VALIDATION OF QUALITY RATING AND IMPROVEMENT SYSTEMS (QRIS):

Examples from Four States

Research-to-Policy, Research-to-Practice Brief
October 2013

DISCLAIMER:

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

ACKNOWLEDGMENTS

The authors would like to thank Ivelisse Martinez-Beck and Naomi Goldstein at the Office of Planning, Research and Evaluation and Linda Smith at the Administration for Children and Families in the U.S. Department of Health and Human Services. Members of the Quality Initiatives Research and Evaluation Consortium (INQUIRE) are also acknowledged for their contributions to the research perspectives provided in this document.

Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States

Research-to-Policy, Research-to-Practice Brief OPRE 2013-036
October 2013

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

Submitted by:
Michel Lahti, LeCroy & Milligan Assoc., Inc.
Terri Sabol, Institute for Policy Research, Northwestern University
Rebecca Starr, Child Trends
Carolyn Langill, Purdue University
Kathryn Tout, Child Trends

Contract Number: GS10F0030R
Project Director: Kathryn Tout
Child Trends
7315 Wisconsin Ave, Ste 1200W
Bethesda, MD, 20814

Suggested Citation:
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.

This Brief was developed by members of the Quality Initiatives Research and Evaluation Consortium (INQUIRE) which is designed to facilitate the identification of issues and the development and exchange of information and resources related to research and evaluation of quality rating and improvement systems (QRIS) and other quality initiatives. INQUIRE is funded by the Office of Planning, Research and Evaluation through the Child Care and Early Education Policy and Research Analysis and Technical Expertise contract with Child Trends.

Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States

"Validation of a QRIS is a multi-step process that assesses the degree to which design decisions about program quality standards and measurement strategies are resulting in accurate and meaningful ratings" (Zellman & Fiene, 2012, p. 1). In a recent Brief produced through the Quality Initiatives Research and Evaluation Consortium – INQUIRE – Zellman and Fiene (2012) provide a framework to guide QRIS validation and examples of the activities that could be conducted as part of validation efforts. The current Brief serves as a companion to the 2012 INQUIRE Brief by providing detailed examples and findings from the validation activities in four states: Indiana, Maine, Minnesota and Virginia. The purpose of this Brief is to demonstrate how different states have approached QRIS validation, to compare findings, and to highlight challenges in designing and conducting QRIS validation studies.

The picture that emerges from the synthesis of findings across the four states and across the validation approaches is mixed. For instance, the results of efforts to validate the quality standards and indicators in QRIS generally have been successful. Efforts to review how well measures are functioning, however, reveal concerns about limited variation on some measures and QRIS structures that are producing skewed distribution of programs across the rating levels. There are some indications that QRIS levels are distinct with respect to measures of observed quality, but only in the QRIS that used the observational measures as part of the rating process. Finally, validation studies that included measures of children's developmental progress indicate limited support for linkages between these measures of children's growth, QRIS ratings and program quality elements. The findings suggest that further work is needed to strengthen the ability of QRIS ratings to serve as meaningful markers of program quality.

A key theme discussed in the brief is that the information gained from validation efforts can serve as a critical tool for guiding initial design of QRIS, redesign efforts and continuous quality improvement. Zellman and Fiene (2012) emphasize that validation studies do not produce "yes" or "no" answers about QRIS but provide data that can support QRIS in a process of refining and improving. As such, validation efforts must be timed appropriately and aligned with a clear decision-making framework for how the findings will be used. In the four states highlighted in this Brief, researchers partnered with state agency leaders and other QRIS stakeholders to assist in developing a validation plan that could support QRIS development as well as a process for reviewing and interpreting findings so that the results could be applied appropriately. As states continue implementation of QRIS, administrators and stakeholders are encouraged to engage in validation efforts that can inform their systems and move progressively toward the provision of effective services.

Reference

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.

Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States Executive Summary

Introduction

"Validation of a QRIS is a multi-step process that assesses the degree to which design decisions about program quality standards and measurement strategies are resulting in accurate and meaningful ratings" (Zellman & Fiene, 2012, p. 1). In a recent Brief produced through the Quality Initiatives Research and Evaluation Consortium – INQUIRE – Zellman and Fiene (2012) provide a framework for QRIS validation and examples of the activities that could be conducted as part of validation efforts. The current Brief serves as a companion to the 2012 INQUIRE Brief by providing detailed examples and findings from the validation activities in four states: Indiana, Maine, Minnesota and Virginia. The purpose of this Brief is to demonstrate how different states have approached QRIS validation, to compare findings, and to highlight challenges in designing and conducting QRIS validation studies.

Approach 1. Examine the Validity of Key Underlying Concepts: Indiana and Maine

This approach involves examining the elements or concepts that are to be included in program ratings, using empirical and/or expert support.

- **Indiana**: Evaluators referred to the literature and other QRIS reports to review the validity of the QRIS quality standards. The review concluded that 75% of the quality indicators examined had substantial evidence for their validity.
- **Maine**: Evaluators conducted "concept mapping" to determine how parents, providers, and professionals viewed the QRIS standards. They also looked to the literature to support the standards. The concept mapping revealed similarly high ratings of several standards by all raters.

Summary of validation approach 1: Both Indiana and Maine conducted reviews of the concepts to be included in the QRIS ratings. Evaluators in both states found evidence supporting the inclusion of certain concepts in the state QRIS.

Approach 2. Examining the Measurement Strategies Used to Assess Quality: Minnesota and Virginia

The second approach to validation discussed by Zellman and Fiene (2012) is to examine the properties of the measures used to assess quality. This type of validation involves looking at the attributes of individual measures and how they are combined to produce a summary rating of quality. This might include examining the distribution and variance of scores for indicators or looking at the inter-correlations among indicators.

Virginia and Minnesota examined the distribution of indicator scores as a means to determine whether a measure provided enough variability to distinguish levels of quality.

- **Minnesota:** The distributions of scores on QRIS categories and total QRIS rating were varied and generally not normally distributed.
- **Virginia:** Total points and half of the categories (staff qualifications and interactions) were relatively normally distributed. The other two categories (ratio and environment) were less so.
- The skewness found in some indicator scores in both states suggests that those measures may be weak in distinguishing levels of quality in the QRIS.

Virginia and Minnesota also examined the correlations among QRIS indicators.

- **Minnesota and Virginia:** Correlations among indicators ranged from non-significant to moderately high correlations.
- The review of correlations in Minnesota and Virginia revealed that components are related to each other but are not providing duplicative information. The small to moderate correlations provide some confidence that the overall categories of quality measures are contributing unique information to the rating.

Summary of validation approach 2: Both Minnesota and Virginia examined the distributions of scores and the correlations among indicators. They found variation in distributions of scores and small to moderately correlated indicators.

Approach 3. Assess the Outputs of the Rating Process: Indiana, Maine, Minnesota, Virginia

A third approach to QRIS validation involves assessing the actual ratings (the "outputs" of the rating process) to understand the degree to which they are producing levels of quality that are distinct in meaningful ways. The degree to which ratings produced distinct levels of quality were examined across the four states by analyzing how ratings were linked to scores on the Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998), a global measure of preschool classroom quality.

- **Indiana**: Differences in mean ECERS-R scores across QRIS levels were noted. Level 4 programs had higher scores than Levels 1 and 2.
- **Maine**: No significant differences between mean ECERS-R scores across QRIS levels were noted.
- **Minnesota**: Differences in mean ECERS-R scores across QRIS levels were found. Level 2 programs had lower scores than Levels 3 and 4.
- **Virginia**: Differences in mean ECERS-R scores were found across Levels 2, 3, and 4.
- Mean ECERS-R scores fell below the "good" level of quality in all four states.

Summary of validation approach 3: Overall, the cross –state findings indicate that ECERS-R scores can discriminate quality levels when the tool or similar indicators are used in the rating process (as seen in three of the four states). The findings in Maine, showing no correlation between quality levels and ECERS-R scores, provide a cautionary note that quality levels may not be distinct on the ECERS-R if indicators used in the QRIS do not align closely with the measure of environmental quality that is used in the validation process.

Approach 4: Examine How Ratings are Associated with Children's Development: Indiana, Minnesota, Virginia

The fourth approach to validation examines the association between QRIS ratings and children's developmental gains. Validation studies using this approach examine whether the QRIS ratings and quality components that comprise the ratings are related in expected ways to measures of children's development and differences in their patterns of growth. These studies are challenging because they must be conducted with a clear understanding of how the particular QRIS operates and characteristics of the quality components that make up the ratings, methods to account for selection biases (in programs, parents and children), recognition that effect sizes are modest in research examining dimensions of quality and children's development, and data collection that allows for analysis of children's gains rather than point-in-time measurements. It is important to review the results with these challenges in mind.

- **Indiana**: No consistent, strong associations between QRIS quality level and young children's development and learning were found. There were some relations between measures of observed quality and child development.
- **Minnesota:** No systematic evidence of strong relations between quality ratings, measures of program quality and children's developmental progress was found.
- **Virginia:** Some evidence was found for an association between QRIS rating and growth in pre-literacy skills in prekindergarten.
- Limitations and implications of these types of analyses are discussed.

Summary of validation approach 4: States are just beginning to examine the relation between QRIS ratings and child development. Examples from Indiana, Minnesota, and Virginia show the difficulty and limitations of these analyses and, as a result, reveal limited evidence for associations between QRIS ratings and child development.

Conclusion

The picture that emerges from the synthesis of findings across the four states and across the validation approaches is mixed. For instance, the results of efforts to validate the quality standards and indicators in QRIS generally have been successful. Efforts to review how well measures are functioning, however, reveal concerns about limited variation on some measures and QRIS structures that are producing skewed distribution of programs. There are some indications that QRIS levels are distinct with respect to measures of observed quality, but only in the QRIS that used the measures as part of the system. Finally, validation studies that included measures of children's developmental progress indicate only limited support for linkages between these measures of children's growth, QRIS ratings and program quality elements. The findings suggest that further work is needed to strengthen the ability of QRIS ratings to serve as meaningful markers of program quality.

A key theme discussed in the brief is that the information gained from validation efforts can serve as a critical tool for guiding initial design of QRIS, redesign efforts and continuous quality improvement. Zellman and Fiene (2012) emphasize that validation studies do not produce "yes" or "no" answers about QRIS but provide data that can support QRIS in a process of refining and improving. As such, validation efforts must be timed appropriately and aligned with a clear decision-making framework for how the findings will be used. In the four states highlighted in this Brief, researchers partnered with state agency leaders and other QRIS stakeholders to assist in developing a validation plan that could support QRIS development as well as a process for reviewing and interpreting findings so that the results could be applied appropriately. As states continue implementation of QRIS, administrators and stakeholders are encouraged to engage in validation efforts that can inform their systems and move progressively toward the provision of effective services.

Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States

Goals of this Brief

Quality ratings are the key output of Quality Rating and Improvement Systems (QRIS) for early care and education and school age care (ECE-SAC) programs. Ratings are used to guide quality improvement supports, the provision of incentives and reimbursement levels to programs and to support parental choice of ECE-SAC programs. Given their central role in QRIS, it is important that processes are in place to ensure the integrity of the ratings. "Validation of a QRIS is a multi-step process that assesses the degree to which design decisions about program quality standards and measurement strategies are resulting in accurate and meaningful ratings" (Zellman & Fiene, 2012, p. 1). Among the pioneer states that launched QRIS in the late 1990's, select validation activities were conducted to examine how ratings were functioning (Bryant, 2001; Norris, Dunn & Eckert, 2003). With over half of the states now implementing QRIS and more systems expected to be piloted or launched in the next five years, validation of QRIS has taken on greater importance. The information gained from validation efforts can serve as a critical tool for guiding initial design of QRIS, redesign efforts and continuous quality improvement.

