ARTICLE IN PRESS

Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs

Michel Lahti [a,*], James Elicker [b], Gail Zellman [c], Richard Fiene [a,b,c,1]

[a] *University of Southern Maine, Muskie School of Public Service, PO Box 9300, Portland, ME 04104-9300, United States*
[b] *Purdue University, United States*
[c] *RAND Corporation, United States*

ARTICLE INFO

ABSTRACT

In recent years, child care quality rating and improvement systems (QRISs) have become an increasingly popular policy tool to improve quality in early childhood education and care (ECEC) settings and have been adopted in many localities and states. The QRIS proposition is that with higher-quality child care settings, it is more likely that children who attend those high-quality programs will benefit in terms of outcomes like school readiness. However, in order to demonstrate this linkage, QRIS standards and ratings must function as intended, i.e. be valid. This paper presents a framework for validating child care quality improvement standards and processes, along with examples from recent QRIS validation studies in two states. The state examples provide useful data about the strengths and limitations of these validation approaches. We discuss the implications of applying these approaches and provide recommendations to researchers, policy-makers, and program leaders who implement QRIS validation studies.

© 2014 Elsevier Inc. All rights reserved.

Introduction

In recent years, child care quality rating and improvement systems (QRISs) have become an increasingly popular policy tool to improve quality in early childhood education and care (ECEC) settings and have been adopted in many localities and states. The QRIS National Learning Network reports that 40 statewide QRISs have launched or piloted, including the District of Columbia (QRIS National Learning Network, 2014). The immediate goal of a QRIS is to raise the quality of care in early learning settings. Existing research suggests that care in higher-quality settings will improve child functioning, including school readiness (Burchinal et al., 2009; Burger, 2010; Howes et al., 2008), especially for children from lower-income families. QRIS logic models that guide these large-scale interventions focus on improving various dimensions of ECEC quality, with the ultimate goal of improving system outcomes, namely; child care program quality, training and technical assistance for child care providers, information and support for families,

and, therefore, improvements to children's cognitive, language, social, emotional, and physical development.

The perceived need for QRIS has grown out of documented gaps in quality in existing ECEC programs, especially those serving children from lower-income families (Fuller, Loeb, Kagan, & Carrol, 2004; NICHD ECCRN, 2000) and the inability of the current ECEC system to promote uniformly high quality (Cochran, 2007). QRISs produce program-level quality ratings based on multi-component assessments designed to make ECEC quality transparent and easily understood to parents and other stakeholders. Most also include feedback, technical assistance, and incentives to both motivate and support providers' efforts toward quality improvement (Tout et al., 2010). To make program quality transparent, QRISs typically rely on a multi-tiered rating system with one to five levels of program quality. Therefore, it is important that these ratings show evidence of validity, so that higher-quality programs are rated higher, and lower-quality programs are rated lower.

Recent research has documented the importance of both specificity and thresholds when testing hypotheses about child care quality impacts on children's developmental outcomes (Burchinal, Peisner-Feinberg, Bryant, & Clifford, 2000; Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Howes, Whitebook, & Phillips, 1992; NICHD ECCRN, 2000, 2002). However, common global measures of classroom quality such as the Early Childhood Environment

* Corresponding author. Tel.: +520-326-5154; fax: +520-326-5155.
  E-mail addresses: mlahti@usm.maine.edu, Michel.lahti@lecroymilligan.com
(M. Lahti).
[1] Professor Emeritus at Pennsylvania State University.

Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 2005) are not always significantly associated with specific child outcomes (Burchinal, Kainz, & Cai, 2011). This may be because these global quality scales do not focus enough on the particular child care quality processes most likely to bring about improved child outcomes (specificity) or they do not provide guidance for the level of quality required to produce improved child outcomes (thresholds). As states implement QRISs, they are using observational measures such as the ECERS-R, and they may also combine other quality measures such as the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) or locally specified quality indicators. Because QRIS quality standards are often complex, including many components and measures at several quality levels, and because they vary from state to state, it is especially important for states to carefully validate their quality rating systems and match measures specifically to the stated outcome goals of the QRIS. For example, if a particular QRIS places more emphasis on the health aspects of children's development, then the ECERS-R and CLASS would not be appropriate tools; but a tool measuring child care health indicators, such as the National Health and Safety Tool being developed by the California Child Care Health Program (Alkon, 2013) would be more appropriate.

Validity data can also enable researchers to test conclusions about whether the quality indicators embedded in QRIS standards lead to adequate quality assessment and whether the methods used to assign quality ratings are working as intended (Cizek, 2007). This paper defines operationally the concept of QRIS validity, presents four general approaches to assessing validity in the context of large-scale QRISs, and critically examines the efforts of two states, Maine and Indiana, to assess the validity of recently implemented QRISs using these approaches.

Validation of a QRIS is a developmental and multi-step process that assesses the degree to which design decisions about program quality standards and measurement strategies are resulting in accurate and meaningful quality ratings. Validation of a QRIS provides designers, administrators, and stakeholders with crucial data about how well the system is functioning. A carefully designed plan for ongoing QRIS validation creates confidence in the system and a climate that supports continuous quality improvement at both the child care provider and system levels (Zellman & Fiene, 2012).

To date, QRIS validation research efforts have been limited, for a number of reasons. First, validation is complex and involves a range of activities, which should include validating standards, measures, and rating protocols. Second, there has been little information available in the field that clarifies the importance and purpose of QRIS validation or identifies recommended strategies. Third, child care quality advocates and policy makers have been extremely busy designing and implementing these statewide systems, often with limited resources. Given these constraints, validation may seem like an abstract luxury that can wait until later. Further, in states with more mature QRISs, there may be some reluctance among stakeholders to assess the validity of an established and accepted quality improvement system. In newer state systems, policymakers may question the need for validation, given arguments recently offered in support of establishing a QRIS system (Zellman & Fiene, 2012; Zellman, Brandon, Boller, & Kreader, 2011). Yet early and ongoing validation research is essential to the long term success of any system.

One challenge is that QRIS validation cannot be determined by a single study. Instead, validation should be viewed as an iterative process with several equally important goals: refining the QRIS quality standards and ratings, improving system functioning, and increasing the credibility and value of rating outcomes and the QRIS system as a whole. A carefully designed validation plan can promote the accumulation of evidence over time that will provide a sound theoretical and empirical basis for the QRIS (AERA, APA, & NCME, 1999; Kane, 2001; Zellman & Fiene, 2012). Ongoing validation activities, carried out in tandem with QRIS monitoring activities (those that examine ongoing implementation processes) and evaluation activities (those that examine specific outcomes) can help a QRIS improve throughout its development, implementation, and maturation (Lugo-Gil et al., 2011; Zellman et al., 2011).

QRIS validation research may produce three important benefits. First, validation evidence can promote increased support for the system among parents, ECEC providers, and other key stakeholders. Ratings that mirror the experiences of parents and providers can build trust and increase the overall credibility of the system. Second, a system that is measuring quality accurately and specifically should better able to target limited quality improvement resources to programs and program elements most in need of improvement. This should result in more targeted and effective supports for programs striving to offer higher-quality services. Third, validation evidence can be used to improve the efficiency of the rating process. If a QRIS is expending resources to measure a component of quality that is not making a unique contribution to a summary quality rating, is not measuring quality accurately, or is not contributing to desired program outcomes, that component can be removed or revised. For example, measures that vary little across providers, whose quality varies substantially in other ways, make little or no contribution to overall quality ratings (Zellman & Fiene, 2012).

