Working Paper

Validation Studies for Early Learning and Care Quality Rating and Improvement Systems

A Review of the Literature

Lynn A. Karoly

RAND Education and RAND Labor and Population

WR-1051-DOEL

May 2014

Prepared for the Delaware Office of Early Learning

Abstract

As early care and education (ECE) quality rating and improvement systems (QRISs) have advanced and matured, a number of states and localities have undertaken evaluations to validate the systems. Such efforts stem from the desire to ensure that the system is designed and operating in the ways envisioned when the system was established. Given that a central component in a QRIS is the rating system, a key concern is whether the rating process, including the use of particular measures and the manner in which they are combined and cut scores are applied, produces accurate and understandable ratings that capture meaningful differences in program quality across rating levels.

The aim of this paper is to review the set of studies that seek to validate QRIS rating systems in one of several ways: by examining the relationship between program ratings and objective measures of program quality; by determining if program ratings increase over time; and by estimating the relationship between program ratings and child developmental outcomes. Specifically, we review 14 such validation studies that address one or more of these three questions. Together, these 14 studies cover 12 QRISs in 11 states or substate areas: Colorado, Florida (two counties), Indiana, Maine, Minnesota, Missouri, North Carolina, Oklahoma, Pennsylvania, Tennessee, and Virginia. In reviewing the literature, we are interested in the methods and measures they employ, as well as the empirical results.

To date, most validation studies have found that programs with higher ratings had higher environment rating scores (ERSs), but the ERS is often one of the rating elements. Independent measures of quality have not always shown the expected positive relationship with quality. The handful of studies that have examined how ratings change over time have generally shown that programs participating in the QRIS did improve their quality or quality ratings. Studies that examine the relationship between QRIS ratings and child development are the most challenging to implement and can be costly to conduct when independent child assessments are performed. Consequently, there has been considerable variation in methods to date across these studies. Among the four studies with the stronger designs, two found the expected relationship between QRIS ratings and child developmental gains. The lack of robust findings across these studies indicate that QRISs, as currently configured, do not necessarily capture differences in program quality that are predictive of gains in key developmental domains.

Based on these findings, the paper discusses the opportunities for future QRIS validation studies, including those conducted as part of the Race to the Top–Early Learning Challenge grants, to advance the methods used and contribute not only to improvement of the QRIS in any given state, but also to add to the knowledge base about effective systems more generally.

1. Introduction

Efforts to design and implement early care and education (ECE) quality rating and improvement systems (QRIS) at the state and local level have been underway for more than a decade (Zellman and Perlman, 2008). Although there is considerable variation in the structure of QRISs across jurisdictions, the general approach is to combine multiple indicators of program quality into a single summary quality rating to make the quality of ECE programs more transparent to parents, funders in the public and private sectors, and other interested parties. Ultimately, the goal of a QRIS is to improve the developmental outcomes of the children who participate in ECE programs by raising quality in those domains that are most relevant for children's social, emotional, cognitive, and physical development.

Given the high stakes often attached to QRISs in the form of publicized ratings and financial supports tied to program quality, efforts to validate and evaluate QRISs have proceeded alongside their design and implementation. Sometimes this is done as part of a pilot phase prior to full-scale implementation and other times once the system is operating at scale (Zellman and Fiene, 2012). Such evaluation efforts can determine if the QRIS is well designed, so that it can be deemed to accurately identify both high quality programs, as well as those in need of further improvement. Validation of a QRIS can provide the motivation for programs to participate in a voluntary QRIS, as they understand that quality will be recognized and any shortfalls can be identified and addressed. When validated, parents as consumers will also view the QRIS as a trusted resource as they make decisions regarding ECE programs for their children.

With these motivations, RAND is undertaking an evaluation of Delaware Stars for Early Success (Delaware Stars), the state's voluntary QRIS for centers and licensed family child care (FCC) providers. The primary focus of the evaluation is to determine if the specific quality elements embedded in the rating standards or the overall summary ratings are predictive of children's developmental progress. To achieve this objective, the RAND evaluation will collect data during the fall and spring of the 2014-15 academic year for a sample of Delaware center- and home-based providers on ECE program quality and on measures of child development for the participating children.

The purpose of this document is to review the relevant literature that can inform the design of the Delaware Stars evaluation.[1] In particular, we focus on prior studies that seek to validate QRIS rating systems in one of several ways: by examining the relationship between program ratings and objective measures of program quality; by determining if program ratings increase over time; and by estimating the relationship between program ratings and child developmental outcomes. In reviewing the literature, we are interested in the methods, measures, and findings in order to inform the design of the Delaware Stars evaluation.

[1] This document extends an earlier review reported in American Institutes for Research (AIR) and RAND (2013).

We begin in the next section with a brief overview of the body of QRIS evaluation studies more generally and the specific studies on which we draw. The two sections that follow center on the results of those studies that focus on quality ratings and program quality, and then on the studies that examine quality ratings in relation to child developmental outcomes. The final section highlights the strengths and weaknesses of the prior validation research and points to the implications for the Delaware Stars evaluation.