In a recent Brief produced through the Quality Initiatives Research and Evaluation Consortium – INQUIRE – Zellman and Fiene (2012) provide a framework for QRIS validation and examples of the activities that could be conducted as part of validation efforts. **The current Brief serves as a companion to the 2012 INQUIRE Brief by providing detailed examples and findings from the validation activities in four states: Indiana, Maine, Minnesota and Virginia. The purpose of this Brief is to demonstrate how different states have approached QRIS validation, to compare findings, and to highlight challenges in designing and conducting QRIS validation studies.**

QRIS Validation in Four States

The QRIS validation framework described by Zellman & Fiene includes four related approaches and activities: (1) examining the validity of key underlying concepts, (2) examining the measurement strategy and the psychometric properties of the measures used to assess quality, (3) assessing the outputs of the rating process, and (4) examining how ratings are associated with children's development. Table 1 (recreated from Zellman & Fiene, 2012) provides a description of these four related approaches.

Table 1. Four Related Approaches to Validating a QRIS (from Zellman & Fiene, 2012)

Approach	Activities and Purpose	Typical Questions	Issues and Limitations
1. Examine the validity of key underlying concepts	Assess whether basic QRIS quality components and standards are the "right" ones by examining levels of empirical and expert support.	Do the quality components capture the key elements of quality? Is there sufficient empirical and expert support for including each standard?	Different QRISs may use different decision rules about what standards to include in the system.
2. Examine the measurement strategy and the psychometric properties of the measures used to assess quality	Examine whether the process used to document and verify each indicator is yielding accurate results. Examine properties of key quality measures, e.g., inter-rater reliability on observational measures, scoring of documentation, and inter-item correlations to determine if measures are psychometrically sound. Examine the relationships among the component measures to assess whether they are functioning as expected. Examine cut scores and combining rules to determine the most appropriate ways to combine measures of quality standards into summary ratings.	What is the reliability and accuracy of indicators assessed through program administrator self-report or by document review? What is the reliability and accuracy of indicators assessed through observation? Do quality measures perform as expected? (e.g., do subscales emerge as intended by the authors of the measures?) Do measures of similar standards relate more closely to each other than to other measures? Do measures relate to each other in ways consistent with theory? Do different cut scores produce better rating distributions (e.g., programs across all levels rather than programs at only one or two levels) or more meaningful distinctions among programs?	This validation activity is especially important given that some component measures were likely developed in low-stakes settings and have not been examined in the context of QRIS.

Approach	Activities and Purpose	Typical Questions	Issues and Limitations
3. Assess the outputs of the rating process	Examine variation and patterns of program-level ratings within and across program types to ensure that the ratings are functioning as intended. Examine relationship of program-level ratings to other quality indicators to determine if ratings are assessing quality in expected ways. Examine alternate cut points and rules to determine how well the ratings distinguish different levels of quality.	Do programs with different program-level ratings differ in meaningful ways on alternative quality measures? Do rating distributions vary by program type, e.g., ratings of center-based programs compared to ratings of home-based programs? Are current cut scores and combining rules producing appropriate distributions across rating levels?	These validation activities depend on a reasonable level of confidence about the quality components, standards and indicators as well as the process used to designate ratings.
4. Examine how ratings are associated with children's development.	Examine the relationship between program-level ratings (and components of the ratings) and selected measures of children's development to determine whether higher program ratings (or components of the ratings) are associated with developmental gains.	Do children who attend higher-rated programs have greater gains in skills than children who attend lower-quality programs?	Appropriate demographic and program level control variables must be included in analyses to account for selection factors. Studies could be done on child and program samples to save resources. Findings do not permit attribution of causality about QRIS participation but inferences can be made about how quality influences children's outcomes.

This Brief reviews QRIS validation activities in four states: Indiana, Maine, Minnesota and Virginia. These states represent a cohort of state QRIS validation studies completed between 2008 and 2011. In addition, the evaluators working in these states participate in the Quality Initiatives Research and Evaluation Consortium (INQUIRE) sponsored by the Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families in the U.S. Department of Health and Human Services . INQUIRE serves as a learning community for researchers conducting research on QRIS and other quality improvement initiatives and provided a hub for collaboration on the Brief. The four states represent different QRIS designs [for example, Indiana and Maine used a building block approach (in which all standards at one level must be met before moving to the next level), while Virginia and Minnesota used a point system (points are earned for each standard and then added up; each level represents a range of points)]and features. The cross-state comparison was conducted with the expectation that the comparisons would result in interesting contrasts as well as areas of similarity across the four QRIS.

Information about each state's QRIS was collected from the state evaluators and from written materials including the April 2010 *Compendium of Quality Rating Systems* (Tout et al., 2010). These background details about each QRIS are provided in the appendix. It is important to note that the current features for each QRIS may be different than the features described in the appendix that were in place at the time the analyses for this Brief were conducted.

Key Details, Similarities and Differences across the Four QRIS

In order to provide context for states' validation activities, we first present the four states' QRIS structures, processes, participation rates, and distributions of quality across programs as of 2011 (when validation reports were completed for each state).

- Indiana implemented QRIS statewide in 2007. Maine launched a QRIS statewide in 2008. Minnesota and Virginia implemented pilots that ended in 2011 (though each system was expanded or implemented statewide beginning in 2012).
- Indiana and Maine used a "building block" structure, while Minnesota and Virginia piloted a "points" structure.
- Indiana, Maine and Minnesota had four levels in their QRIS. Virginia piloted five levels.
- Indiana had a rating period of one year as did Minnesota. Virginia had a rating period of two years. Ratings in Maine were certified for three years.
- The states varied in their enrollment and density of participation. Table 2 presents the numbers of programs participating in each QRIS and the percent participating with respect to the number of eligible programs as of April 2011. The percent of participating programs out of eligible programs, or "density", is an important indicator of QRIS effectiveness in penetrating the early care and education market. There is a good deal of variation across the four states in enrollment and density of participation.

Table 2 – QRIS Program Enrollment and Density (April 2011)

	Total Number of Participating Programs	Number of Centers, including Head Start	Number of Family Child Care Programs	Number of Other Programs* Participating	Density: Percent of Participating Programs Out of Total Eligible
INDIANA	2,019	490	1,494	35	46%
MAINE	885	410	475	0	40%
MINNESOTA	354	226	79	49	26%
VIRGINIA	285	285	N/A	N/A	8%

*Other programs include unlicensed child-care ministries (Indiana), school-age programs (Maine), and School Readiness programs (Minnesota). The Virginia pilot served only center-based programs.

The distribution of programs across quality levels also varied across the states. Table 3 provides the approximate enrollment by quality level in April 2011. It is notable that the two QRIS with building block designs (Indiana and Maine) had a distribution of ratings skewed toward the lower quality levels, whereas the two QRIS with points structures (Minnesota and Virginia), had a distribution of ratings skewed toward the higher quality levels. A similar pattern was noted in the Compendium which looked across 26 QRIS (Tout et al., 2010). The pattern suggests that the design of the QRIS is an important contributor (though not the sole determinant) to the distribution of quality ratings, and must be considered in the validation of the QRIS. It is important to identify the expected distribution of programs in the QRIS and to examine the extent to which the actual distribution represents a rating process that is too difficult or too simple for programs or whether the participants in QRIS are similar on selection characteristics (such as initial level of quality). Validation studies conducted in QRIS with high density of participation and even distribution of programs across program types and tiers will likely produce different findings than studies conducted in QRIS with low enrollment, low density of participation and limited distribution of programs among the tiers of program quality.

Table 3 – QRIS Program Enrollment by Star/Step Level (April 2011)

	Number and Percentage of All Program Types at each Star / Step Level	Density: Percent in Top Two Levels
INDIANA	Level 1 – 1262 35% Level 2 – 767 21% Level 3 – 577 16% Level 4 – 445 12%	28% in Top Two Levels
MAINE	Step 1 – 517 58% Step 2 – 169 19% Step 3 - 85 9% Step 4 - 154 17%	26% in Top Two Levels
MINNESOTA	1 Star – 5 1% 2 Stars – 26 6% 3 Stars – 53 13% 4 Stars – 330 80%	93% in Top Two Levels
VIRGINIA	Piloting in 2011	Piloting in 2011

State Examples and Cross-State Comparisons of Four Validation Approaches

In the next sections, we present state examples of studies conducted in Indiana, Maine, Minnesota and Virginia that illustrate the four validation approaches described by Zellman and Fiene (2012). Each section concludes with a summary of key lessons learned about each approach by looking across four states.

Approach 1. Examine the Validity of Key Underlying Concepts

This approach involves the examination of the elements or concepts that are to be included in program ratings. States typically refer to these as their quality standards and indicators. This approach is a logical first step to take in the design of a QRIS because it results in the core set of standards that will be measured in a QRIS. It is also an activity that can be done on a periodic basis to inform revisions to a QRIS. As new findings or best practices emerge in the literature on ECE-SAC program quality, they can be tapped to inform QRIS design or revision.

Among the four states included in this Brief, Indiana and Maine conducted validation activities using this approach.

Indiana

Purdue University's review of the quality standards in Paths to QUALITY was done to assess the "scientific validity" of the program quality standards (see Elicker et al, 2007). The evaluators referred to the published literature and other reports on statewide QRIS. They defined a "quality indicator" as:

1. a concrete, observable, or otherwise documentable aspect of child care settings or practices;

2. a feature identified as a "best practice" in national policies or professional position statements; and

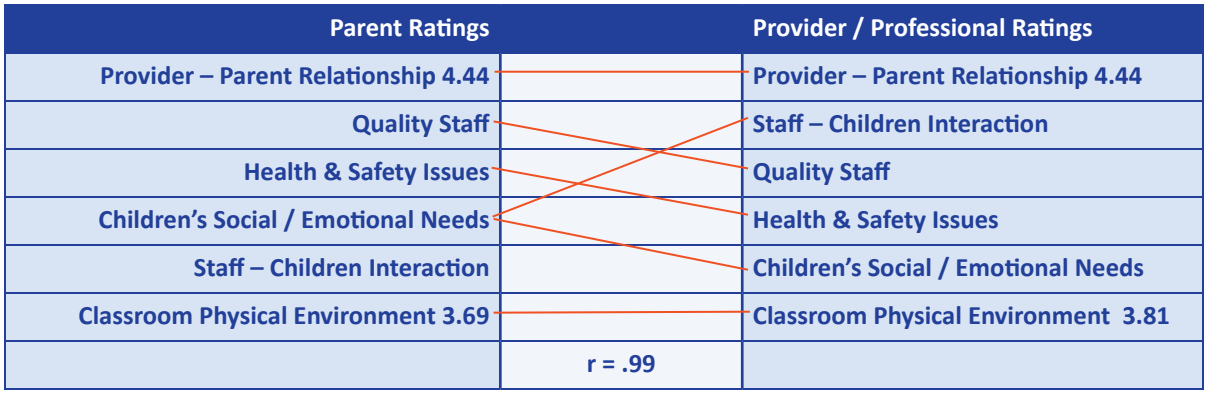
3. a feature that has been evaluated specifically in the published scientific early education and child care literature.

The review concluded that 75% of the quality indicators examined had substantial evidence for their validity. The results provided a basis for QRIS administrators to proceed with Paths to QUALITY. The report was used with QRIS stakeholders to demonstrate the research basis for the QRIS (Elicker et al., 2013).