Four approaches to validation

A comprehensive QRIS validation plan includes multiple studies that rely on different sources of information and ask different but related questions. We suggest QRIS validation research be organized around four complementary approaches: key quality concepts; quality measurement; ratings outputs; and links to child outcomes (Zellman & Fiene, 2012). Summaries of these approaches are provided in Table 1, which includes the purpose of each validation approach, the types of research that can be undertaken, the questions that are asked, and some limitations of each approach. The four approaches are also elaborated later in the paper, as we summarize results of validation research in Indiana and Maine.

In reviewing the table, and throughout this paper, we use three key QRIS terms: component, standard, and indicator. The term 'quality component' refers to broad quality categories used in QRIS (such as staff qualifications, family engagement, or learning environment). A 'quality standard' is defined as a specific feature of quality, such as specialized training in the use of developmentally appropriate curriculum or developmental assessment training within the staff qualifications component. A set of quality standards comprise each quality component. 'Quality indicators' are the specific metrics used for each quality standard. A given quality standard may have one or more quality indicators. An indicator related to the curriculum/assessment staff training standard may be, for example, "At least 50% of teaching staff have completed the two-course statewide training session on developmentally-appropriate curriculum."

QRIS validation in Indiana and Maine

This section will describe efforts at QRIS validation in two states in order to explore current validation efforts using these four approaches and to identify the successes and challenges experienced in these early QRIS validation studies. In Indiana and Maine, the QRIS designs are similar, but some aspects of the states' child care contexts, specific QRIS quality components, standards, and rating processes employed are somewhat different. Both states launched their QRIS statewide in 2008, and both systems have four quality tiers, referred to as "levels" in Indiana and "steps" in Maine, organized into a "building block" framework, meaning that child

Table 1
Four related approaches to validating a QRIS.

Approach	Activities and purpose	Typical questions	Issue and limitations
1. Examine the validity of key underlying concepts.	Assess whether basic QRIS quality components and standards are the "right" ones to include 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?	*Process subject to interpretation and to political pressure *Limited empirical evidence available; few established links to outcomes of interest.
2. Examine the measurement strategy and psychometric properties of measures used to assess quality.	*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 relationships among quality measures to assess whether they function as expected.	*What is the reliability and accuracy of indicators collected using different methods? *Do quality measures perform as expected? (e.g., do expected subscales emerge?) *Do measures of similar standards relate more closely to each other than to other measures? *Do measures relate to each other in theoretically consistent ways?	*This validation activity is especially important given that some quality component measures were likely developed in low-stakes settings and have not been examined in the high-stakes context of QRIS.
3. Assess the outputs of the rating process	*Examine variation and patterns of program-level ratings within and across program types, to assess if QRIS distinguishes levels of quality. *Examine relationship of program-level ratings to other validated quality indicators to determine if ratings are assessing quality in expected ways. *Examine alternate cut points and combining rules to determine how well the ratings distinguish different levels of quality.	*Do rating distributions vary by program type, e.g., center-based programs vs. home-based programs? *Do programs with different program-level ratings differ in meaningful ways on alternative quality measures? *Do levels cut scores and combining rules produce expected rating distributions and meaningful distinctions among programs?	*Measurement error is an important issue that should be examined. *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. *Comparing QRIS measures to other measures is frequently constrained by the absence of validated alternative measures of the same constructs.
4. Examine how ratings are associated with children's outcomes.	Examine the relationship between program-level ratings and selected child outcomes to determine whether higher program ratings are associated with better child outcomes.	Do children who are attending higher-rated programs have greater levels or 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 may be done on selected child and program samples to save resources. *Given staff turnover and movement, children may not spend substantial time with quality-rated caregivers (dosage.) *Measurement limitations noted for Approach 3 apply here as well. *Findings from non-experimental studies do not permit attribution of causality about QRIS participation, but preliminary inferences can be made about how quality influences children's outcomes.

care providers must enter at the lowest level and meet all quality standards and indicators at each level in order to advance to the next higher level. The focus on these two states in this paper is to help illustrate the application of these four approaches to operationalizing validation in a QRIS. While the QRIS evaluations in Maine and Indiana have resulted in other kinds of information disseminated for policy makers in these states and publications for other audiences, this paper is unique in that it is only intended to focus on these four concepts of validation.

Both states partnered with university-based researchers to conduct validation research, after piloting aspects of their QRIS design. However, there are also key differences between these two states. For example, the Indiana QRIS standards were developed based on a local community-based model that was then modified by a state stakeholder committee for statewide expansion. The Maine quality standards were developed to align with program-type-specific national accreditation standards. The Maine standards were also vetted through review and comment by many stakeholders and technical assistance was provided by University researchers based on reviews of the scientific literature. Maine QRIS ratings are generated by provider self-report, then verified by state agency staff, while Indiana employs independent raters who directly assess the standards by visiting child care settings. Provider voluntary participation rates are higher among state-licensed providers in Indiana. However, Indiana also has significant numbers of license-exempt child care providers, whereas license exemption is not a prominent feature of the Maine child care system. The key features of each state QRIS are summarized in Table 2. These two states provide useful examples, because while the state child care contexts are different, they each used strategies contained in the four validation approaches discussed above and outlined in Table 1. The successes and limitations of these states' approaches will inform future validation research on QRIS.

Table 2
Key features of Indiana and Maine QRISs.

QRIS feature	Indiana	Maine
System name	Paths to QUALITY (2008)	Quality for ME (2008)
Eligible child care types (participation rate)	Licensed centers (89%) Licensed homes (62%) Unlicensed registered ministry centers (12%)	Licensed centers (68%) Licensed homes (44%)
Participation rules	Voluntary for all providers (Not required for CCDF participation)	Required for programs participating in federal CCDF/Voluntary for all others
QRIS structure	Building Block; 4 quality levels	Building Block; 4 quality levels
QRIS standards (examples)	Level 1 – Licensed, or completes voluntary certification program	Step 1 – Meets all regulatory standards, in operation for more than one year, and all staff registered in Maine Roads to Quality Registry (MRTQ).
	Level 2 – Learning environment and materials requirements; daily literacy activities; 25% of staff have CDA or equivalent; 15 hrs. in-service training/yr.; etc.	Step 2 – Learning Environment/Developmentally Appropriate Practice requirements; program improvement plan in place; 50% of staff at level 5 on MRTQ career lattice; etc.
	Level 3 – Written curriculum focused on whole child; provision for special needs; 50% of staff have CDA or equivalent; 20 hrs. in-service training/yr.; etc.	Step 3 – Documented use of Early Childhood Learning Guidelines and/or Infant-Toddler Learning Guidelines; Evidence collected at least three times per year on child's development; etc.
	Level 4 – National accreditation; Provide mentoring to other QRIS providers (see www.in.gov/fssa/carefinder/2554.htm#).	Step 4 – National accreditation; written parent involvement plan; etc. (See: https://www.maine.gov/dhhs/ocfs/ec/occhs/qualityforme.htm).
QRIS standards development process	Aligned with national accreditation standards, expansion of community pilot program, modified by state stakeholder committee.	Aligned with national accreditation standards; based on available empirical evidence; and from results of stakeholder reviews and comment.
QRIS rating procedure	Independent ratings contractor, annual site visits, using Paths to QUALITY standards checklist	(1) Provider self-assessment, online system; (2) enrollment system uses linked files from state licensing and registry; and (3) verified by state agency staff.

