport2.pdf

Mathias, D. (2015). Impact of the Early Learning Challenge on state Quality Rating and Improvement Systems. In The Build Initiative (Ed.), *Rising to the challenge: Building effective systems for young children and families, a BUILD E-Book*. Retrieved from http://www.buildinitiative.org/Portals/0/Uploads/Documents/Chapter8Mathias.pdf

Maxwell, K. L., Blasberg, A., Early, D. M., Li, W., & Orfali, N. (2016). *Evaluation of Rhode Island's BrightStars child care center and preschool quality framework*. Chapel Hill, NC: Child Trends.

McClelland, M. M., & Cameron, C. E. (2012). Self-Regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives*, *6*, 136–142.

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the Preschool Learning Behaviors Scale. *Psychology in the Schools*, *39*, 353–365.

Mitchell, A. W. (2005). *Stair steps to quality: A guide for states and communities developing Quality Rating Systems for early care and education.* Alexandria, VA: United Way of America. Retrieved from http://www.earlychildhoodfinance.org/downloads/2005/MitchStairSteps_2005.pdf

Mitchell, A. W. (2009). *Quality Rating & Improvement Systems as the framework for early care and education system reform*. The Build Initiative. Retrieved from http://www.earlychildhoodfinance.org/downloads/2009/QRISasSystemReform_2009.pdf

Monahan, S., Atkins-Burnett, S., Wasik, B. A., Akers, L., Hurwitz, F., & Carta, J. (2016). *Developing a tool to examine teachers' use of ongoing child assessment to individualize instruction*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Mullen, E. M. (1995). *Mullen scales of early learning.* Circle Pines, MN: American Guidance Service, Inc.

Neuman, S. B., & Dickinson, D. K. (2010). *Handbook of early literacy research.* New York: Guilford Publications.

Pianta, R. C., La Paro, K. M., & Hamre, B. (2008). Classroom assessment scoring system (CLASS): Pre-K version. Baltimore, MD: Brookes Publishing

Quick, H. E., Hawkinson, L. E., Holod, A., Anthony, J., Muenchow, S., Parrish, D., . . . Haggard, M. S. (2016). Independent evaluation of California's Race to the Top – Early Learning Challenge Quality Rating and Improvement System: Cumulative technical report. Santa Monica, CA: RAND Corporation.

Race to the Top – Early Learning Challenge. (2011). Race to the Top – Early Learning Challenge, request for proposals. Washington, DC: U.S. Department of Education, U.S. Department of Health and Human Services.

Schilder, D., Iruka, I., Dichter, H., & Mathias, D. (2015). Quality Rating and Improvement Systems: Stakeholder theories of change and models of practice. Retrieved from http://qrisnetwork.org/resource/2016/quality-rating-and-improvement-systems-stakeholder-theories-change-and-models-practice

Smith, S., Robbins, T., Schneider, W., Kreader, J. L., & Ong, C. (2012). Coaching and quality assistance in Quality Rating Improvement Systems: Approaches used by TA providers to improve quality in early care and education programs and home-based settings. National Center for Children in Poverty: Columbia University, Mailman School of Public Health, Department of Health Policy and Management.

Soderberg, J., Joseph, G. E., Stull, S., & Hassairi, N. (2016). Early Achievers standards validation study: Final report. Washington (State), Department of Early Learning.

Stoney, L. (2012). *Unlocking the potential of QRIS: Trends and opportunities in the Race to the Top Early Learning Challenge applications*. QRIS National Learning Network.

Swanson, C., Carran, D., Guttman, A., Wright, T., Murray, M., Alexander, C., Nunn, J. (2017). *Maryland EXCELS Validation Study*. Johns Hopkins University, School of Education, Center for Technology in Education.

Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2011). *ECERS-E The Four Curricular Subscales Extension to the Early Childhood Environment Rating Scale (ECERS-R), 4th Edition with Planning Notes.* New York: Teachers College Press.

The Build Initiative & Child Trends. (2016). A catalog and comparison of Quality Rating and Improvement Systems (QRIS) [Data System]. Retrieved from http://qriscompendium.org/ on March 1, 2017.

Tout, K., Chien, N., Rothenberg, L., & Li, W. (2014). Implications of QRIS design for the distribution of program ratings and linkages between ratings and observed quality. OPRE Research Brief #2014-33. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., Cleveland, J., Li, W., Starr, R., Soli, M., & Bultinck, E. (2016). The Parent Aware evaluation: Initial validation report. Minneapolis, MN: Child Trends.

Tout, K. & Starr, R. (2013). *Key elements of a QRIS validation plan: Guidance and planning template*. (OPRE 2013-11). Washington, DC: U.S. Administration for Children and Families, Office of Planning, Research and Evaluation. Retrieved from http://www.acf.hhs.gov/sites/default/files/opre/key_elements_of_a_qris_validation_plan_final_2_21_13.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,* OPRE Report. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from http://www.childcareresearch.org/childcare/resources/18554

US Department of Education. (2016, October 29). *Race to the Top – Early Learning Challenge: Purpose and Program Description.* Retrieved from https://www2.ed.gov/programs/racetothetop-earlylearningchallenge/index.html

Wellesley Centers for Women & UMass Donahue Institute Applied Research & Program Evaluation. (2017). Massachusetts Quality Rating and Improvement System (QRIS) validation study: Final report. Massachusetts Department of Early Education and Care. Retrieved from https://pilot.mass.gov/files/2017-08/Revised%20Validation%20Study%20ReportfinalFORMATTED.pdf

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement* (Third Edition). Rolling Meadows, IL: Riverside Publishing.