Maine

Two years before the QRIS was piloted, researchers at the University of Southern Maine conducted a concept mapping (Trochim, 2012) process to guide selection of program quality standards. This process involved; 1) generation of statements relevant to program quality standards that parents and child care providers reflected upon through focus groups and interviews, 2) sorting and rating of statements by participants and expert panelists, 3) computation of maps, and 4) interpretation and utilization of maps for identifying program quality components and standards. Figure 1 presents a "pattern match" that was generated through the concept mapping process which illustrates how parents and providers/ professionals were similar and different in their rating of program quality components. This type of information is helpful for QRIS program designers because it shows areas of agreement and importance. The ratings were done on a five point scale. The highest rated concept, by both groups, was "Provider-Parent Relationship" with a mean of 4.44. The lowest rated concept, also by both groups, was "Classroom Physical Environment" with a mean of 3.69 for parents and 3.81 for providers. For the rest of the concepts, the order is different comparing parents to professionals, with "Staff-Children Interaction" the most different in terms of order. However, this type of data display also indicates that the size of difference between these concepts is not very large (from 3.69 to 4.44 for parents and just 3.81 to 4.44 for professionals). Results from the concept mapping were used to identify specific program quality standards, as well as a literature review process similar to that described for Indiana's QRIS.

Figure 1 – Maine Parent and Provider Pattern Match



Summary of Approach 1

In both Indiana and Maine, validation activities to examine the key concepts or standards in the QRIS were conducted to inform design of the systems. However, as described in the following sections, these activities were part of a larger set of validation activities that were conducted. Validation of key concepts provides important information but typically has not served as the only source of validation for a QRIS.

Approach 2. Examining the Measurement Strategies Used to Assess Quality

The second approach to validation discussed by Zellman and Fiene (2012) is to examine the properties of the measures used to assess quality. This type of validation involves looking at the attributes of individual measures and how they are combined to produce a summary rating of quality. The overall question is whether each measure of quality, such as a QRIS indicator or a score from an observational tool, measures what it is intended to measure and whether it contributes uniquely to a summary rating.

One strategy to employ in this approach is to examine the distribution and variance of scores for a given indicator or set of indicators. If distributions are skewed or there is insufficient variance, then it is likely that the measures will not distinguish meaningful levels of quality.

A second approach is to examine the inter-correlations among the indicators. The strength of correlations between indicators allows researchers to determine whether a given indicator is contributing unique information in measuring quality. That is, if two indicators are highly correlated with each other, they are providing similar information and may not both be necessary in the rating. According to Zellman et al. (2008), indicators ideally would be moderately correlated showing that they are related to each other but are not providing redundant information for the rating.

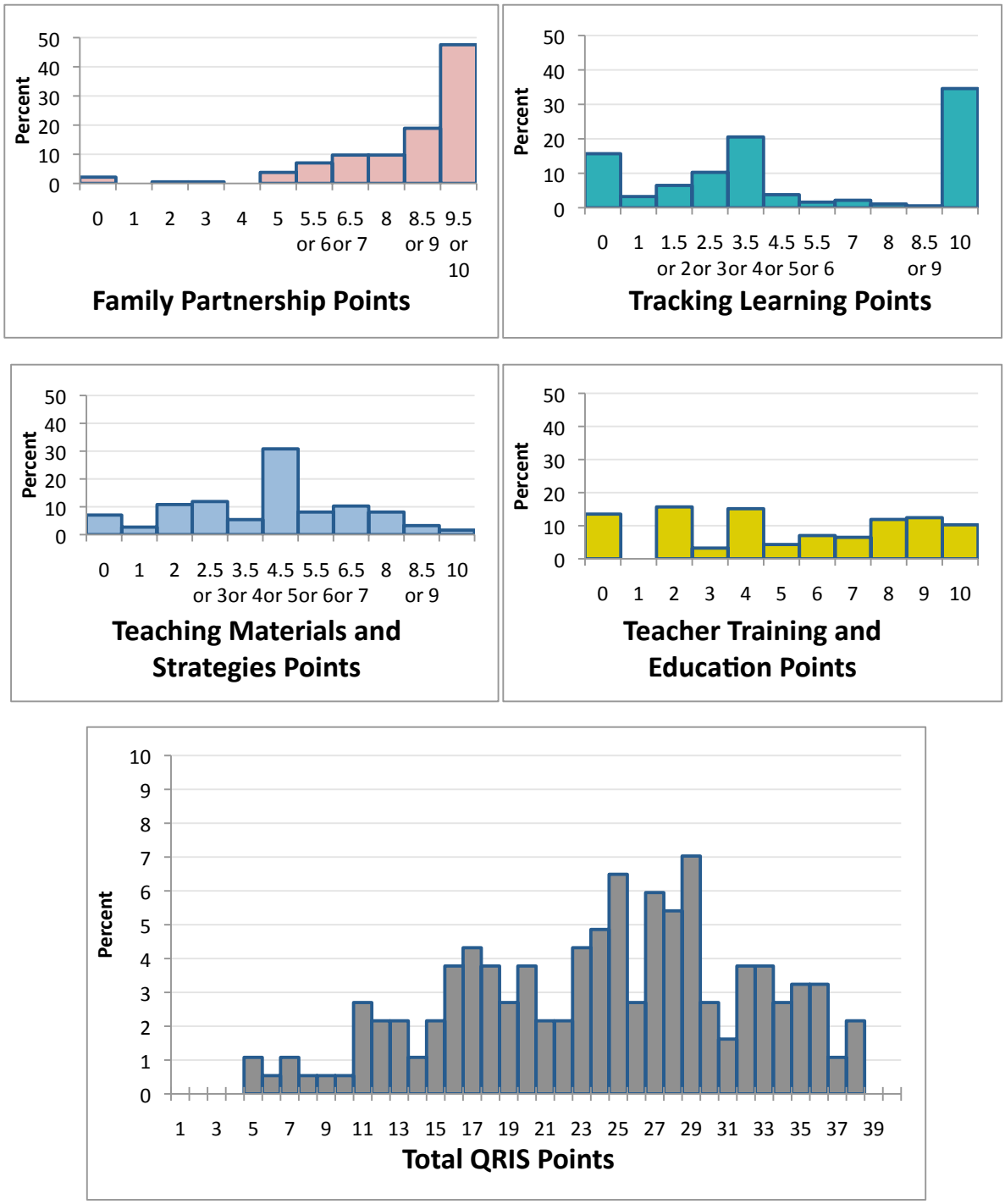
In this section, we present information on how two strategies – examination of the variance within measures of quality and the inter-correlations among measures of quality – were used in the Virginia and Minnesota validation studies.

Distribution of Measures and Variance

Virginia and Minnesota examined the distribution of indicator scores as a means to determine whether a measure provided enough variability to distinguish levels of quality. This analysis could not be done in Indiana and Maine because the building block structure did not provide the necessary data elements.

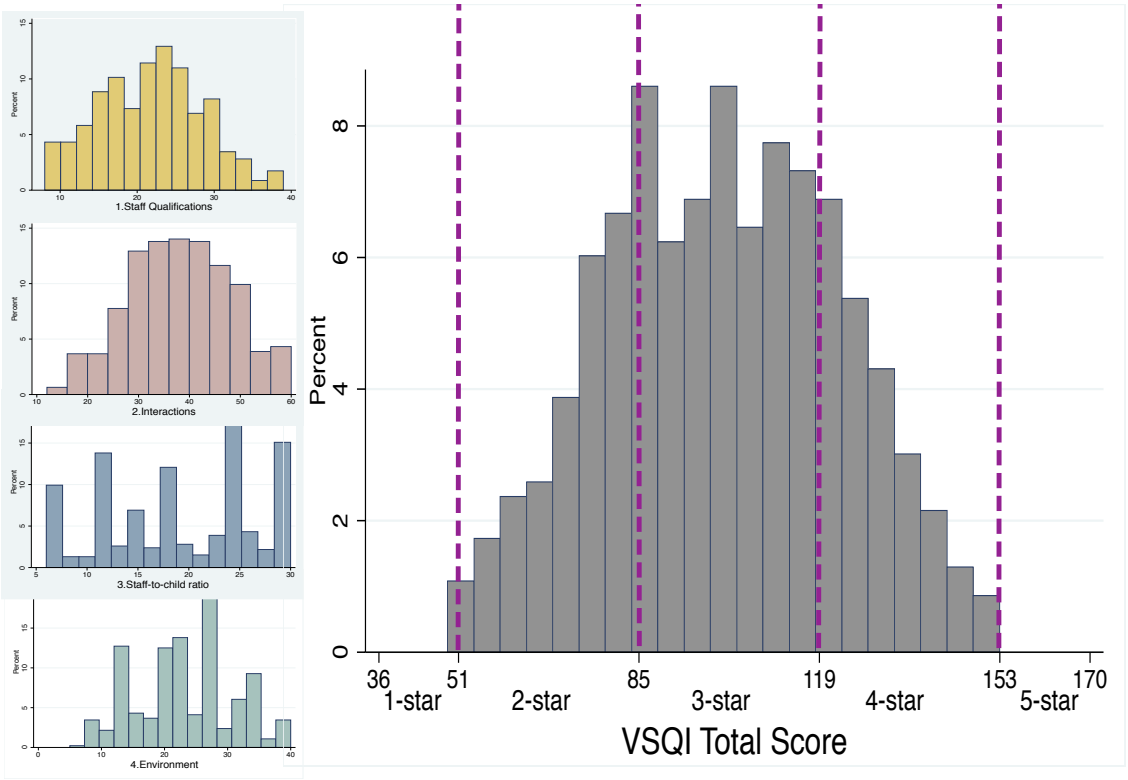
The distributions for Minnesota's QRIS categories and total QRIS points are presented in Figure 2. The distributions varied and were not normally distributed. Family Partnerships, in particular, was skewed to the high end of the points scale. For Tracking Learning, programs either tended to score relatively low or they earned all 10 points. Teaching Materials and Strategies had the best approximation of a normal distribution, while Teacher Training and Education had a flat distribution.

Figure 2. Distribution of indicator category and total points (Minnesota): Percent of programs scoring at each point level.



The distributions for Virginia's QRIS categories and total QRIS score are presented in Figure 3. Whereas Total Points, Staff Qualifications, and Interactions were relatively normally distributed, points for Ratio and Environment were less so.

Figure 3. Distribution of indicator category and total points (Virginia)



Analysis of point distributions provides important insights into how sets of indicators are working to assess quality. In Virginia, the distributions indicated potential issues in the areas of Ratios and Environment. In Minnesota, the Family Partnerships component was an area for further investigation as the scores were skewed to the high end of the scale. These findings can identify weaknesses in the measures themselves or in the strategy used for measurement. Prior to conducting a review of distributions, it is useful to assess characteristics of the programs in the QRIS so that the context for the review is clear.

Correlations among QRIS Indicators

Another strategy to examine the quality measures is to assess the relationship of indicators within the QRIS to other indicators within the QRIS and to indicators external to the QRIS. The purpose of these analyses is to determine if indicators are related to each other in ways that make sense. Correlation matrices were reviewed in both Minnesota and Virginia.

The correlations among categories of QRIS standards for center-based programs were examined in Minnesota and are presented in Table 4. Correlations among indicators in Minnesota ranged from non-significant to moderately high correlations. Family Partnerships was significantly correlated with Teaching Materials and Strategies (environment) and Tracking Learning (child assessment), but not with Teacher Training and Education. Teaching Materials and Strategies was also highly correlated with both Tracking Learning and Teacher Training and Education. Finally, Tracking Learning and Teacher Training and Education were significantly correlated.

Table 4. Correlations for Indicator Categories for Center-based Programs in Minnesota's QRIS

	1. Family Partnerships	2. Teaching Materials and Strategies	3. Tracking Learning	4. Teacher Training and Education
1. Family Partnerships	--			
2: Teaching Materials and Strategies	0.47*	--		
3. Tracking Learning	0.33*	0.58*	--	
4. Teacher Training and Education	0.21	0.49*	0.49*	--

*p < .05

For family child care programs, most of the correlations between indicator categories were also significant (see Table 5) with some differences from the patterns for child care centers. For example, Teacher Training and Education was significantly correlated with Family Partnerships, but was not correlated with Tracking Learning, whereas the opposite was true for center-based programs.