Method

Indiana

The Indiana QRIS is called "Paths to QUALITY™." The validation research reported here includes a preliminary literature review and an empirical field study including a stratified random sample of 276 child care providers who had voluntarily entered the QRIS during 2008–2009, including 135 classrooms in 95 licensed child care centers, 169 licensed family child care homes, and 14 classrooms in 12 unlicensed registered child care ministry centers. Independent, on-site assessments were completed by university researchers approximately one year after QRIS entry and included: observational global quality assessments of the child care environment using the Environmental Rating Scales (ERS: ITERS-R, ECERS-R, FCCERS-R; Harms et al., 2005); observations of adult–child interaction quality (Caregiver Interaction Scale, CIS; Arnett, 1989); surveys and interviews with child care providers; and interviews with parents whose children had been placed with QRIS providers. Observers were trained to reliability level of 80% exact agreement (*Kappa* = .70) or higher, and maintained reliability during the study. Child development assessments were completed using standardized research-validated measures, with two randomly selected children from each participating child care center classroom or family child care home. For children under three years, measures included the Mullen Scales of Early Learning (Mullen, 1995) for cognitive and language development and the Brief Infant Toddler Social Emotional Assessment (BITSEA; Briggs-Gowan & Carter, 2002) for social-emotional development. For children three to five years, the measures included the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 1997) and Woodcock–Johnson Applied Problems and Letter–Word Identification subtests (Berry, Bridges, & Zaslow, 2004) for language and cognitive development and the Social Competence and Behavior Evaluation (SCBE; LaFreniere & Dumas, 1997) for social–emotional development. (For a detailed description of the Indiana evaluation methodology, see Elicker et al., 2013; Elicker, Langill, Ruprecht, Lewsader, & Anderson, 2011.)

Maine

The Maine QRIS is called "Quality for ME." The Maine validation research reported in this paper is based on a literature review of quality variables, focus group interviews with providers and parents, and a field study including a stratified random sample of 255 providers who enrolled in the QRIS in 2008 through 2011, including: 153 classrooms in 105 licensed child care centers; 113 licensed family child care homes; and 41 classrooms in 37 Head Start sites. Assessments were completed as soon as possible after a program enrolled into the QRIS, however, this varied based on the length of time required for the state agency to verify enrollment information and schedule on-site observations. Like Indiana, Maine researchers used the ERS global quality assessment scales and conducted surveys with providers and parents. Unlike Indiana, Maine did not collect any child-level outcome data. Assessors were trained to reliability annually by authors of the ERS scale and maintained a 85% inter-rater reliability during the study. (For a detailed description of the evaluation methodology, see Lahti et al., 2011.)

Results

Results of the QRIS validation research in Indiana and Maine are presented in relation to the four approaches to validation recommended by Zellman and Fiene (2012; refer to Table 1).

Approach 1: examine the validity of key underlying concepts

As noted above, the quality components included in a QRIS (e.g., staff qualifications, learning environment, family engagement) essentially define how child care quality will be viewed in each state. Conceptual validation provides justification and support for these chosen elements. This first validation approach asks whether the quality components, standards, and indicators included in a QRIS are the "right" ones; that is, if together they define quality of care. Many state QRISs have adopted similar, though not identical, concepts and program quality standards (Smith, Robbins, Stagman, & Kreader, 2012).

One approach that can help to validate the underlying concepts of quality in a QRIS involves assessing the degree to which the quality components used in the QRIS ratings include standards and indicators that are based on empirical evidence that links them to desired program, family and child outcomes. A literature review weighs the existing research evidence and on that basis provides a judgment about whether a particular quality component should be included or excluded from the QRIS. Like many validation activities, such reviews ideally would be updated from time to time to determine if revisions to the QRIS are advisable in light of new research findings. As noted in Table 1, this approach may be limited by available data. Further, available data may be subject to more than one interpretation. Politics can also play a role; supporters of particular elements, e.g., nutrition, accreditation, may want to ensure that such measures are included, regardless of the strength of the research evidence. This literature review approach of conceptual validation was a key method used in developing both Indiana's and Maine's QRIS quality standards.

Indiana: examining the validity of underlying concepts

Standards and indicators for each QRIS level in Indiana were drafted by a state committee of child care providers and stakeholders. The standards were based on an existing community-level Paths to QUALITY model, but also made accommodations for statewide use and integration into the existing state child care licensing and training/technical assistance systems. The highest level quality goal for QRIS in Indiana is national accreditation, so proposed quality standards and criteria at each QRIS level were constructed to help child care providers work toward accreditation in steps.

The Purdue University research team conducted a review of previous evaluations of the Indiana QRIS community-level pilot programs and an in-depth analysis of the proposed QRIS quality standards based on the published child development and child care literature. This literature-based analysis is summarized here. (For a full report, see Elicker, Langill, Ruprecht, & Kwon, 2007; Elicker et al., 2013.)

First, the evaluators looked at each proposed QRIS quality standard and indicator for each type of child care. Ten broad quality components were identified that encompassed all of the proposed quality indicators: regulation; teacher education/training; structural/environmental quality; process quality/interactions; assessment; provisions for children with special needs; program policies; director/owner professional development; parent–teacher communication; and national accreditation. These ten components were then used as key terms to guide an extensive search of the research literature to collect and weigh the available evidence that each component was: (1) generally considered a valid aspect of quality; and (2) empirically associated with children's well-being or positive developmental outcomes. Based on the amount and quality of evidence, each quality component was the rated as follows: (1) some or limited evidence (one or two well-designed studies); (2) moderate evidence (3–5 well-designed studies); or (3) substantial evidence (more than five well-designed studies). The results of this analysis were reported to the state QRIS planning committee, including a conclusion that most of the proposed quality indicators had "substantial evidence" for their validity.

Maine: examining the validity of underlying concepts

Researchers at the University of Southern Maine worked with state agency leaders and other key stakeholders through a process that involved the use of Concept Mapping (The Concept System®, 2012). This process allows for the development of a conceptual framework that can guide planning, and in this case led to the selection of the underlying quality concepts and standards for Maine's QRIS. Similar to what was done in Indiana, University of Southern Maine research staff identified key quality concepts from the literature and national accreditation standards. In addition, concepts emerged from results of eight focus groups with parents and ECEC professionals across the state, including participants from various types of settings, e.g., family child care homes, after school programs, centers, etc. Statements of program quality were developed; these statements were the focus of a mapping process which involved more than twenty-four experts reviewing and rating the statements. The Concept Maps that result from this process allowed participants to visually identify which concepts of program quality were most favored by specific key stakeholder groups. In addition, the mapping software illustrated how closely related the concepts were to each other, based on reviews from the select experts. From this process, a set of components and standards was developed. The final step in selecting program quality standards involved a formal review and comment process that the state agency implemented in various locations across the state (Maine DHHS, 2008).

Approach 2: examine measurement strategies and psychometric properties of quality measures

A second type of validation focuses on the attributes of the individual quality measures used in the QRIS and the way these measures are combined to produce a summary rating of program quality. This approach addresses how well measures are working in the context of the QRIS. These efforts attempt to answer questions such as, "Is there evidence that a given indicator measures what it purports to measure?" "If the QRIS claims to have a specific number of dimensions, do we find those dimensions in the output data?" "Is there sufficient variance in scores on this quality indicator to justify its inclusion in the QRIS?" Addressing these issues involves an examination of the distribution of participating provider quality scores and the internal consistency of multi-item measures.