Zaslow, M. & Tout, K. (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.

Zellman, G. L. & Fiene, R. (2012). *Validation of Quality Rating and Improvement Systems for early care and education and school-age care*, Research-to-Policy, Research-to-Practice Brief OPRE 2012-29. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

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.

Zill, N., & Resnick, G. (2000). *FACES Story and Print Concepts – Version 2.* Rockville, MD: Westat, Inc.

Appendix: Observation and Child Development Tools

Environment Rating Scales (ERS)

The Environment Rating Scales are designed to measure process quality in early childhood education. Process quality is defined as the quality of interactions between children and the various aspects of a classroom learning environment. The **Early Childhood Environment Rating Scale-Revised (ECERS-R)** is intended for use in preschool or kindergarten classrooms, the **Early Childhood Environment Rating Scale-Extension (ECERS-E)** includes a curricular extension to the tool, and the **Family Child Care Environment Rating Scale (FCCERS-R)** is designed to assess the quality of family child care.

Classroom Assessment Scoring System (CLASS)

The Classroom Assessment Scoring System is an observation tool which assesses the quality of emotional support and instruction. There are three subscales of the CLASS-PreK: Emotional Support, Instructional Support, and Classroom Organization. The toddler version has Emotional and Behavioral Support and Engaged Support for Learning subscales.

Caregiver Interaction Scale (CIS)

The Caregiver Interaction Scale is a tool that assesses the interactions between providers and children. The scale measures sensitivity, harshness, detachment, and permissiveness in interactions.

Preschool Quality Assessment (PQA)

The PQA is designed to measure early childhood program quality in 7 key domains assessed through the following subscales: Learning Environment, Daily Routine, Adult-Child Interactions, Curriculum Planning and Assessment, Parent Involvement and Family Services, Staff Qualifications and Development, and Program Management.

Peabody Picture Vocabulary Test (PPVT)

The PPVT is a test of receptive vocabulary for Standard American English.

Test of Preschool Early Literacy (TOPEL)

This instrument is intended to measure children's oral vocabulary, print knowledge, and phonological awareness. The TOPEL was designed to identify preschoolers at risk of literacy problems, thereby allowing for early intervention.

Individual Growth and Development Indicators (IGDI) – Picture Naming

Picture Naming measures expressive language – how many pictures a child can name in a minute.

Woodcock-Johnson-III (WJ-III)

The Woodcock-Johnson test of cognitive abilities is a set of intelligence tests which are also commonly used individually to measure specific skills. The **Letter-Word** subtest measures oral reading skills (e.g., pronunciation), the **Picture Vocabulary** subtest measures object recognition, and the **Applied Problems** subtest measures mathematics skills and quantitative reasoning.

Story and Print Concepts

Story and Print Concepts is an assessment of pre- or early-literacy skills.

Peg/Pencil Tapping

In the Peg/Pencil Tapping test, children must do the opposite of what the experimenter does. The task is designed to measure executive function and inhibitory control over prepotent behaviors.

Head, Toes, Knees, and Shoulders (HTKS)

In HTKS, children are asked to play a game in which they do the opposite of what the experimenter does. Depending on how the child does in the initial part of the task, additional rules are added to the game. The task is intended to measure inhibitory control, working memory, and attention.

Bracken School Readiness Assessment (BSRA)

The Bracken School Readiness Composite assesses basic concepts related to school readiness including colors, letters, numbers/counting, size/comparison, and shapes.

Mullen Scales of Early Learning

The Mullen consists of five individual scales, four of which comprise an overall Early Learning Cognitive Composite, measuring development in the area of cognitive functioning. The four scales are: Expressive language, receptive language, fine motor, and gross motor skills.

Body Mass Index

Body mass index is a measure of physical growth based on height and weight. It provides an estimate of how much body fat a child has.

Social Competence and Behavior Evaluation (SCBE-30)

The Social Competence and Behavior Evaluation short form (SCBE-30) is a teacher report consisting of 30 questions that provide an assessment of preschool emotional adjustment and social competence. Three subscales are measured: Social Competence (emotionally mature, pro-social behaviors), Anger Aggression (oppositional behaviors, poor frustration tolerance), and Anxiety Withdrawal (anxious, depressed).

Preschool Learning Behavior Scale (PLBS)

The Preschool Learning and Behavior Scale (PLBS) persistence subscale is a teacher report checklist that assesses children's observable approaches to learning, specifically attention/persistence.

Devereux Early Childhood Assessment (DECA)

The DECA is a social/emotional development assessment that measures positive behaviors in children indicative of resilience.

Child Behavior Checklist (CBCL)

The CBCL is a caregiver report form used to identify problem behavior in children. Subscales include internalizing behavior, externalizing behavior, and total problems scales.

Tools for Early Assessment in Math (TEAM)

TEAM is a diagnostic tool used to determine where a child is proficient in math knowledge and skills.