Table 5. Correlations for Indicator Categories for Family Child Care Programs in Minnesota's QRIS

	1. Family Partnerships	2. Teaching Materials and Strategies	3. Tracking Learning	4. Teacher Training and Education
1. Family Partnerships	--			
2: Teaching Materials and Strategies	0.33*	--		
3. Tracking Learning	0.30*	0.38*	--	
4. Teacher Training and Education	0.30*	0.41*	0.16	--

Correlations among the points awarded in each category of QRIS standards were also examined in Virginia. As shown in Table 6, the correlations ranged from non-significant to moderately high correlations. There was no significant correlation between staff qualifications and staff- to-child ratio. Staff qualifications were significantly and positively correlated with the environment and the quality of interactions. The quality of interactions was moderately correlated with staff-to-child ratio, indicating that higher quality of interactions was slightly higher in classrooms with lower staff-to-child ratio. The measures of environment quality (ECERS-R) and the quality of interactions (CLASS) were positively and significantly correlated, which matches previous work with raw scores indicating a moderately strong correlation between the two process-oriented measures (Mashburn et al., 2008; Pianta, La Paro, & Hamre, 2008).

Table 6. Correlations for indicator categories for center-based programs in Virginia's QRIS

	1. Staff Qualifications	2. Interactions	3. Staff-to-child ratio	4. Environment
1. Staff Qualifications	--			
2: Interactions	0.24*	--		
3. Staff-to-child ratio	0.07	0.28*	--	
4. Environment	0.42*	0.61*	0.10*	--

The review of correlations in Minnesota and Virginia revealed that components are related to each other but are not providing duplicative information. The small to moderate correlations provide some confidence that the overall categories of quality measures are contributing unique information to the rating.

Approach 3: Assess the Outputs of the Rating Process

A third approach to QRIS validation involves assessing the actual ratings (the "outputs" of the rating process) to understand the degree to which they are producing levels of quality that are distinct in meaningful ways. In this Brief, we examine the rating outputs in four states by analyzing how ratings were linked to scores on the Early Childhood Environment Rating Scale – Revised (ECERS-R; Harms, Clifford, & Cryer, 1998), a global measure of preschool classroom quality. The ECERS-R assesses structural components of the classroom, such as the physical environment and basic care of children, as well as more process-oriented components, such as the interactions among staff, children and parents. The ECERS-R is the most widely used measure of preschool classroom quality and has been used as a standard measure of quality to which other measures can be compared and validated (e.g. Burchinal, Kainz, and Cai, 2011; Bryant et al., 2003). Results from analyses across the four states address whether the levels of each state's QRIS were associated with progressively higher ECERS-R scores at higher levels of quality. We focus only on center-based programs because comparable data were available across the four states.

It is important to note that Virginia and Minnesota included the ECERS-R as part of the QRIS rating. In Minnesota, the ECERS-R counted for up to 10% of the total points in the rating system. In Virginia, the ECERS-R counted for around 15% of the total points. Indiana and Maine did not use the ECERS-R to determine program ratings, but rather collected ECERS-R data as part of the validation study. This variation presents an interesting comparison between states that use the ECERS-R in the ratings and states that do not. Before examining the relation between ECERS-R and program ratings, we present basic descriptive statistics, including number of observations, mean, standard deviation and distribution, on the total ECERS-R scores in each state (see Figure 4) in order to understand the extent to which ECERS-R scores are comparable among states.

In Indiana, researchers conducted ECERS-R observations in 90 randomly selected child care centers that were rated in the QRIS from 2008-2011. One classroom in each center was randomly selected for observation. The average ECERS-R total score was 4.21 (SD=0.71; Range 1.70-5.74). Almost half of all programs received a "4", which is considered between minimal and good according to the ECERS-R authors (Harms, Clifford, & Cryer, 1998), with very few programs receiving very low (1-2) or very high (6-7) total score .

In Maine, researchers randomly selected 106 classrooms for ECERS-R observations from 2008-2011. On average, 1.37 classrooms were observed at each center. The average ECERS-R total score was 4.25 (SD=.70, Range 2.49-6.06). A score of "5" is considered good quality on this scale. The distribution of ECERS-R scores indicates that the majority of classrooms (62%) scored at or below 4.50, 33% scored between 4.51 and 5.50 and 5% scored above a 5.51.

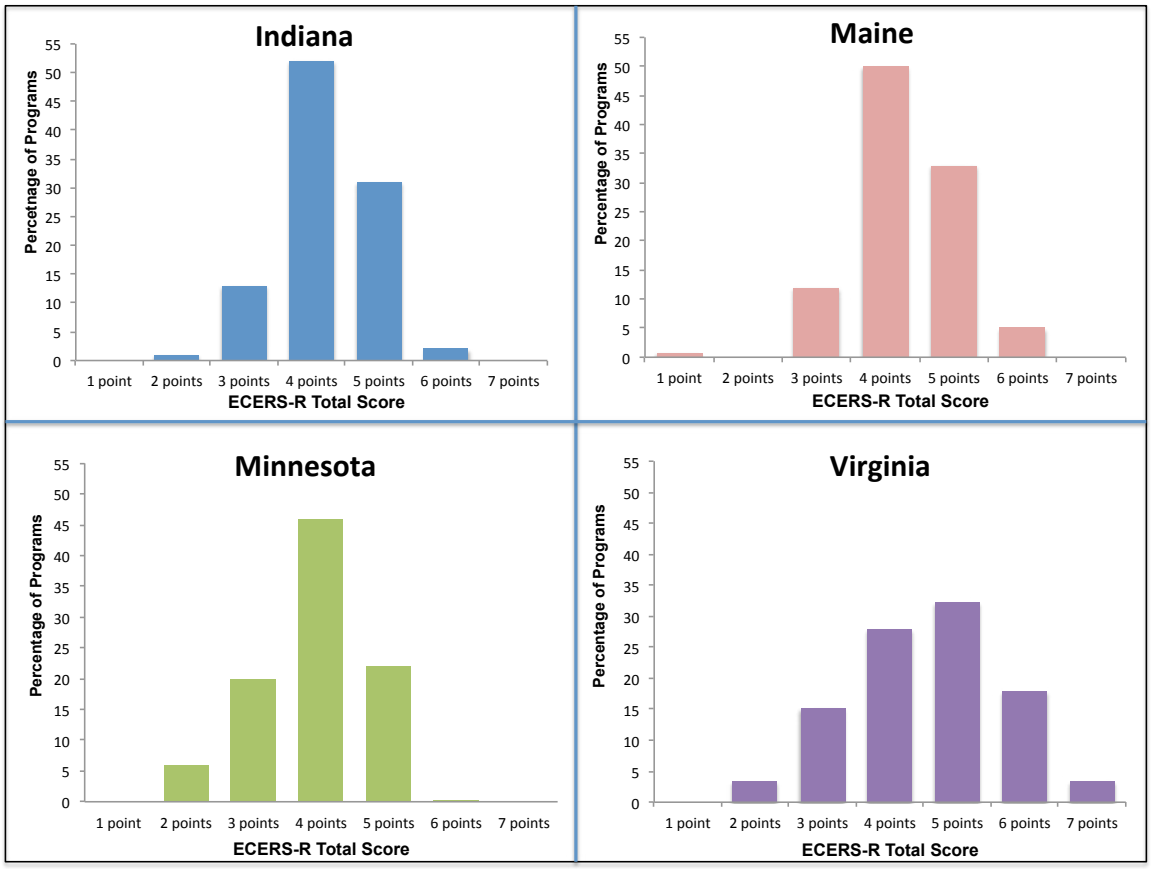
In Minnesota, the ECERS-R is included in the rating process for non-accredited centers completing the full rating process; the evaluation team collected the ECERS-R in accredited centers so that comparable data would be available.[1] The average ECERS-R total score across all ratings of center-based programs rated in years 2008-2011 was 3.82 (SD=0.79, Range 2.17-5.81). Only 25% of the programs received scores below a 4, with the majority of programs scoring between 4 and 6.

In Virginia, the average ECERS-R total score for preschool programs in Virginia's QRIS (2007-2009) was 4.55 (SD=1.12, Range 1.7-6.9; n=350), which is considered moderately good according to the ECERS-R authors. On average, 1.38 classrooms were observed in each center (SD=0.66; Range 1-4). The distribution of the ECERS-R scores demonstrates a fair amount of variation in VSQI programs' environmental quality, with around half of the centers demonstrating good or excellent quality and half demonstrating minimal or moderate quality.

[1] During the pilot, accredited programs in Minnesota received an automatic 4-star rating.

Overall, the four states had relatively normally distributed ECERS-R scores. The means among the four states ranged from 3.82-4.55, indicating that the average classroom among the four states falls within the minimal-good quality range. Virginia demonstrated the most variation (SD=1.12), whereas the remaining three states had most of the classrooms close to the mean (range SD=0.70-0.79). Despite these differences, the four states appear to have relatively comparable levels of environmental quality among center-based programs in the QRIS.

Figure 4. Cross-State Comparisons of the Distribution of ECERS-R Total Scores



For each state, we conducted a Pearson's correlation between QRIS rating levels and ECERS-R total score. Correlations ranged from 0.24 (Maine) to 0.67 (Virginia), indicating a somewhat large variation in the magnitude of the correlation (see Table 7).

Table 7. Correlations between ECERS-R Total Score and Program Rating within Each State

ERS Total Score

	Indiana	Maine	Minnesota	Virginia
Program Rating	0.45***	0.24**	0.44*	0.67***

*p<.05 **p<.01 ***p<.001

In addition, for each state, an analysis of variance (ANOVA) was performed to compare the mean ECERS-R scores in each rating level. When the ANOVA was significant ($p<.05$), follow-up tests were used to test significance of the means of the program levels using a family-wise error rate of .05. The mean level findings are presented in Figure 5 with superscripts to indicate the significant differences between levels within each state. Each level was required to have at least 10% of the programs in order to be included in the analysis. For instance, in Virginia, only 1 program received a Level 5 rating, and only 2 programs received a Level 1 rating, so Level 1 and Level 5 programs were rounded to the nearest level. In Minnesota, only 4 programs received a Level 1 rating and therefore Level 1 was omitted from the analysis.

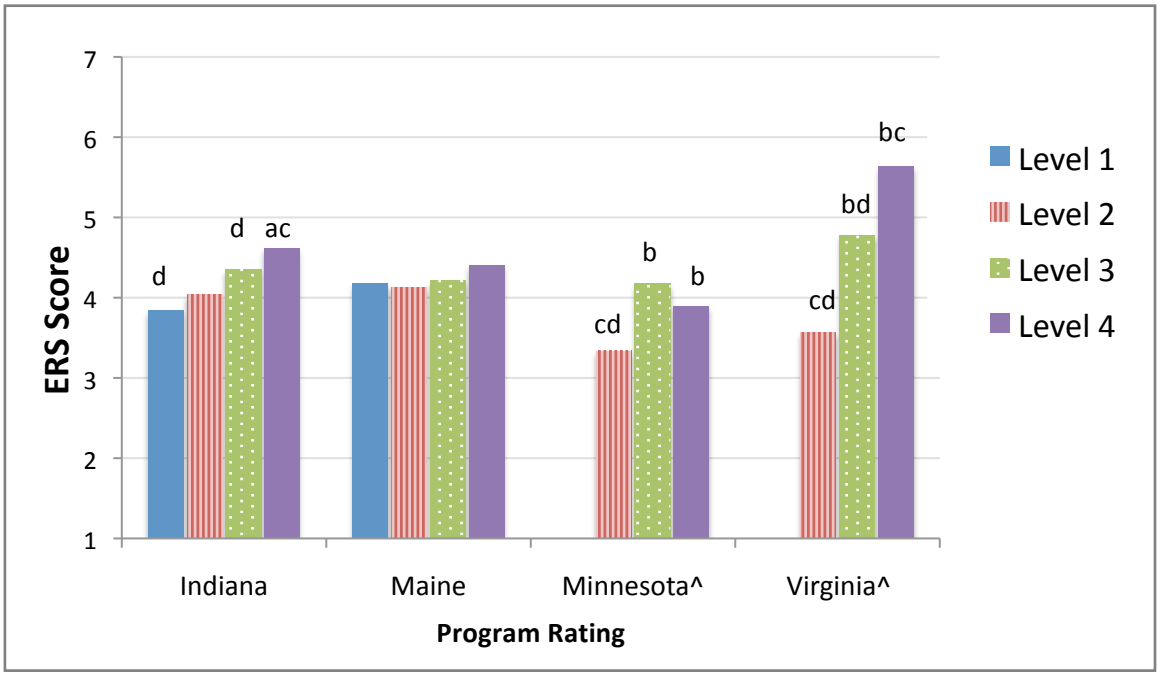
In Indiana, the ANOVA revealed a statistically significant difference in mean ECERS-R total scores across levels ($F(3, 86) = 4.96$, $p<.01$). Post-hoc analysis indicated that Level 4 (highest level) programs had higher ECERS-R scores compared to Level 1 and Level 2 programs. Level 3 was not significantly different from the remaining three levels (Level 1, 2 or 4). Interestingly, observed ECERS-R quality, while related to the QRIS ratings, was highly variable within each level. For example, preschool classrooms at Level 1 had an average global quality score of 3.8, but a range of 1.7 to 5.5. Level 4 preschool classrooms had an average global quality score of 4.6, but ranged from 2.9 to 5.7.

