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.

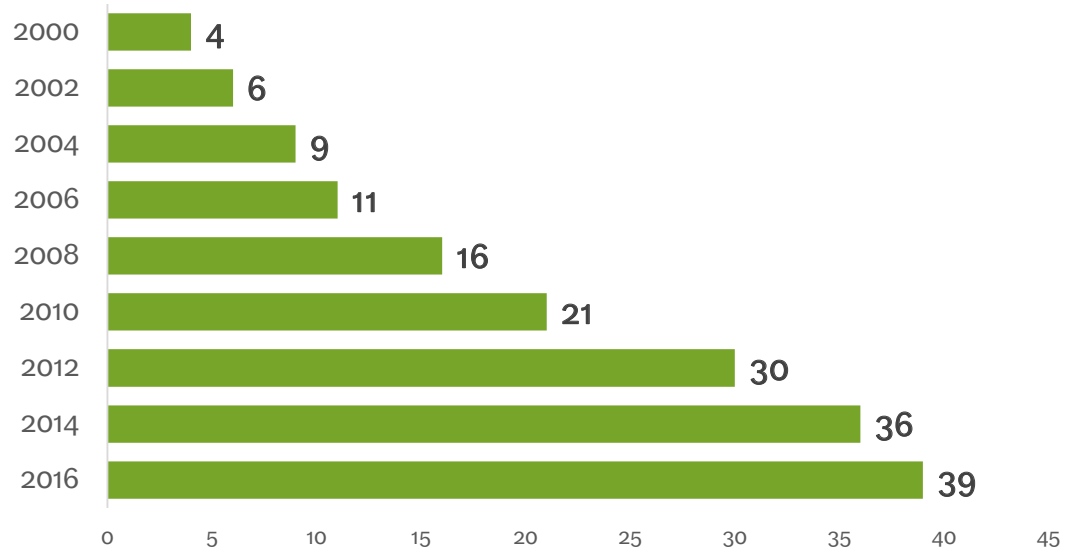
Background on QRIS and the Purpose of Validation Studies

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.