The research literature provides limited guidance concerning the most appropriate ways to combine measures of quality indicators into summary ratings (Lugo-Gil et al., 2011; Tout, Zaslow, Halle, & Forry, 2009; Zellman, Perlman, Le, & Setodji, 2008). Yet this process is crucial to producing meaningful overall program quality ratings, the key output of the rating assessment process. At minimum, it is important to consider whether certain elements *should* be treated as more important, and if so, how this can be assured in the process of combining them. If this issue is not addressed, unexamined weighting may occur anyway. For example, if measures of individual quality elements are combined without any weighting, then those measures that are longer (e.g., include more items) will count for more in a final rating.

At the time both the Maine and Indiana QRISs were being designed, in the mid-2000s, the predominant global quality measures in use in both states were the Environmental Rating Scales (ERS) (Harms et al., 2005; Harms, Cryer, & Clifford, 2006; Harms, Cryer, & Clifford, 2007). ERS use was predominant in the accreditation quality improvement efforts in both child care centers and child care homes. So there was some familiarity with the measures on the part of providers. This was an important political consideration in terms of developing and promoting the design of the QRIS. In addition, in reviewing emerging QRIS work from other states, it appeared that the ERSs were the predominant global classroom quality measure in use at that time. While ERS was influential in the design of the QRIS quality standards in both Maine and Indiana, it is important to note that the ERS are not used to determine the step or level quality ratings. Many other quality indicators are included in the QRIS standards of both states, including staff qualifications, annual staff training hours, and other indicators that help providers make progress toward the ultimate quality goal of national accreditation.

Choosing the points at which individual measures (in block design QRISs) and summary ratings are assigned to rating levels is another exercise that has received limited attention. Cut scores can be assessed in a number of ways. One relatively simple one is to use existing data to conduct a "virtual pilot" (Zellman & Karoly, 2012a) in which existing data are used and cut scores are altered and the effects are examined in terms of distributions of summary ratings across programs. A downward limit on cut scores is the need for some variation within each quality component; without it, a component provides no useful information in overall ratings. Designers may compare program distributions using different cut scores, although it is not always clear what an appropriate rating levels distribution should be. However, it is reasonable to assume that an appropriate distribution in the early phase of a QRIS would be one in which there are programs placed at all levels, with decreasing numbers of programs at each succeeding higher level.

Another validation activity might involve an assessment of the relationship of a given indicator to other indicators of quality included in the QRIS. In studies that examine measures to be included together in a QRIS, it is important to look at the degree of correlation found among these measures: ideally, measures will be moderately correlated so that each measure both contributes to an overall assessment of quality yet also provides some non-redundant program quality information (Zellman et al., 2008). Correlation patterns should make sense. For example, two measures of interaction quality should be more closely related to each other than to a measure of adult–child ratios. If such studies reveal for example that the correlation between ratios and interaction processes is very high ($r = .90+$) this result might argue for eliminating one or the other indicator from the QRIS, as they may not be providing unique information (although some QRISs include certain quality elements to ensure that they are paid attention to for other policy related reasons, even if their psychometric properties are not ideal). To date, the Maine and Indiana validation research has not included a comparison of measures internal to the QRIS rating systems, but this is recommended in future research as the systems mature and stabilize.

Measurement error presents another potential challenge in assessing QRIS validity. Most QRISs assume that observational measures are relatively stable over time absent quality improvement efforts. This assumption is consistent with empirical evidence for at least one widely used instrument, the ERS (Clifford, 2005).

A related measurement issue concerns inter-rater reliability. In the twenty systems reviewed by Tout et al. (2010), nearly all QRISs require 80–85% agreement with a master coder (either exact agreement or agreement within one scale point) on ERS; this degree of reliability does not eliminate errors in ERS measurement (Bryant, 2010; Bryant, Burchinal, & Zaslow, 2011). For instance, two raters could be 100% reliable under a standard of 85% agreement within one scale point, but one might give a classroom a score of 3.5 and the other a score of 4.5, a difference that is large enough to affect an overall program rating (Karoly, Zellman, & Perlman, 2013). Based on the range and degree of variability in ERS quality scores at each rated level in both Maine and Indiana QRISs (see Tables 3 and 4), we recommended that program managers strive to increase the reliability of the rating process by clearly defining quality indicators and rating procedures, and conducting regular reliability checks.

Approach 3: assess the outputs of the rating process

A third validation approach focuses on assessing the *outputs* of the rating system: the scores and levels assigned to providers who undergo a rating, and the distributions of those scores within and across different types of providers. Studies conducted under this approach examine the degree to which the quality levels in the QRIS are meaningfully distinct from each other. The results of these
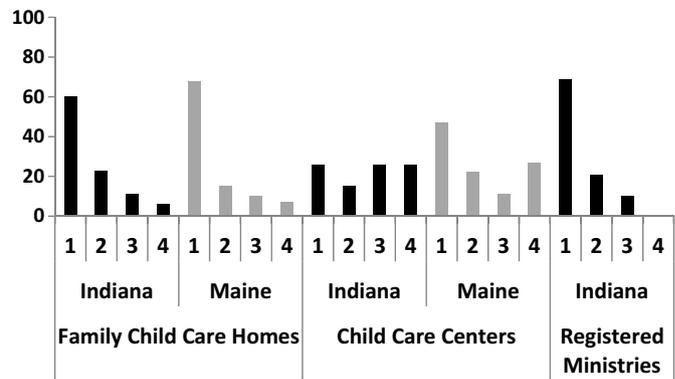

Fig. 1. Percent enrolled providers rated at four quality levels, by state and type of child care.

studies may provide data that suggest that measures, cut scores, or rules for combining measures need to be changed in order to distinguish the rated quality levels effectively. Because these studies can result in proposals for significant changes to the standards for QRIS levels, it is helpful for these studies to occur prior to studies that examine associations between quality levels and children's development.

Output studies may focus on individual indicator scores, such as how providers score on an environmental rating, as well as on the overall quality level that is the final output of the rating process. These studies may also utilize a measure of quality not included in the QRIS rating process to make an evaluation of concurrent validity, by examining whether assessments on both measures co-vary in predictable ways. The following section provides examples of the two states' examinations of the distribution of quality ratings and rating-level advancement patterns for each program type enrolled in the state QRIS.

Examining initial QRIS rating distributions and cut points

While evaluators in Indiana and Maine did not conduct a detailed examination of the weighting or internal consistency of specific quality indicators, they did analyses to reveal the distribution of quality levels. After three years of system implementation, both Indiana and Maine QRIS child care providers were predominately rated at Level 1 or Level 2 (see Fig. 1). It is important to note that in Indiana, all providers enter the system at Level 1, and in Maine, providers can enter the system at any level based on their program rating, and then may advance at will from that level. A recent in-depth study of five state quality rating and improvement systems that were fully implemented found a similar pattern, with four of the five states reporting 40–76% of all programs enrolled in the lower tiers of the system (Mathematica Policy Research, 2011).

In Indiana, licensed child care centers were evenly distributed across the four QRIS levels approximately two years after the program inception. However licensed family child care homes were most frequently found at Level 1, with steeply declining numbers at the other three levels. This higher proportion of Indiana licensed centers rated at Level 3 or Level 4 may have been due to a greater historical emphasis in child care centers than in homes on regulation and attaining national accreditation, greater organizational capacity to complete the requirements of advancement in QRIS, or possibly that QRIS standards more closely reflect center quality than family child care home quality. Unlicensed registered child care ministries, a unique type of child care center in Indiana that is not licensed due to religious affiliation, participated at a much lower rate, and none had yet attained Level 4, reflecting significant challenges facing these unlicensed centers in meeting the Level 1 standards needed to enter the QRIS. These data patterns in Indiana

Table 3
Indiana QRIS: mean global quality ERS scores[a] as a function of program type and rated quality level.