In Maine, the ANOVA failed to find a significant difference between mean ECERS-R scores across levels ($F(3, 130)1.051$, $p=.372$). As with Indiana, observed ECERS-R quality was highly variable within each level. Level 1 programs scored 4.18 (SD=0.73; range 2.97-6.06; n=35), and Level 4 programs scored 4.41 (SD=0.71, range 2.49 – 6.06, n=38).

In Minnesota, the ANOVA indicated statistically significant differences in mean scores across Levels ($F(3, 116) = 5.83$, $p < .01$). There was also some evidence in post hoc analyses for between-level differences. Level 2 programs had significantly lower ECERS-R scores than Level 3 and Level 4. There was no significant difference between Level 3 and Level 4 ECERS-R scores. Level 2 programs scored 3.33 (SD=0.78; range 2.17-4.86, n=28), and Level 4 programs scored 3.90 (SD=0.70, range 2.67- 5.25, n=67).

In Virginia, an ANOVA indicated a significant difference in ECERS-R scores across Level 2, Level 3, and Level 4 centers ($F(2,137)=184.01$, $p<.001$). A follow-up comparison indicated significant differences between all levels. For example, Level 3 programs had significantly higher ECERS-R scores than Level 2 programs, and significantly lower ECERS-R scores than Level 4 programs. Level 1 programs scored 3.57 (SD=0.87; range 1.67-5.70, n=111), and Level 4 programs scored 5.64 (SD=0.73, range 4.0- 6.76, n=58).

Figure 5. Cross-State Comparison of Average ECERS-R by Quality Rating Level



^State included the ERS in the program rating

[a] significantly different from Level 1 programs
[b] significantly different from Level 2 programs
[c] significantly different from Level 3 programs
[d] significantly different from Level 4 programs

Cross-site comparisons of differences in ECERS-R scores among program levels revealed a number of interesting patterns (see Figure 5). Virginia had the strongest correlation between QRIS level and ECERS-R scores and weighted the ECERS-R the most when determining programs' ratings. Minnesota and Indiana had relatively similar patterns of association between ECERS-R and levels; however, Minnesota included the ECERS-R in the rating and Indiana did not. Maine demonstrated very little association between ECERS-R and levels. All states had a relatively large variation of scores within each level, indicating that ECERS-R was not the only driving force behind programs' ratings.

There are a number of issues to consider in understanding why the associations between ECERS-R and program ratings may be stronger in some states than in other states. For instance, the variation in ECERS-R scores was greater in Virginia than in the other states which likely is a factor in the stronger correlations noted in Virginia between ECERS-R and quality levels. In addition, the number of programs at each level as well as the total number of programs in the analysis may influence the results. For instance, if a state QRIS has a high percentage of programs at Level 1, a lower correlation with the ECERS-R may be detected.

Another issue to consider is that the reliability of observers and the protocol for conducting ECERS-R may vary across the states. For example, the states varied in the number of classrooms observed, the length of the observation and the time period in which the observation took place (e.g. beginning of the school year or at any point in the year), all of which have implications for validity.

The differences in associations may also be explained by the differing role of ECERS-R in determining the program rating. States that weight the ECERS-R scores in the program rating may expect to have stronger correlations between ECERS-R and levels compared to states that do not. This appears to be the case for Minnesota and Virginia which include the ECERS-R in the ratings. It is important to note that many of the indicators in the Indiana QRIS are similar to items on the ECERS-R which may also account for the relatively higher correlation in Indiana between ECERS-R and quality levels.

Summary of Approach 3

Overall, the cross –state findings indicate that ECERS-R scores can discriminate quality levels when the tool or similar indicators are used in the rating process (as seen in three of the four states). The findings in Maine, showing no correlation between quality levels and ECERS-R scores, provide a cautionary note that quality levels may not be distinct on the ECERS-R if indicators used in the QRIS do not align closely with the measure of environmental quality that is used in the validation process.

Perhaps the most important finding to observe across the four states is the relatively low ECERS-R scores, even among programs that received high ratings. These scores indicate that, at least on measures of global quality, programs are scoring well below scores that indicate "good" quality according to the authors of the scale.

Approach 4: Examine How Ratings are Associated with Children's Development

After validating the concepts, measures, and outputs used to determine program rating, a key task of QRIS validation is to examine whether the ratings relate to outcomes of interest. There are many potential outcomes of QRIS that may be important for stakeholders. For instance, states may have a goal of improving programs' ratings to increase the professionalization of early childhood educators thus leading to lower levels of turnover in early childhood programs. Additionally, states may be interested in supporting child care decision making and choices among parents selecting child care for their children.

Across many QRIS, a central goal is to improve child outcomes. In order to ensure that QRIS have the capability to support improvements in programs and in practices that can support children's developmental progress, it is essential to demonstrate that the quality ratings are related to child growth and development in meaningful ways. To account for different starting points on measures of key developmental domains such as language and literacy, early math and social-emotional development, it is best practice to track how ratings are correlated with measures of growth or progress rather than a set benchmark or score.

To date, researchers have only begun to examine whether QRIS program quality standards relate to measures of children's growth and development (Tout et al., 2009; Elicker and Thornburg, 2011; Zellman & Fiene, 2012). Validation of the ratings to children's progress is particularly challenging because it relies on the validity of all of the components that go into the ratings. This requires a clear understanding of how the system operates, and all of the components that drive the program ratings.

In addition, validation work linking ratings or rating components to children's progress relies on accurately accounting for selection bias. Selection bias can occur at a number of stages in the QRIS. For instance, most QRIS in the country are voluntary. Programs that volunteer may believe that they will have higher ratings than programs that choose not to volunteer, and this self-selection may explain relations between star ratings and outcomes. Additionally, the programs in the QRIS may not represent a random pool of programs in a state, thus limiting generalizability.

Additionally, there may be parental selection into certain programs. Parents with certain characteristics and experiences may select into certain types of programs. These characteristics are also likely related to children's skills. Thus, failing to take these important contextual factors into account when considering relations between program ratings and children's development may produce misleading results (Elicker and Thornburg, 2011; Zellman & Fiene, 2012).

Finally, reviews of research demonstrate that the quality of early care and education programs is associated with higher language, academic, and social skills and fewer behavior problems, but the effect sizes of these associations are small (Burchinal, Kainz & Cai, 2011). It is important to put validation results in the context of this research and to note the efforts in the field to strengthen the breadth, depth and content of the available quality measures (Zaslow, Martinez-Beck, Tout & Halle, 2011).

In this final section, we examine Indiana's, Minnesota's and Virginia's efforts to validate their QRIS ratings by examining linkages between QRIS ratings and measures of children's development. (Maine did not include measures of children's development in their validation studies.) These results are provided to illustrate the challenges of analyses relating ratings to children's development. We also emphasize that validation studies that include measures of children's progress are not attempting to identify causal linkages between QRIS participation and children's outcomes (Zellman & Fiene, 2012). The validation studies described below were not outcome evaluations aimed at assessing the effectiveness of the QRIS. Instead validation studies using this approach examine whether the QRIS ratings and quality components that comprise the ratings are related in expected ways to measures of children's development and differences in their patterns of growth. In essence, when reviewed in the context of information about the QRIS (findings from other validation activities, selection of programs and parents), these studies tell program developers and policymakers whether the tools used in the QRIS to measure and rate quality are working as intended.

We present the details of each study (excerpted from final project reports) and then identify lessons learned and implications for QRIS evaluators conducting similar studies in the future.

Indiana

Two children from each classroom or family child care home were randomly selected for a developmental assessment. The children were assessed by trained research assistants in a 20-45 minute time period during the quality assessment visit [ECERS-R; ITERS-R, or FCCERS, and Caregiver Interaction Scale (CIS)]. Analyses were conducted to investigate the relation between QRIS Path to Quality (PTQ) level and children's development as well as relations between observed quality (ECERS-R; ITERS-R, or FCCERS, and CIS) and children's development. For the full report, see Elicker et.al, 2011.

Infants and Toddlers. To examine infant-toddler development, 249 children ages 6 to 35 months were assessed statewide. The Brief Infant Toddler Social and Emotional Assessment was used to assess social competence and problem behavior. The Mullen Scales of Early Learning was used to assess cognitive development. Analyses were conducted to determine if children's developmental levels on these measures were higher at PTQ Level 4 vs. Level 1.

Findings indicated that infant-toddler development did not differ by type of care or PTQ level, even when parental education and household income were taken into account. Although these associations for infants/toddlers did not reach statistical significance, the average scores indicated a trend in the expected direction – infants and toddlers in Level 4 sites had higher average social competence, fewer reported behavioral problems, and scored higher on the cognitive assessments.

Several significant findings were noted in analyses linking observed quality and infant-toddler development. For example, when environmental quality (as measured by several ITERS-R scales) was higher, infants/toddlers displayed higher levels of social competence. When caregivers' interactions with children were rated as higher quality (according to the CIS), infants/toddlers' cognitive and language scores were higher. Children who scored higher on the Mullen Scales of Early Learning tended to have caregivers who were less permissive and less detached and displayed more sensitivity and positive interactions with children than the caregivers of children who scored lower on these cognitive measures.

Preschoolers. To examine preschool children's development, 308 children ages 36 to 60 months were assessed statewide. The Social Competence and Behavior Evaluation was used to assess social competence and problem behavior. The Woodcock Johnson III Applied Problems and Letter Word Identification Subtests were used to assess cognitive development. The Peabody Picture Vocabulary Test – 4 was used to measure receptive vocabulary (comprehension). Analyses were conducted to determine if children's developmental levels on these measures were higher at PTQ Level 4 vs. Level 1.

The analyses revealed significant findings. PTQ level was negatively related to children's anxiety/withdrawal behaviors, $r = -.12$, $p = .03$. Children with providers at higher PTQ levels displayed fewer anxiety/withdrawal behaviors than children with providers at lower PTQ levels. However, this finding disappeared when analyses controlled for family income.

Several significant findings also were noted in analyses linking observed quality and preschool children's development. For example, when providers were rated higher on the Language/Reasoning scale of the ECERS-R or FCCERS, children displayed greater language skills. When providers were rated higher on the Parents/Staff scale of the ECERS-R or FCCERS, children displayed less anxiety or aggression. When caregivers were observed to interact with children more positively and less punitively or permissively, children displayed higher levels of social competence and greater language ability.