2. Overview of QRIS Evaluation Studies

As QRIS systems have advanced and matured, a number of states and localities have undertaken evaluations to validate the systems and, where possible, measure the impact of specific system components. The primary goal of a validation study is to determine if the QRIS is well designed and operating in the ways envisioned when the system was established (Zellman et al., 2011; Lugo-Gil et al., 2011; Zellman and Fiene, 2012). In particular, a central component in QRISs is the rating system, so a key concern is whether the rating process, including the use of particular measures and the manner in which they are combined and cut scores are applied, produces accurate and understandable ratings that capture meaningful differences in program quality across rating levels.

To date, QRIS validation studies have primarily focused on one or more of the three validation questions listed in Table 1 (Schilder, 2013).[2] First, one test of the validity of program ratings is to ask whether programs with higher ratings indeed have higher observed quality (see V1 in Table 1). If this pattern is not observed, it may mean that the quality elements included in the rating system do not capture relevant dimensions of quality or that the quality elements are not aggregated in the most effective way. Another validation test is to determine if program ratings improve over time as would be expected from participation in technical assistance (TA) and other supports (see V2 in Table 1). If ratings do not show improvements over time, it may indicate that the rating scale is not sufficiently sensitive to changes in program quality or that the TA and other supports are not effective in improving quality. Yet another test of whether the ratings are meaningful is to determine whether programs that receive higher ratings are actually producing better outcomes in terms of child development (Elicker and Thornburg, 2011; Zellman and Karoly, 2012) (see V3 in Table 1). If they do not, it may be that the rating system is not capturing the dimensions of quality that are most relevant for child development.

[2] Drawing on the validation framework outlined by Zellman and Fiene (2012), Table 1 does not include exercises to determine the validity of the underlying quality concepts in a QRIS or studies that examine the measurement strategies and psychometric properties of the quality measures used in the QRIS. Examples of studies addressing these questions can be found in Lahti et al. (2013). Other relevant questions that have received less attention include whether parents as consumers know about, understand, and use the ratings in making care choices; whether providers value and benefit from participating in the system; and whether the reporting, accountability, and financial aspects of the QRIS are operating effectively.

Table 1. Key Evaluation Questions for Validation Studies

Number	Question
V1	Do programs with higher QRIS ratings have higher observed quality?
V2	Do QRIS ratings or other indicators of program quality for participating programs increase over time?
V3	Do programs with higher QRIS ratings have better child developmental outcomes?

SOURCE: Authors' analysis.

Table 2 summarizes the set of 14 QRIS validation studies, published to date, that address one or more of the three validation questions listed in Table 1.[3] (Studies are listed in alphabetic order by state; one study covers multiple states and has an entry for each state covered.) For each study, the table lists the geographic coverage, the QRIS name (where relevant), the time period covered by the evaluation, the ECE settings included in the study, and the validation questions addressed (referencing V1, V2, and V3 in Table 1).

Together these 14 studies cover 12 QRISs in 11 states or substate areas: Colorado (Zellman et al., 2008), Florida (Shen, Tackett, and Ma, 2009; Malone et al., 2011), Indiana (Elicker et al., 2011), Maine (Lhati et al, 2011), Minnesota (Tout et al., 2010, 2011), Missouri (Thornburg et al., 2009), North Carolina (Bryant et al., 2001), Oklahoma (Norris, Dunn, and Eckert, 2003; Norris and Dunn, 2004), Pennsylvania (Barnard et al., 2006; Sirinides et al., 2010), Tennessee (Malone et al., 2011), and Virginia (Sabol and Pianta, 2012).[4] Other states have evaluations that are unpublished, currently underway, or in the planning stages. In effect, QRIS evaluation, specifically a validation study, has come to be viewed as a required component of QRIS implementation.

For the most part, validation studies have been undertaken in the states that were among the first to implement QRIS, as reflected in the list of states in Table 2. North Carolina and Oklahoma were two of the earliest QRIS adopters (1998), while Colorado, Florida, Indiana, Missouri, Pennsylvania, and Tennessee initiated systems a few years later (between 2000 and 2003). As leaders in the QRIS movement, these states have had more time to plan for and execute validation studies. More recent adopters include Maine and Minnesota, two states that integrated validation efforts as part of a pilot or early implementation phase. Consistent with this sequencing, Table 2 illustrates that the validation efforts have been concentrated in the first

[3] Lahti et al. (2013) review the validation studies for four of the states listed in Table 2: Indiana, Maine, Minnesota, and Virginia. In counting studies, when the same validation study has produced more than one publication, we count that as one study for one QRIS (e.g., Tout et al., 2010, 2011). When a single validation study covers more than one QRIS, we count one study for each QRIS analyzed, even though results may be available in one publication (e.g., Malone et al., 2011). If distinct validation studies are performed for the same QRIS, we count each study separately (e.g., Barnard et al., 2006, and Sirinides, 2010; Norris, Dunn, and Eckert, 2003, and Norris and Dunn, 2004). We retain this counting convention throughout.