	Level one ($n = 84$)	Level two ($n = 90$)	Level three ($n = 74$)	Level four ($n = 66$)
All providers ($N = 314$)	3.2 (.87)	3.7 (.76)	3.8 (.73)	4.3 (.80)
Family child care homes ($n = 167$)	2.9 (.64)	3.4 (.75)	3.6 (.67)	4.0 (.89)
Licensed child care centers ($n = 133$)	4.0 (.77)	4.0 (.68)	4.3 (.66)	4.5 (.67)
Unlicensed registered child care ministries ($n = 14$)	3.2 (.95)	4.1 (.45)	4.0 (.18)	NA

[a] Possible range = 1–7.

Table 4
Maine QRIS: mean global quality ERS scores[a] as a function of program type and rated quality level.

	Step one ($n = 82$)	Step two ($n = 99$)	Step three ($n = 79$)	Step four ($n = 82$)
All providers ($N = 342$)	3.7 (.77)	3.9 (.84)	4.0 (.80)	4.3 (.79)
Family child care homes ($n = 129$)	3.3 (.67)	3.5 (.80)	3.8 (.91)	4.2 (.83)
Licensed child care centers ($n = 165$)	3.9 (.72)	4.1 (.80)	4.2 (.68)	4.4 (.84)
Head start centers ($n = 48$)	NA	NA	4.1 (.75)	4.5 (.71)

[a] Possible range = 1–7.

supported the validity of the QRIS rating system in that they showed variation in quality ratings across participating providers, they reflected the increasing effort necessary to meet quality standards at higher levels, and they were interpretable within the state's child care context.

Another gauge of overall quality rating system utility is the amount of program advancement to higher rated levels. It is reasonable to expect, if the QRIS is viable, that at least some providers will advance in quality level. In the Indiana evaluation, 19% of the licensed centers, 24% of the licensed homes, and 27% of the unlicensed ministry centers had advanced at least one QRIS quality level in a 6-month period between assessments, during which mentoring was provided by local training providers. This advancement pattern, if maintained over time, suggests that even though attaining the highest levels may be challenging, quality improvement is feasible.

For Maine, as Fig. 1 illustrates, center-based programs and family child care type programs are most frequently found at Step or Level One. A disproportionately small number of family child care programs have attained Step Four, the highest quality level, and a disproportionately large number of child care centers and Head Start programs are enrolled at Step Four. This pattern of fewer family child care homes enrolled at higher Step levels has existed throughout QRIS implementation in Maine. Maine family child care home providers argued that some of the program standards were not "a good fit," despite designers' beliefs that standards were well-matched to setting type. The large number of center-based and Head Start programs at the higher Step levels was expected, given that QRIS quality standards closely align with accreditation standards, and center-based programs are more likely to be nationally accredited than family child care homes.

An assumption of the designers of the QRIS in Maine was that programs engaged with QRIS will improve their tier levels consistently over time (Lahti et al., 2011). Approximately 80% of all programs ($n = 1118$) in the QRIS observed during the study period 2008 through 2011 did not experience a move up from one Step Level to the next. Results indicated that 95 of the 103 events or changes in Step Level from level one to two occurred during the first 23 months of enrollment. Moving from a Step One to Two, center-based care programs had a hazard probability of just .02 while family child care homes stayed virtually flat during this early period of enrollment in the QRIS. For movement from Step Two to Three, neither program type ($p = 290$) nor regional location ($p = 195$) appear to be significant in explaining Step level movement. For movement from Step Three to Four, the highest tiers in Maine's QRIS, the analysis indicated that only type of program is a significant covariate explaining advancement. Family child care homes

appeared to have a significantly lower probability of advancing a Step Level at this highest quality tier, compared with center-based and Head Start programs. These types of analyses of program movement in the system are relevant to the validation of a QRIS as they illustrate whether or not the way the system as designed is meeting its goals of supporting program advancement, leading to statewide improvement of program quality.

Studies may also be conducted to examine the degree to which given measures relate to other measures that purportedly assess the same concept. Here, strong correlation is desired, as they suggest that measures are measuring the concepts that they purport to measure in ways that are consistent with other measures of the same concepts.

Indiana: assessing the output of the rating process

The Indiana evaluation research included one validation test of state committee-generated quality standards, indicators, and levels by comparing the outputs of the QRIS rating system with independently gathered assessments of quality using validated quality measures, the environmental rating scales (ERS, Harms et al., 2005, 2006, 2007) and the Caregiver Interaction Scale (CIS; Arnett, 1989). The results, originally published by Elicker et al. (2011) and shown in Table 3, indicate that ERS scores co-varied as expected with QRIS level ratings, with a significant mean difference in global scores of 1.1 scale points between Level 1 and Level 4. Table 5 shows that caregiver interaction as observed using the CIS was less related to the rated QRIS quality levels. The overall correlation between the 4-level QRIS ratings and global ERS quality scores was moderate ($r = .42$, $p < .01$). The correlation between CIS adult–child positive interaction scores and QRIS level was more modest, but positive and significant ($r = .24$, $p < .01$).

Taken together, and looking across all types of providers, these results suggest that the QRIS ratings distinguish levels of quality in somewhat similar ways as two time-tested, validated measures of child care quality. However, mean quality levels at Level 4 were mostly found to be below the "good" rating threshold, suggesting the need to strengthen standards and/or rating procedures at the highest QRIS levels. In addition, finer analysis of the data suggested specific recommendations about quality standards and rating procedures that might be improved for each type of child care. Summaries of individual ERS item means for Level 3- and 4-rated providers led to the identification of a number of ERS items with scores below 4. Program planners are currently improving standards and QRIS rating procedures in light of these findings (Elicker et al., 2013).

In Indiana, patterns of association between QRIS ratings and ERS ratings were not the same for all types of child care. While the

Table 5
Indiana QRIS: mean total positive interaction (CIS) scores[a] as a function of program type and rated quality level.

	Level one ($n=84$)	Level two ($n=87$)	Level three ($n=74$)	Level four ($n=65$)
All providers ($N=310$)	3.0 (.44)	3.1 (.46)	3.2 (.37)	3.2 (.52)
Family child care homes ($n=164$)	2.9 (.37)	3.1 (.43)	3.2 (.36)	3.3 (.48)
Licensed child care centers ($n=132$)	3.2 (.44)	3.1 (.50)	3.2 (.41)	3.2 (.54)
Unlicensed registered child care ministries ($n=14$)	2.8 (.58)	3.2 (.48)	3.5 (.11)	NA

[a] Possible range = 1–4.

global ratings were significantly correlated in both licensed centers and licensed family child care homes, the strength of association was stronger for homes, meaning QRIS level ratings in homes more clearly distinguished levels of ERS-related quality, meaning at each QRIS-rated level, the ERS quality differences were generally greater than they were for centers. Second, the overall ERS quality levels for center-based preschool classrooms (using ECECRS-R; $M = 4.6$ at Level 4) were somewhat higher than for center-based infant classrooms (ITERS-R; $M = 4.4$ at Level 4) and family child care homes (FCCERS-R; $M = 4.0$ at Level 4). While the equivalence of quality scores across these three ERS scales is not supported by research evidence, the results taken together suggest the need to strengthen quality standards and assessment procedures for all types of care, so that child care providers at the highest rated levels are providing care that is at or above threshold levels recommended to impact children's developmental outcomes (Zaslow, Martinez-Beck, Tout, & Halle, 2011).