Indiana summary. At an early stage of PTQ implementation, consistent, strong associations between PTQ quality level and young children's development and learning were not found. Considering all of the cognitive, language, and social-emotional child assessments, there were only small and inconsistent trends that children participating in programs at higher PTQ levels were doing better developmentally. These trends were not statistically significant, after parent education and household income were controlled. The study had several limitations including a small sample size and highly variable observed quality assessed within each PTQ level.

While PTQ levels did not predict children's development in this study, specific measures of child care quality did predict children's development and learning. For infants and toddlers, higher observed quality predicted higher levels of social competence; more positive, responsive interactions with caregivers predicted more advanced cognitive and language skills. For preschoolers, those who were in settings rated higher on ECERS-R Language/Reasoning sub-scale displayed higher language ability. Preschoolers in settings rated higher on the ECERS-R Parents/Staff sub-scale displayed fewer problem behaviors. When caregivers interacted more positively and responsively with preschoolers, the children tended to display more social competence and higher language abilities.

Minnesota

Participants. Children in Parent Aware rated programs were recruited into the evaluation in three cohorts: Fall 2008, fall 2009, and fall 2010. Parent Aware-rated programs assisted with the recruitment of eligible children (the majority were children completing the year prior to starting Kindergarten), with priority given to low-income children. Across the three cohorts, 701 children attending 138 Parent Aware-rated programs (including fully-rated and automatically-rated programs) participated in the evaluation. The sample of children was diverse with respect to race, income and language. Less than half of the children (42 percent) were White; nearly a quarter (24 percent) were African-American and the remaining 34 percent were from other racial and ethnic groups including Hispanic (8 percent), Hmong (5 percent) and Asian (4 percent). Eighty percent of the sample spoke English as their primary language. Other languages included Hmong, Spanish, Somali, and Karen. Sixty-one percent had a household income of less than $50,000 per year, and over one-third (37%) reported receiving some type of scholarship, subsidy, or other assistance for their early care and education expenses.

Method. Children attending programs rated in Minnesota's QRIS, Parent Aware, were assessed with measures of receptive and expressive language (Peabody Picture Vocabulary Test-4, Individual Growth and Development Indicators – Picture Naming), early literacy skills (Tests of Preschoolers Early Literacy (TOPEL), and early math and numeracy skills (Woodcock-Johnson – III: Applied Problems and Quantitative Concepts). Measures of social/emotional development and approaches to learning were completed by the children's teachers (Social Competence and Behavior Evaluation-short form and the Persistence scale from the Preschool Learning and Behavior Scales). Child assessments were collected in the fall and spring to assess children's gains across the pre-kindergarten school year (Tout, Starr, Isner, Cleveland, Albertson-Junkans, Soli, and Quinn (2011).

Analytic approach. Several approaches were used to understand the linkages between characteristics of early care and education programs and children's developmental progress. Analyses examined the relations between Parent Aware quality category scores, observational measures, and Parent Aware star rating with developmental gains. If Parent Aware ratings and observational measures successfully distinguished levels of quality that are linked to children's development, it was expected that children in programs with higher rating levels and scores on observational measures would make greater developmental gains. Multilevel modeling (accounting for children's nested in programs) was used to test the linkages between program quality and children's development, accounting for child and family characteristics such as household income and parental education. Models were run for all children and separately for low-income children to allow additional tests of the robustness of the findings.

Findings. Overall, the validation analyses did not show systematic evidence of strong relations between quality ratings, measures of program quality and children's developmental progress. Several significant relations were found, but in many cases they were not in the predicted direction, and they were not robust across different models and sub-samples. Findings for the star ratings were mixed. The star rating was positively related to children's receptive language (for low-income children only) and print knowledge scores. However, star rating was related in the unexpected direction to measures of social competence (negative relationship), anger/aggression (positive relationship) and attention/persistence (negative relationship). Analyses also examined linkages between children's development and scores on the quality categories that comprise the ratings as well as the observational measures used in the study (including the ECERS-R, subscales of the ECERS-Extension and the Classroom Assessment Scoring System – CLASS). Few significant findings were noted for the quality category scores in Parent Aware. Associations between observed quality scores and gains in early math skills were largely in the expected direction (with quality scores and ratings positively related to children's gains). Results were more inconsistent for language/literacy, social/emotional outcomes, and approaches to learning. Looking across the findings of the validation analyses, there were nearly as many unpredicted associations (with quality scores and ratings negatively associated with children's positive gains) as expected associations between program characteristics and children's development.

Minnesota summary. Overall, there was no systematic pattern of linkages between the star ratings, the quality measures and children's development in the validation analyses for Minnesota. Yet, the data about children's progress in Minnesota have been an important component of the research activities on the QRIS. For example, QRIS stakeholders in Minnesota have used the descriptive data on gains in preschool children's development – with stronger gains noted for low-income children – to identify overall patterns of strengths and concerns. For example, while children make relatively strong gains in aspects of language and literacy development, gains in early math and in social-emotional development are less strong (with teacher reports of anger/aggression *increasing* over the preschool year). These descriptive data can be used to identify focal areas for providing supports such as training and technical assistance to early childhood programs.

Virginia

The Center for Advanced Study of Teaching and Learning (CASTL) at the University of Virginia conducted a study on pre-kindergarten programs in the Virginia's QRIS, the Virginia Star Quality Initiative (VSQI; for a full report, see Sabol & Pianta, 2012). The primary purpose of the study was to investigate the extent to which the ratings in the Virginia Star Quality Initiative relate to growth across pre-kindergarten and kindergarten. Due to methodological concerns for the validation question of this study, and to better understand selection into the VSQI, the study also investigated the characteristics of communities, programs, and children in the VSQI.

Method. Analyses drew upon a database constructed from the following datasets: (1) quality standard scores and star ratings on the VSQI from 2007-2009; (2) center addresses, licensure information, and demographic information from the Virginia Department of

Social Services; (3) block group census data from the 2000 U.S. Census; and (4) pre-kindergarten and kindergarten performance on the Phonological Awareness Literacy Screening (PALS), and (5) center and child characteristics from the Virginia Department of Education and the University of Virginia PALS Office, from 2007-2010.

Participants. The 71 VSQI-rated pre-kindergarten programs were located predominately in urban neighborhoods (70%), with 25% of the residents having a Bachelor's degree or more, and relatively high percentages of White residents (67%), followed by African American (24%) and Hispanic (4%) residents. The average income of neighborhood residents was $54,910 with 11% living below the poverty level.

Among the 71 VSQI rated programs, the average rating score on the VSQI was 111.59 (S.D. 19.02). There were 8 2-star programs (11%), 38 3-star programs (54%), and 25 4-star programs (35%), with no 1-star or 5-star programs. The programs had approximately 41.68 children enrolled with a fair amount of variation (S.D. 63.11). Programs were composed of almost equal averages of white (38%) and African American children (33%) followed by Hispanic children (15%). Children in the sample were on average 57 months of age. Fifty-four percent of the children were non-white, 50% were boys, 7% had disabilities at prekindergarten entry, and 8% had limited English Proficiency. Twelve percent of the children received Title I funding.

Teachers assessed children's pre-literacy skills using the Phonological Awareness Literacy Screening (PALS) across four time points: fall and spring of prekindergarten, and the fall and spring of kindergarten. PALS Pre-K and PALS-K is a criterion-referenced assessment and demonstrates high test–retest reliability (.99), and high internal consistency (.99; Huang, Invernizzi, & Drake, 2012). Two factors were used, Alphabet Knowledge (2 items) and Phonological Awareness (2 items), to allow for longitudinal analysis (Townsend et al., 2010).

Analytic approach. To test the relation between VSQI ratings and children's development, a multilevel model of lagged performance scores, as well as a multilevel growth model, with child- and center-level controls and community fixed-effects was employed. A two-level model was employed to account for the hierarchical

data structure where children were nested in programs. This approach is consistent with numerous rigorous examinations of relations between program quality and children's functioning (Barnett, Lamy, & Jung, 2005; Gormley, Gayer, Phillips, & Dawson, 2005; Mashburn et al., 2008).

All analyses controlled for children's pre-kindergarten performance. A second set of analyses controlled for community-, center-, and child-level controls in addition to children's pre-kindergarten performance. Additionally, the study only explored the relations between program ratings and children's development in the first year the center was rated, reducing the potential that parents are differentially selecting programs based on the center rating. Although these analytic techniques may reduce selection, they by no means eliminate bias. Thus, the results from the VSQI validation study present a non-causal, descriptive exploration of the relation between ratings and child development.

Findings - Differences in star rating by center and neighborhood characteristics. A one-way ANOVA was employed to test for mean differences of neighborhood characteristics, center characteristics, and child characteristics between the three star ratings (1-star and 5-star programs were omitted because there were so few). There were a number of differences between 2-star, 3-star and 4-star programs with regard to center/ neighborhood characteristics among pre-kindergarten programs (see full report for table). Two-star programs tended to be in areas with more African American residents and more single mother households compared to 3-star and 4-star programs. Four-star programs, and to a lesser extent 3-star programs, were located in areas with the most Hispanic residents, and more rural areas.

Findings - Child characteristics by star rating. Child characteristics also differed by star rating. In terms of racial demographics; there were higher proportions of African American children in 2-star programs compared to 3-star and 4-star programs and higher proportions of Hispanic and Limited English Proficiency (LEP) children in 4-star programs compared to 2-star and 3-star programs. Children were also slightly older in 4-star programs. Most surprisingly, children in the 2-star programs had higher performance on all measures of literacy performance at the start of pre-kindergarten compared to 3-star and 4-star programs. In aggregate, the differences in center and child characteristics and performance demonstrate that children enrolled in 2-star, 3-star, and 4-star programs did not have the same characteristics.

Findings - Predicting literacy growth in early childhood. Researchers then examined children's growth rate between the fall of pre-kindergarten to spring pre-kindergarten, spring of pre-kindergarten to the fall of kindergarten, and fall of kindergarten to spring kindergarten. Within each of these time points, they examined whether the growth rate differed as a function of star rating. This type of analysis allowed an examination of questions such as: did children in 4-star programs have a sharper increase in their literacy skills growth compared to children in 2-star programs in the pre-kindergarten year?

Children in 3-star and 4-star programs had significantly lower Alphabet Knowledge and Phonological Awareness skills at the start of the pre-kindergarten year, starting at least half a standard deviation behind children in 2-star programs (see Table 8). Although children in 3-star and 4-star programs started out with lower performance, children in these higher rated programs were characterized by sharper growth in the pre-kindergarten year. Children in 3-star programs gained around three-fourths of a standard deviation in Alphabet Knowledge (letter sounds and names) and Phonological Awareness, over and above children in 2-star programs in the pre-kindergarten academic year. Additionally, children in 4-star programs grew one third of a standard deviation more in Alphabet Knowledge compared to children in 2-star programs. Children in 4-star programs grew slightly more in Phonological Awareness in pre-kindergarten compared to children in 2-star programs, yet the relation was not significant. There was no significant difference in growth in pre-kindergarten between 3-star and 4-star programs.

Table 8. Estimates of Association between Star Rating and Growth in Alphabet Knowledge and Phonological Awareness

	Alphabet Knowledge	Phonological Awareness
Fall PK to Spring PK		
3-star	0.43**	0.37*
4 -star	0.40**	0.20
Spring PK to Fall K		
3-star	-.12*	0.04
4 -star	-0.18*	-0.02
Fall K to Spring K		
3-star	-0.02	-0.07
4 -star	0.06	-0.12

Note: Effect sizes are presented. All effect sizes are in comparison to 2-star programs. * p<.05 ** p<.01 ***p<.001

Surprisingly, children in 3-star and 4-star programs had a somewhat steeper decline in alphabet knowledge skills over the summer between pre-kindergarten and kindergarten compared to children in 2-star programs. There was no difference in Phonological Awareness growth rates over the summer among the programs. Additionally, results suggested that there was no significant difference in growth during the kindergarten year between 2-star, 3-star and 4-star programs. Importantly, it is difficult to parse out whether the non-significant finding in kindergarten are because of fading relations to quality, or because of ceiling effects from the measure, which may not allow for substantial growth in kindergarten.

Virginia summary. The validation findings suggest that VSQI ratings are a modest, but reliable, predictor of growth in pre-literacy skills during prekindergarten. Although the methods may not have entirely eliminated the threats of selection, the results of this study suggest that the criteria used to distinguish 3-star or 4-star program may have positive relations to children's pre-literacy trajectories, at least in the short run; 3-star and 4-star programs are related to slightly steeper decline in alphabet knowledge over the summer, and by kindergarten, there are no differences in growth.

In addition, the descriptive component of this study had important implications for linking programs' rating to children's development. For instance, descriptive findings suggested that selection into the VSQI varied by neighborhood and program characteristics, and that programs received differential ratings based on these neighborhood characteristics. Centers in more disadvantaged neighborhoods were more likely to participate in the VSQI than centers in more advantaged neighborhoods. This limited the generalizability of the validation work to programs that serve mainly disadvantaged children.

Descriptive findings also indicated somewhat stark differences in the racial composition between 2-star and higher rated programs, where 3-star and 4-star programs had higher percentages of children who were Hispanic and English language learners, and 2-star programs had more African American children. Although researchers employed a number of controls for this differential selection into quality levels, as well as employed a number of robustness checks (e.g. omitting English language learners from the model), it remains a possibility that the differences in the characteristics of children in each of the levels may in part explain the differences in growth in language and literacy skills.

This Virginia study only begins to unpack the relation between VSQI ratings and children's development. There are notable limitations to this study, including the non-causal approach, the lack of multiple dimensions of children's development, the lack of control for dosage of quality, and the exclusive focus on pre-kindergarten programs. More work is needed with a broader sample to understand the effects of selection into the VSQI and the links between program ratings and children's development.

Summary of Approach 4

The three state studies presented in this section provide useful examples of an approach to QRIS validation that includes examination of children's development. Taken together, the studies demonstrate both the opportunities and the limitations of this approach to validation. The opportunities are evident in the rich data that are available to QRIS stakeholders about the functioning of the QRIS and the developmental status of children participating in QRIS-rated programs. In Indiana, for example, the research teams included infants and toddlers in the data collection, a subgroup of children that was not studied in the other states or in previous validation studies in Colorado or Missouri. In Virginia, the researchers conducted selection analyses so that patterns of program enrollment and child participation could be understood. In Minnesota, researchers presented descriptive data on children's development to inform stakeholders about the developmental gains (or lack of) children made overall across QRIS-rated programs.

The studies also demonstrate the limitations that must be considered when conducting validation studies that include measures of children's development. All of the studies were limited by small sample sizes, unequal distribution of programs across the rating levels and low variation in the observed quality scores across levels (which could limit the possibility of finding significant linkages with children's developmental progress). Additionally, collection of data on developmental progress requires a significant investment of project resources, not only for deploying research assistants to collect direct assessments but also for the time needed to create IRB protocols and to recruit programs and families into the study. The relative value of a child development analysis should be weighed against these other significant costs and limitations.

Overall, each state effort produced relevant information that was used by QRIS administrators and stakeholders to examine whether the ratings related to measures of children's developmental progress in expected ways. The findings indicate a need for new evidence to support the development and refinement of rating processes and QRIS quality levels that are linked to greater distinctions in children's developmental progress. For example, researchers working with states on validation efforts might recommend piloting new quality measures along with the existing set of measures in a QRIS so that comparisons can be made between different measurement strategies and their linkages with constructs of interest. Piloting can help inform these decisions before measurement changes are brought to scale. Also, as noted in the state work presented above, researchers can conduct analyses to learn more about the components that make up the rating. Measures of some quality components and indicators may be working more effectively than others to distinguish children's developmental progress. Alternatively, some measures used in the rating process may have stronger linkages with measures of workforce development (teacher qualifications or retention), parent engagement or other construct than with measures of children's progress. Data from validation studies can be used to understand how the tools in the rating system are functioning and whether and how adjustments should be made to improve the measurement and rating processes.

Conclusion

The purpose of this Brief is to serve as a companion document to a descriptive overview of four related approaches to QRIS validation produced through the Quality Initiatives Research and Evaluation Consortium (INQUIRE) by Zellman and Fiene (2012). The results of validation studies conducted in four states – Indiana, Maine, Minnesota and Virginia – are presented to provide examples of the efforts that QRIS might invest in to learn more about whether the program measures and rating are functioning to produce distinct markers of quality.

The picture that emerges from the synthesis of findings across the four states and across the validation approaches is mixed. For instance, the results of efforts to validate the quality standards and indicators in QRIS generally have been successful. Efforts to review how well measures are functioning, however, reveal concerns about limited variation on some measures and QRIS structures that are producing skewed distribution of programs. There are some indications that QRIS levels are distinct with respect to measures of observed quality, but only in the QRIS that used the measures as part of the system. In Maine, where the ECERS-R was used only for validation purposes, no linkages were found between observed quality and QRIS levels. Finally, validation studies that included measures of children's developmental progress indicate only limited support for linkages between these measures of children's growth, QRIS ratings and program quality elements. The findings suggest that further work is needed to strengthen the ability of QRIS ratings to serve as meaningful markers of program quality.

To return to a key theme from the beginning of this Brief, **the information gained from validation efforts can serve as a critical tool for guiding initial design of QRIS, redesign efforts and continuous quality improvement.** Zellman and Fiene (2012) emphasize that validation studies do not produce "yes" or "no" answers about QRIS but provide data that can support QRIS in a process of refining and improving. As such, validation efforts must be timed appropriately and aligned with a clear decision-making framework for how the findings will be used. In the four states highlighted in this Brief, researchers partnered with state agency leaders and other QRIS stakeholders to assist in developing a validation plan that could support QRIS development as well as a process for reviewing and interpreting findings so that the results could be applied appropriately. As states continue implementation of QRIS, administrators and stakeholders are encouraged to engage in validation efforts that can inform their systems and move progressively toward the provision of effective services.

References

Barnett, W. S., Lamy, C., & Jung, K. (2005). *The effects of state prekindergarten programs on young children's school readiness in five states.* New Brunswick, NJ: National Institute for Early Education Research.

Bryant, D. M. (2001). *Validating North Carolina's 5-Star child care licensing system*. Chapel Hill, NC: Frank Porter Graham Child Development Center.

Bryant, D., Maxwell, K., Taylor, K., Poe, M., Peisner-Feinberg, E., & Bernier, K. (2003). *Smart Start and preschool child care quality in North Carolina: Change over time and relation to children's readiness*. Chapel Hill, NC: FPG Child Development Institute.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow (Ed.) *Quality measurement in early childhood settings.* Baltimore, MD: Brooks Publishing.

Elicker, J.G., Langill, C.C., Ruprecht, K.M., Lewsader, J., and Anderson, T. (2011). *Evaluation of "Paths to QUALITY", Indiana's Child Care Quality and Rating and Improvement System; Final Report.* Dept. of Human Development and Family Studies, Center for Families, Purdue University.

Elicker, J., Langill, C.C., Ruprecht, K. and Kwon, K-A. (2007). *Paths to QUALITY – Child Care Quality Rating System for Indiana What is its Scientific Basis?* Child Development and Family Studies, Purdue University. Retrieved from: http://www.cfs.purdue.edu/cff/documents/project_reports/07_paths_to_quality.pdf

Elicker , J. , Ruprecht , K. M. , Langill , C. , Lewsader , J. , Anderson , T. , & Brizzi , M. (2013). Indiana Paths to QUALITY™: Collaborative evaluation of a new child care quality rating and improvement system. *Early Education & Development*, *24*(1), 42–62.

Elicker, J., & Thornburg, K. (2011). *Evaluation of Quality Rating and Improvement Systems in Early Childhood Programs and School Age Care: Measuring Children's Development.* Research-to-Policy, Research-to-Practice Brief OPRE 2011-11c. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Gormley, W. T., Gayer, T., Phillips, D. & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology, 41,* 872-884.

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

Huang, F. L., Invernizzi, M. A., & Drake, E. A. (2012). The differential effects of preschool: Evidence from Virginia. *Early Childhood Research Quarterly*, *27*(1), 33–45. doi:10.1016/j.ecresq.2011.03.006

Indiana, Paths to QUALITY (2012) Retrieved from: http://www.childcareindiana.org/childcareindiana/ptq.cfm

Lahti, M., Cobo-Lewis, A., Dean, A., Rawlings, S., Sawyer, E., and Zollitsch, B. (2011) *Maine's Quality for ME – Child Care Quality Rating and Improvement System (QRIS): Final Evaluation Report.* Edmund S. Muskie School of Public Service, University of Southern Maine. Submitted to ME, Dept. of Health and Human Services.

Maine, Quality for ME QRIS (2012) Retrieved from: http://www.maine.gov/dhhs/ocfs/ec/occhs/qualityforme.htm.

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

Minnesota Parent Aware Ratings (QRIS) (2012) Retrieved from: http://www.parentawareratings.org/

Norris, D. J., Dunn, L., & Eckert, L. (2003). *"Reaching for the stars": Center validation study final report.* Norman, OK: Early Childhood Collaborative of Oklahoma.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment Scoring System.* Baltimore, MD: Paul H. Brookes Publishing Co., Inc.

Sabol, T.J. & Pianta, R.C. (2012). *Improving child care quality: A validation study of the Virginia Star Quality Initiative.* Charlottesville, VA: Center for the Advanced Study of Teaching and Learning.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *Compendium of Quality Rating Systems and Evaluations.* Washington, D.C.: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of quality rating and improvement systems.* OPRE Issue Brief #3. Washington DC: Child Trends.

Tout, K., Starr, R., Isner, T., Cleveland, J., Albertson-Junkans, L., Soli, M., & Quinn, K. (2011). *Evaluation of Parent Aware: Minnesota's Quality Rating and Improvement System pilot: Final evaluation report*. Saint Paul, MN: Minnesota Early Learning Foundation

Townsend, M., & Konold, T. R. (2010). Measuring early literacy skills: A latent variable investigation of the Phonological Awareness Literacy Screening for Preschool. *Journal of Psychoeducational Assessment, 28*, 115-128.

Trochim, W. (2012). *An Introduction to Concept Mapping for Planning and Evaluation.* Retrieved from: http://www.socialresearchmethods.net/research/epp1/epp1.htm.

Virginia Start Quality Initiative (2012) Retrieved from: http://www.smartbeginnings.org/Home/StarQualityInitiative/AboutStarQuality.aspx

Zaslow, M., Martinez-Beck, I., Tout, K., & Halle, T. (2011). *Quality measurement in early childhood.* Paul H. Brookes Publishing Co. Baltimore, MD.

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., Le, V., & Setodji, C.M. (2008). *Assessing the validity of the Qualistar Early Learning Quality Rating and Improvement System as a tool for improving child care quality.* RAND. Santa Monica, CA.

APPENDIX

Background on QRIS and Evaluation in each State

Indiana

Indiana's rating system, Paths to QUALITY (PTQ), was launched in 2008 and offers four rating levels for licensed child-care centers; licensed family home providers; and unlicensed, registered child-care ministries. Level 1 means that the provider meets the requirements for state licensing, which includes proper adult and child ratios, program development, and adherence to fire and safety guidelines. Levels 2 and 3 focus on improvements in facilities' physical environment and educational opportunities, respectively. Level 4, the highest level, includes national accreditation criteria; see http://www.childcareindiana.org/childcareindiana/ptq.cfm for more information about Indiana's QRIS. The child care provider must pass every PTQ standard on the checklist to be awarded the next level. The four levels address many quality criteria, but the main emphasis at each level is:

- Level One: Health and safety needs of children met.
- Level Two: Environment supports children's learning.
- Level Three: Planned curriculum guides child development and school readiness.
- Level Four: National accreditation is achieved

The evaluation study, with data collection completed between July 2008 and September 2011 included all eleven Child Care Resource and Referral Service Delivery Areas (SDAs) in Indiana. The overall goals of the evaluation research were to validate the QRIS and describe the experiences of child care providers, parents, and children with the QRIS as it was implemented. During the course of the research, Purdue provided program leaders with periodic reports that described aspects of PTQ implementation in each SDA region, so that they could better monitor the acceptance and influence of PTQ and make program adjustments as needed. The final evaluation sample comprised a total of 276 child care providers: 95 licensed child care centers (including 135 classrooms assessed); 169 licensed family child care homes; and 12 unlicensed registered child care ministries (including 14 classrooms assessed). Within these selected child care providers, the research team interviewed or assessed 270 child care teachers/providers, and 557 children and their parents.