Maine: assessing the output of the rating process

As in Indiana, differences in program quality were measured using Environmental Rating Scales (ERS) mean scores at the classroom level, and these scores were not part of the QRIS standards or ratings. The results presented here are from factorial ANOVAs to examine the effects of Step Level, ERS scale type and child care program type on the dependent variable ERS mean score. Table 4 provides the adjusted mean scores for all 307 classrooms and by each program type by Step Level. Table 4 is extracted from the full report on Maine's QRIS (see Lahti et al., 2011).

The results show an overall significant difference between Step Level and ERS mean score at the classroom/setting level ($F = 5.02$; $df = 3, 307$; $p = .002$). Results of post hoc Bonferroni tests showed a significant difference between Step One and Step Four programs ($p = .001$) and between Step Two and Step Four programs ($p = .001$). The total variance of the mean ERS score explained by Step Level was only 5%, indicating weak relationships between the variables. Comparisons of the program type mean ERS quality scores indicated a difference only between the family child care home scores and the center-based scores ($p < .001$). The family child care home mean scores were lower at each Step Level than the center-based setting scores with the exception of scores at Steps Three and Four. There did not appear to be any significant differences at Step Three or Four between the center-based and Head Start type settings ($p = .97$). The results provide some evidence for differences in rated quality, with higher ERS means for higher tier or step programs, most distinctly for family child care homes. Overall these mean scores suggest the need for considerable efforts at quality improvement, considering that the majority of settings are scoring below the "5" or "good" level on the ERS measures.

Maine: parent level data on QRIS program quality

In the Maine validation study, parents in programs selected for observation were asked to complete an anonymous survey that focused on services received by the parent and the parent's perceptions of the quality of the program. The belief was that parents served by higher Step level programs should be receiving more supports and services and therefore may rate the program higher in

level of quality. The response rate over the three-year study period was approximately 26% ($N = 1478$). These results are extracted from the full report on Maine's QRIS (see Lahti et al., 2011).

Parental perception of program quality was measured by the 15-item Emlen scale, see Emlen, Koren, and Schultze (2000), and was found not to be correlated to Step Level rating (Pearson's $r = .010$, $p = .68$). In terms of services parents should have received according to program standards, a majority of parents reported not receiving: information about other government services for their child; opportunities for parent engagement with the program; daily communication from the program about their child; and being provided an up to date written parent hand-book from their provider. There did not appear to be any difference in step level in terms of parents not consistently receiving these types of services as required by the program quality standards according to parent reports. The use of these data by QRIS administrators was primarily for monitoring purposes focused on services and or supports parents should have received based on requirements in the QRIS standards. While parents were asked about perceptions of program quality, due to a strong desire to reach out to parents as a key stakeholder in the QRIS, that information was not relied upon for program planning or program improvement.

Approach 4: relate ratings to children's development

The fourth approach to validation focuses on children's development. In many respects, this is the final step in validating a QRIS, and one that arguably should be delayed until the questions raised in the earlier approaches are addressed and changes made to the system as necessary. It may even be possible that new data will emerge that makes the costly and difficult effort involved in assessing child outcomes unnecessary. For example, if studies begin to show consistently that certain inputs, e.g., ratings-based coaching lead to substantial improvements in indicators such as instructional support, and if instructional support or other indicators is found to consistently promote improved child outcomes, it may be possible to argue that the inclusion of those inputs and measures of those outputs may suffice.

The logic models that underlie QRISs typically assert that higher quality care will be associated with improved child outcomes. Therefore, one important piece of validation evidence concerns whether children make greater developmental gains in programs with higher program-level QRIS ratings than in programs with lower ratings. While a definitive evaluation of QRIS impact on child outcomes would consist of an experimental study with random assignment of providers and children to QRIS levels, ethical and practical considerations often make experiments impractical, at least on a state-level scale. Instead, current studies evaluating QRIS validity in terms of child outcomes using this approach do not attempt to evaluate causal linkages. Instead, they examine whether the QRIS ratings and the quality components that comprise the ratings are associated in expected ways to measures of children's development. Showing significant associations between QRIS-rated quality would be a first step, a necessary but not sufficient result to demonstrate causal inferences about how QRIS quality influences children's outcomes.

To date, few QRIS validation studies have incorporated children's outcomes. Maine did not include this approach to validation. As Elicker and Thornburg (2011) note, results from such studies are mixed, at least in part because of the challenges of conducting them. A primary challenge is the inability to control for all the factors that may confound the quality-outcome correlations for children whose families have selected programs in a non-random way. Additional challenges include the difficulty of recruiting of programs and children across all quality levels; lack of information about the amount of care children received in each setting (dosage); lack of appropriate outcome measures for children of diverse ages, abilities, cultures and linguistic backgrounds; and, lack of variation in the quality of participating QRIS programs. As noted above in the discussion of Approach 3, measurement error remains a problem.

Indiana: examination of ratings associated with children's outcomes

To examine validity-related questions about children's development in the context of the Indiana QRIS, the evaluators assessed the developmental status of 557 children (249 infants/toddlers; 308 preschoolers) who were in the care of QRIS providers. Two children per classroom or home were randomly selected in approximately equal numbers at all four QRIS levels. Data from parent interviews describing annual family income and parents' education levels and participation in the CCDF voucher program were used as control covariates in the analyses. The basic validity question explored was: are children in higher-rated QRIS care functioning at higher levels, socially and cognitively, than children in lower rated care?

It is important to point out that this study of quality and child outcome associations was cross-sectional, with all data collected at one point in time. As mentioned earlier, exploring these correlational relationships does not substitute for longitudinal or experimental designs that can better evaluate the causal impact of the QRIS on child outcomes. However in the implementation phase of QRIS, it is useful to explore the developmental status of participating children, how they are distributed in the child care system, and whether associations between quality measures and measures of children's functioning are occurring in the expected direction (Elicker & Thornburg, 2011).

Bivariate correlations and multiple regression models were used to explore the associations between children's development and the three measures of child care quality: QRIS ratings (4 levels); ERS global quality scores; and a CIS composite rating of positive adult–child interactions. All regression models included parent education, household income, and type of child care. No significant correlations were found between the four-level QRIS quality ratings and either infant/toddler or preschooler developmental status. Some of the researcher-observed quality measures were mildly but significantly correlated with child development measures. For preschoolers, CIS positive interactions were correlated with social competence ($r = .17^{**}$) and receptive language ability ($r = .17$, $p < .01$). For infants and toddlers, ERS global quality scores were associated with social competence ($r = .15$, $p < .01$), and total CIS positive interactions were associated with cognitive/language competence ($r = .17$, $p < .01^{**}$). These significant correlates were entered as predictors in regressions of child outcomes on the quality variables, controlling for the family SES variables (parent education level and household income) that were also significantly correlated with the child outcome variables. As a result, for preschoolers, CIS positive adult–child interactions significantly predicted children's receptive language ability, after controlling for family SES ($b = .12$, $p < .05$). For infants and toddlers, CIS positive adult–child interactions significantly predicted children's cognitive/language competence, after controlling for family SES ($b = .14$, $p < .05$).

Family income was also a significant predictor, $b = .23$, $p = .009$.