The evaluation questions addressed by the Purdue research team were:

- When providers attain higher PTQ levels, does this result in higher quality care for children?
- Are child care providers entering the PTQ system? What are the incentives and the challenges for providers? Are providers using available training/technical assistance (T/TA) resources? Are providers advancing to higher PTQ levels?
- Are parents aware of PTQ? Will PTQ affect their parents' child care decisions?
- Are children and families at all education and income levels gaining access to child care at the highest PTQ levels? Are children in higher PTQ levels developing more optimally than children in lower PTQ levels?

The evaluation plan included activities using all four validation approaches outlined in Table 1.

Maine

Maine has a building block QRIS with program standards defined at four tiers or Steps. If a program is in operation for more than one year, and has no significant licensing violations, then it is eligible for enrollment. Programs that are serving families supported by government subsidy, Child Care Development Fund supports, are required to enroll into the state QRIS. For all other programs, enrollment is voluntary. Standards at Step One consist of; being licensed for one year with no substantiated serious licensing violations, and all members of program must be members of Maine Roads to Quality Registry. Standards at Steps Two through Four are designed by type of program; School Age, Head Start, Center Based and Family Child Care. As a building block type of QRIS, programs only advance after meeting all of the standards at each level. For each program type and Step Level, standards are defined in these component areas:

- Compliance with licensing / Membership in Registry
- Training / education levels of staff and director/owner
- Focus on curriculum and completion of training: infant/toddler and early learning state guidelines
- Program Structural Components (regular staff meetings, self-evaluation, policies/procedures for staff, etc.)
- Family Involvement
- Child Level Assessments for Children's Development and Curriculum Planning
- Step Four level programs are required to attain national certification

Providers complete an online application form that allows them to self-report as to whether or not they meet each standard. Providers are required to have on site a portfolio that provides documentation for how each standard is met. From the online application, a form is generated immediately that provides an initial Step Level rating as well as information for program lower than a Step Four identifying standards needed to be met to move to the next Step. The online system is built from a linked administrative data base that connects state licensing data and state Registry data. This allows for data verification and enhances data quality for items associated with provider training and education, and some program components. The state agency program specialist visits programs each year on a random basis to check the program portfolio for evidence on the indicators for each standard. For more information about Maine's QRIS see; http://www.maine.gov/dhhs/ocfs/ec/occhs/qualityforme.htm.

The evaluation and monitoring of Maine's QRIS is designed to monitor the enrollment patterns and movement of the QRIS. The evaluation is designed to answer the following questions:

- What are the characteristics of programs enrolled in the QRIS?
- What is the quality of the program learning environment as measured by the Environmental Rating Scales (ERS)?
- What are the differences in program characteristics at each Step Level?
- What are the differences in program quality comparing similar program types between Step Levels?
- What are parent perceptions of program services and quality?
- What are the characteristics and perspectives on learning of center-based program teachers / staff and family child care home providers?

Maine's QRIS evaluation plan includes activities using the first three validation study approaches outlined in Table 1; Maine's QRIS evaluation and validation studies do not include assessment of children's development. The evaluation was conducted through random selection of programs by type and step level over a three year period, 2008-2011. Evaluation and validation activities are continuing through November 2012.

Data is collected at the levels of program, staff and parents. Program level data includes information on teacher education and training, licensing status and information from their QRIS application. In addition, classroom level observations are done using the Environmental Rating Scales for each program type. The results of these global rating scales are used in the validation study; the scores are not part of the QRIS quality indicators. Confidential parent and staff surveys are conducted at each site. The final evaluation sample comprised of 255 childcare programs, 307 individual classrooms / family child care homes, 1,478 parents who completed surveys, and 424 staff who completed surveys. For more information on the evaluation and access to results (Lahti et al, 2011), see http://muskie.usm.maine.edu/cutler/cyf/mainechildcare/.

Minnesota

Minnesota's QRIS, Parent Aware, is a voluntary system for licensed family child care programs, child care centers, Head Start, and School Readiness programs. Parent Aware was piloted in four Minnesota communities including the cities of Minneapolis and St. Paul, as well as select suburban and rural communities from 2007 – 2011. During the pilot, for fully-rated programs, ratings were assigned in a points system based on four categories of standards: Family Partnerships, Teaching Materials and Strategies, Tracking Learning, and Teacher Training and Education. Accredited programs, Head Start, and School Readiness programs were automatically awarded four stars.

The primary purpose of Parent Aware is to support parents by providing information about the quality of early care and education programs. Parent Aware uses ratings to recognize quality and promotes quality improvement using a variety of resources. Together, these strategies aimed at parents and early care and education programs target an ultimate goal of improving children's school readiness. For more information see; http://www.parentawareratings.org/.

In the pilot, Parent Aware calculated ratings for fully-rated programs based on points earned in four categories: Family Partnerships, Teaching Materials and Strategies, Tracking Learning, and Teacher Training and Education. Family Partnership points could be earned by collecting and using feedback from parents, implementing strategies for communicating with families, and other opportunities for communication with parents. In Teaching Materials and Strategies, programs earned points by using a research-based curriculum (with staff trained on the curriculum) and by earning points on the Environment Rating Scales and Classroom Assessment Scoring System observation tools. Tracking Learning points were earned by using a research-based assessment tool, having staff trained on the tool, sharing assessment information with families, and using assessment information to guide instruction. Finally, Teacher Training and Education points were based on teachers' education as indicated by their step on the Career Lattice. For each category, 10 points were possible. Program who achieved a total of 0-11.5 points were rated 1 star, 12-23.5 points earned 2-stars, 24-31.5 points earned 3-stars, and 32-40 points was equal to 4-stars. In addition, the use of an approved curriculum in preschool classrooms was mandatory to earn 3-stars and programs must have received a score of 3 or higher for each CLASS subscale in order to achieve 4-stars.

Information on the indicators was collected through program observation, data from the Minnesota Professional Development System Registry, and documentation provided by the programs. Parent Aware raters at the Department of Human Services reviewed the documentation and issued a rating.

Accredited programs, School Readiness programs, and Head Start programs did not go through the full rating process. Rather, they were automatically rated 4-stars in Parent Aware.

In 2007, the Minnesota Early Learning Foundation (MELF) contracted with Child Trends to conduct a four-year evaluation of Parent Aware. The evaluation used multiple data sources to examine implementation and outcomes of Parent Aware. Specifically, the evaluation examined: (1) implementation of Parent Aware (participation, quality improvement supports, parent and provider perceptions, marketing, the rating process), (2) the effectiveness of the rating tool in distinguishing levels of quality, (3) quality improvement, and (4) the relation between quality and measures of children's developmental progress.

Data were collected at the level of the community, early care and education programs, and families and children. Data collection methods included interviews, surveys, program observations, administrative data, and child assessments. For the final evaluation report on Parent Aware see Tout, Starr, Isner, Cleveland, Albertson-Junkans, Soli, and Quinn (2011).

The evaluation plan for Minnesota's Parent Aware addressed three of the four validation approaches. Efforts to examine the concepts and quality standards in Parent Aware were conducted directly by the Parent Aware program prior to design.

Virginia

The main goal of Virginia's QRIS, the Virginia Star Quality Initiative (VSQI), is to provide a consistent way to distinguish the level of quality in early education programs within the Commonwealth of Virginia. A non-for-profit, the Virginia Early Childhood Foundation, and a state agency, the Virginia Department of Social Service's Office of Early Childhood Development, jointly manage the VSQI. Differing from other states, the program is currently funded by a complex combination of federal grants (the Child Care Development Fund), private foundation grants, and state and local funds. For more information, see; http://www.smartbeginnings.org/Home/StarQualityInitiative/AboutStarQuality.aspx.

The VSQI was first implemented as a pilot program in the 2007-2008 school year. In the pilot program, the VSQI solely focused on center-based programs that served three- and four-year-old children. Participating programs received information on their performance, but the ratings in the pilot year were never publicized. The VSQI officially began in 2008-2009. Differing from other QRIS which typically include home-based programs, only licensed child day care programs, Head Start/Early Head Start, prekindergarten, early childhood programs, licensed-exempt faith-based providers, and military settings are eligible to participate in the VSQI (Tout et al., 2010).

Programs in the VSQI are assessed on a five star scale based on performance on four quality standards: (1) staff education and qualifications; (2) teacher-child interactions; (3) structure (i.e. staff-to-child ratio); and (4) environment. The quality standards are derived from direct observations and program documentation. Star Quality Raters directly assess Standard 2, the quality of teacher-child interactions, with the CLASS (Pianta, La Paro, & Hamre, 2008) and Standard 4, environment, with the Early Childhood Environmental Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998) Star Quality Raters are extensively trained and are tested for inter-rater reliability once every seven visits. In the pilot year, raters observed one out of every three classrooms. In subsequent years (2008-2011), one toddler classroom, one three year-old classroom and one four year-old classroom was observed when available. Additionally, programs send in documentation on their staff qualifications (Standard 1) and staff-to-child ratio and group size (Standard 3).

The program levels in the VSQI are derived by a complex weighting and aggregation scheme based on performance on the four performance standards. The raw data are converted to points for each standard, which in turn are converted to star ratings. Each standard is out of 40 possible points except for Standard 2, teacher-child interactions. Because the stakeholders were particularly concerned with the quality interactions, the VSQI weights the interaction standard area 1.5 times more than the other three standards (60 points). As such, the total score is out of 170 possible points. The total scores are converted to a star based on the following cut-scores: Star 1: 34-50 points; Star 2: 51-84; Star 3: 85-118, Star 4: 119-152, and Star 5: 153-170.

In 2008, the Center for Advanced Study of Teaching and Learning (CASTL) at the University of Virginia partnered with the non-for-profit organization, the Virginia Early Childhood Foundation, and the Virginia Office of Early Childhood Development to study the VSQI. The primary purpose of this partnership was to validate the VSQI and investigate the extent to which the rating structure in the VSQI relates to children's concurrent functioning, as well as growth across prekindergarten and kindergarten.

The validation work addressed three main research areas: (1) the characteristics of the VSQI rating system; (2) the characteristics of the communities, neighborhoods, centers and children in the VSQI; and (3) the relation between star ratings in the VSQI and children's literacy skills. The data for validation study came from a variety of different sources, including the Virginia Early Childhood Foundation, the Virginia Department of Social Services, the 2000 U.S. Census, and the Virginia Department of Education.

At the broadest level, the study descriptively explored the characteristics of the rating system among 464 programs rated in the VSQI from 2007-2011. Researchers then narrowed in on programs only rated from 2007-2009 (n=255) that they had access to a richer set of information and explore the community, neighborhoods, and classroom characteristics. Lastly, researchers examined the center and classroom characteristics of 71 pre-kindergarten programs in the VSQI, and then examined the relation between the star ratings and children's literacy performance in early childhood. After the validation study was complete, researchers prepared a final report for the VSQI stakeholders. VSQI stakeholders discussed the findings with researchers and asked recommendations in effort continuing to refine and improve the system.

Validation efforts in Virginia included a focus assessing the outputs of the rating process and examining how ratings are associated with children's development.