Therefore while QRIS rated levels were not significantly associated with any child development measures for either infants/toddlers or for preschoolers, ERS and CIS quality measures were moderately associated with aspects of children's development. Specifically, after controlling for family SES, it was the positive quality of interaction between adults and children that was associated with language and cognitive functioning, for both preschoolers and infants and toddlers.

Therefore it appeared that the specific aspects of child care quality assessed by the ERS and CIS measures are more likely to be associated with children's development than are the composite of quality indicators represented by the 4-level QRIS ratings. This was true even though the QRIS ratings and the ERS and CIS were significantly correlated with each other. As a result, in Indiana, further refinement of QRIS standards and procedures is taking account of these findings, especially by identifying ways to strengthen QRIS standards and ratings to inlcude the quality of adult–child interactions.

Discussion

Limitation to validation study designs

Both of these state studies provide results that describe linear associations among variables. The study designs are limited due to the fact that the investigators have no control over how the QRIS systems are implemented which affects enrollment and therefore sample sizes and selection of measurement strategies were also not in the sole control of the investigator. It will be interesting as additional studies are done and where non-linear associations are found to determine the impact this has on outcomes. These field studies were conducted with all the limitations associated with working in a developing system with multiple stakeholders. While the design presents a limitation in terms of arguing for causality and application of more sophisticated analytic approaches, it should be noted that the state agency program managers and other stakeholders in both states found the information generated from these studies of high value in terms of system planning, program improvements, and resource allocation. Depending upon context and resources, limitations to these two study designs can be remedied in future studies by such design choices as having programs that are on a waiting list be compared to programs already participating in their state's QRIS.

Validation of QRIS is a process that needs attention over time, using more than one approach

The examples from Indiana and Maine illustrate how these validation approaches can work in practice, with tangible benefits for system improvement. These validation activities are specific to the design and implementation of each state QRIS. We believe it is important to stress to QRIS policy leaders that each of the four validation approaches needs to be used appropriately, considering the developmental stage of the QRIS and the unique features of the setting and QRIS. For example, states with QRIS in development can use the four approaches as a framework for planning how to validate their system. Developmentally then, an initial focus of QRIS design would be to validate the key concepts used in the QRIS design. The four approaches highlighted in this paper need to be considered as part of an ongoing process, not a one-time event. As states progress in their implementation of QRISs, more descriptive research is necessary to understand better how these validation approaches work in other settings, for example with point-based QRISs (rather than "building block" QRISs, like those or Indiana and Maine). Use of these approaches enables cross system comparisons,

which will allow for the identification of common threats to validity and useful strategies to enhance the validity of a state QRIS.

Validation and early care and education system constraints

A QRIS is not merely a program-level quality-improvement intervention, it is a policy lever for strengthening a state's overall early care and education system that reaches beyond child care (Schaack, Tarrant, Boller, & Tout, 2012). The two state validation efforts highlighted in this paper reflect the challenges and constraints common to other state experiences with validation activities (Lahti, Sabol, Starr, Langill, & Tout, 2013). On-site observations of global program quality and establishing and maintaining inter-rater reliability for QRIS raters is a time consuming and costly endeavor. Keen interest in school readiness may pressure program administrators to collect child outcome level data before a QRIS is well established. Current research on the measures that are in the widest use to predict child outcomes appear to do so consistently, especially for children at-risk, but with modest levels of association with program quality as measured by the ERS's (for example, Burchinal et al., 2011). We recommend to policy makers to always take into account that any validation study is occurring within a dynamic system. System-level constraints such as varying resources available to programs, different type and design of programs, and challenges to measuring quality and reliably collecting information about program quality all influence the design and implementation of state-level QRIS validation studies.

Validation research is critical for performance measurement and improvement for a state QRIS

The 2012 Child Care Development Fund (CCDF) Plan preprint for fiscal years 2014–2015 includes a much larger focus on QRISs (U.S. Department of Health & Human Services, Administration for Children and Families, 2011). In this document, a QRIS is defined as a ". . .systematic framework for evaluating, improving, and communicating the level of quality in early childhood programs." States are expected to provide a self-assessment based on current program quality initiatives from a set of questions that are also organized according to a "QRIS framework." Validation of program standards or assessment tools is mentioned specifically in relation to information states must provide about data and performance measures on program quality.

The information generated from QRIS validation activities can be used to inform efforts for continuous quality improvement. For example, both Indiana and Maine found that, for at least some types of providers, enrollment patterns in the QRIS, and lack of movement by programs once they are in the QRIS, is resulting in a large proportion of providers at the lower-rated quality levels of the quality tiers. System-level, quality-improvement responses to this information could be to re-assess the design of the system in terms of the ability of programs to meet standards at each tier, or to focus training/technical assistance on specific quality standards that are most challenging for providers to meet. At the same time, care should be taken to ensure that standards reflect current knowledge about the specific indicators and levels of quality most likely to produce the desired child outcomes. Findings from validation studies can be part of the information that state child care administrators use to assess the overall performance of the state early care and education system. The performance data could then be used by program administrators in making decisions about monitoring programs in a differential manner by visiting those programs more often who are having difficulty meeting QRIS standards (Fiene, 2013). We recommend that one focus of future research be learning more about whether and how information from the results of validation studies are used to improve system and program level performance in QRIS.

It will be interesting to determine as more validation studies are completed to analyze the differences between levels and how often lower quality is present in the top level which is the case with measuring compliance with licensing standards (Fiene & Nixon, 1985). Key areas to look at will be the movement of programs from one level to another, how long this takes, and are the increments equal or not in terms of quality improvement.

Another area to be explored which may have an impact on overall QRIS implementation are the fiscal constraints that many states are experiencing due to the recent recession and lower levels of federal funding. It would be interesting to note differences amongst states with large investments in quality improvement initiatives and those states with smaller investments.

It is important to remember that the QRIS is a policy lever and the validation of child care quality standards in a QRIS is a new phenomenon in early care and education policy-making. These four recommended approaches to QRIS validation, illustrated by validation research in these two states, even with their limitations, did provide policy makers and program administrators with information that guided efforts at system quality improvement. The use of these approaches in other studies will create a common nomenclature for better understanding threats to validity in a QRIS and ultimately increase our understanding of how best to design a QRIS that meets the needs of the parents, providers and children it serves.

References

Alkon, A. (2013). *National health and safety tool.* San Francisco, CA: California Child Care Health Program.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*, *10*, 541–552. http://dx.doi.org/10.1016/0193-3973(89)90026-9

Berry, D. J., Bridges, L. J., & Zaslow, M. J. (2004). *Early childhood measures profile.* Retrieved from: http://aspe.hhs.gov/hsp/ECMeasures04/report.pdf

Briggs-Gowan, M. J., & Carter, A. S. (2002). *Brief infant-toddler social and emotional assessment (BITSEA) manual, version 2.0.* New Haven, CT: Yale University.

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

Bryant, D. M., Burchinal, M., & Zaslow, M. (2011). Empirical approaches to strengthening the measurement of quality: Issues in the development and use of quality measures in research and applied settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 33–47). Baltimore, MD: Paul H. Brookes.

Burchinal, M. R., Peisner-Feinberg, E. S., Bryant, D. M., & Clifford, R. (2000). Children's social and cognitive development and child-care quality. Testing for differential associations related to poverty, gender, or ethnicity. *Journal of Applied Developmental Science*, *4*, 149–165. http://dx.doi.org/10.1207/S1532480XADS0403_4

Burchinal, M. R., Vandergrift, N., Pianta, R., & Mashburn, A. J. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in prekindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. http://dx.doi.org/10.1016/j.ecresq.2009.10.004

Burchinal, M. R., Kainz, K., Cai, Y., Tout, K., Zaslow, M., Martinez-Beck, I., et al. (2009). *Early care and education quality and child outcomes (Research-to-Policy Research-to-Practice Brief: OPRE Research-to-Policy Brief #1).* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services, and Child Trends.

Burchinal, M. R., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11–27). Baltimore, MD: Paul H. Brookes.

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. *Early Childhood Research Quarterly*, *25*(2), 140–165. http://dx.doi.org/10.1016/j.ecresq.2009.11.001

Cizek, G. J. (2007, August). Introduction to validity. In *Presentation to the National Assessment Governing Board of NAEP* Washington, DC.

Clifford, R. (2005). Structure and stability in the early childhood environment rating scale. In H. Schonfeld, S. O'Brien, & T. Walsh (Eds.), *Question of quality: Proceedings of a conference on defining, assessing and supporting quality in early childhood care and education* (pp. 12–21). Dublin, Ireland: Centre for Early Childhood Development and Education, St. Patrick's College.

Cochran, M. (2007). *Finding our way: The future of American early care and education.* Washington, DC: Zero to Three.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Elicker, J., Langill, C. C., Ruprecht, K., & Kwon, K. (2007). *Paths to QUALITY™: A child care quality rating system for Indiana. What is its scientific basis?* Lafeyette, IN: Purdue University, Center for Families.

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: Department of Health and Human Services, Administration of Children and Families, Office of Planning, Research, and Evaluation.

Elicker, J., Langill, C. C., Ruprecht, K., Lewsader, J., & Anderson, T. (2011). *Evaluation of "Paths to QUALITY" Indiana's child care quality rating and improvement system.* West Lafayette, IN: Purdue University Center for Families.

Elicker, J., Langill, C., Ruprecht, K. M., Lewsader, J., Anderson, T., & Brizzi, M. (2013). Paths to QUALITY: Collaborative evaluation of a new child care quality rating and improvement system. *Early Education and Development, 24*(1), 42–62. http://dx.doi.org/10.1080/10409289.2013.736127

Emlen, A., Koren, P., & Schultze, K. (2000). *A packet of scales for measuring quality from a parent's view.* Portland, OR: Portland State University, Regional Research Institute for Human Services, The Oregon Child Care Research Partnership.

Fiene, R. (2013). *Differential monitoring logic model and algorithm (DMLMA): A new early childhood program quality indicator model (ECPQIM) for early care and education regulatory agencies.* Middletown, PA: Research Institute for Key Indicators.

Fiene, R., & Nixon, M. (1985). Instrument based program monitoring and key indicators in child care. *Child Care Quarterly, 14*(3), 198–214.

Fuller, B., Loeb, S., Kagan, S. L., & Carrol, B. (2004). Child care in poor communities: Early learning effects of type. *Quality, and Stability, Child Development, 75*(1), 47–65.

Harms, T., Clifford, R., & Cryer, D. (2005). *Early childhood environmental rating scale–revised.* New York, NY: Columbia University Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2006). *Infant toddler environmental rating scale–revised.* New York, NY: Columbia University Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2007). *Family day care environmental rating scale–revised.* New York, NY: Columbia University Teachers College Press.

Howes, C., Whitebook, M., & Phillips, D. (1992). Results on the National Child Care Staffing Study concerning effectiveness and characteristics of teachers across the United States. *Child & Youth Care Forum, 21*(6), 399–414.

Howes, C., Burchinal, M., Pianta, R. C., Bryant, D. M., Early, D. M., Clifford, R. M., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Child & Youth Care Forum, 37*(1), 27–50. http://dx.doi.org/10.1016/j.ecresq.2007.05.002

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Education Measurement, 38,* 319–342.

Karoly, L. A., Zellman, G. L., & Perlman, M. (2013). Understanding variation in classroom quality within early childhood centers: Evidence from Colorado's quality rating and improvement system. *Early Childhood Research Quarterly, 28,* 645–657. http://dx.doi.org/10.1016/j.ecresq.2013.05.001

LaFreniere, P. J., & Dumas, J. E. (1997). Social competence and behavior evaluation in children ages 3 to 6 years: The short form (SCBE-30). *Psychological Assessment, 8*(4), 369–377. http://dx.doi.org/10.1037//1040-3590.8.4.369

Lahti, M., Cobo-Lewis, A., Dean, A., Rawlings, S., Sawyer, E., & Zollitsch, B. (2011). *Maine's quality for ME – child care quality rating and improvement system (QRIS): Final evaluation report.* Portland, ME: Cutler Institute for Health and Social Policy, Muskie School of Public Service, University of Southern Maine.

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.

Lugo-Gil, J., Sattar, S., Boss, C., Boller, K., Tout, K., & Kirby, G. (2011). *The quality rating and improvement system (QRIS) evaluation toolkit (OPRE Report #2011-31).*

Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

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

NICHD Early Child Care Research NICHD Network. (2000). Characteristics and quality of child care for toddlers and preschoolers. *Applied Developmental Sciences, 4,* 116–135.

NICHD Early Child Care Research NICHD Network. (2002). Child care structure → process → outcome: Direct and indirect effects of child care quality on young children's development. *Psychological Science, 13,* 199–206.

Maine, Department of Health and Human Services. (2008). *Quality for ME: Child care quality rating system program standards.* Retrieved from: http://www.maine.gov/dhhs/ocfs/ec/occhs/quality_cbc.pdf

Mathematica Policy Research. (2011). *Defining and measuring quality: An in-depth study of five child care quality rating and improvement systems (OPRE Report 2011-29).* Retrieved from: http://www.whitehouse.gov/sites/default/files/omb/budget/fy2011/assets/management.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system manual, Pre-K.* Baltimore, MD: Paul H. Brookes.

QRIS National Learning Network. (2014). Retrieved from: http://qrisnetwork.org/qris-state-contacts-map

Schaack, D., Tarrant, K., Boller, K., & Tout, K. (2012). Quality rating and improvement systems, frameworks for early care and education systems change. In S. L. Kagan, & K. Kauerz (Eds.), *Early childhood systems, transforming learning* (pp. 71–86). New York, NY: Teachers College Press.

Smith, S., Robbins, T., Stagman, S., & Kreader, J. L. (2012). *Practices for promoting young children's learning in QRIS standards.* Retrieved from: http://www.nccp.org/publications/pdf/text_1070.pdf

The Concept System® (2012). BT computer software. PN Concept Systems, Incorporated, PL Ithaca, NY.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the next decade of quality rating and improvement systems, OPRE Issue Brief.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *Compendium of quality ratings systems and evaluations.* Washington, DC: Child Trends and Mathematica Policy Research, for Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

U.S. Department of Health and Human Services, Administration for Children and Families. (2011). *Program instruction (Child Care and Development Block Grant Act of 1990).* Retrieved from: http://www.acf.hhs.gov/sites/default/files/occ/pi2011_03.pdf

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

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 (MG-650-QEL).* Santa Monica, CA: RAND Corporation.

Zellman, G. L., Brandon, R. N., Boller, K., & Kreader, J. L. (2011). *Effective evaluation of quality rating and improvement systems for early care and education and school-age care (Research-to-Policy, Research-to-Practice Brief OPRE 2011-11a).* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zellman, G. L., & Karoly, L. A. (2012). *Incorporating child assessments into state early childhood quality improvement initiatives.* Santa Monica, CA: RAND Corporation.