[4] We exclude the Boller et al. (2010) experimental evaluation of Washington state's Seeds for Success model as it is properly viewed as an impact evaluation of the professional development component of Seeds for Success, rather than a validation study of the QRIS itself.

Table 2. Features of the QRIS Validation Studies Reviewed

Study	Geographic Coverage	QRIS Name	Time Period	Settings	Questions Addressed
Zellman et al. (2008)	Colorado	Qualistar	2003–2007	Centers (I, T, P) FCCH	V1, V2, V3
Shen, Tackett, and Ma (2009)	Florida (Palm Beach County)	n.a.	2004–2007	Centers (I, T, P) FCCH	V2, V3
Malone et al. (2011)[a]	Florida (Miami-Dade County)	Quality Counts	2008–2010	Centers (I, T, P)	V1
Elicker et al. (2011)	Indiana	Paths to Quality (PTQ)	2008–2011	Centers (I, T, P) FCCH	V1, V2, V3
Lahti et al. (2011)	Maine	Quality for ME	2008	Centers (I, T, P, S) FCCH	V1
Tout et al. (2010) Tout et al. (2011)	Minnesota (Minneapolis, Saint Paul, Wayzata school district, Blue Earth County, and Nicollet County)	Parent Aware	2008–2011	Centers (I, T, P) FCCH	V1, V2, V3
Thornburg et al. (2009)	Missouri (Columbia, Kansas City, and St. Joseph)	Missouri Quality Rating System (pilot)	2008–2009	Centers (P) FCCH	V3
Bryant et al. (2001)	North Carolina	n.a.	1999	Centers (P)	V1
Norris, Dunn, and Eckert (2003)	Oklahoma	Reaching for the Stars	2001–2002	Centers (I, T, P, S)	V1, V2
Norris and Dunn (2004)	Oklahoma	Reaching for the Stars	2002	FCCH	V1
Barnard et al. (2006)	Pennsylvania	Keystone STARS	2006	Centers (P) FCCH	V1
Sirinides (2010)	Pennsylvania	Keystone STARS	2004–2009	Centers (I, T, P, S) FCCH	V1, V2, V3
Malone et al. (2011)[a]	Tennessee	Star-Quality Child Care Program	2008–2010	Centers (I, T, P)	V1
Sabol and Pianta (2012)	Virginia	Virginia Star Quality initiative	2007–2009	Centers (P)	V3

[a] Illinois is included in the study but results are not available for the validation questions included in Table 1.
NOTES: All studies are statewide unless otherwise noted. Question numbers refer to Table 1. Abbreviations: I = infants, T = toddlers, P = preschool age, S = school age. n.a. = not applicable.
SOURCE: Cited studies.

decade of the century, with the exception of North Carolina's QRIS validation dating back to 1999.

At a minimum, all of the state validation efforts have included center-based providers serving preschool-age children, the most prevalent age-group in nonparental care and the dominant setting at that age (Karoly, 2012). But most studies have expanded validation efforts to examine both center- and home-based programs in their analyses and specifically those center-based providers serving infants, toddlers, and preschool-age children. A few studies have also included center-based programs providing care to school-age children.

Across the 14 studies, all but three address the first validation question: is quality higher in more highly rated programs? Less common are those studies, six in total, that address the second validation question to determine if program ratings change over time. A total of seven studies concern the third validation question that asks if program ratings are consistent with child developmental outcomes. Such studies are less common, in part, because of the added cost involved when direct child assessments are performed.

3. Prior Studies Validating the Relationship Between QRIS Ratings and Program Quality or Validating Changes in Program Quality Over Time

Tables 3 and 4 provide more detail on the set of studies addressing validation questions V1 and V2, respectively. Both tables report on the ECE settings covered, sizes of the samples, and the findings. Table 3 details the measures of ECE quality that are used to compare with program ratings, while Table 4 notes the method used for examining quality changes over time. We begin by discussing the studies that examined the relationship between QRIS ratings and program quality (question V1) in Table 3.

Evaluations Examining QRIS Ratings and Program Quality

Eleven studies (reported in 12 publications) covering nine state or local-area rating systems have examined the relationship between QRIS program ratings and measures of ECE program quality. Typically these studies examine the average level of an "independent" measure of program quality for programs in each of the respective QRIS rating tiers. The expectation is that the average level of the independent quality measure will rise as programs move up on the QRIS rating scale. Some studies also examine the range of independently-derived quality scores within each rating tier to identify the amount of variability in quality among the programs in a given tier.

Settings and samples. Most of the V1 validation studies examine both center- and home-based programs, drawing distinctions between setting types that match with the state licensing system or correspond to the program types distinguished in the rating system quality standards. Typically studies employed samples of providers because of the expense associated with performing new quality assessments to compare with the QRIS ratings. For the most part, the

Table 3. Evaluations of QRIS Ratings and Program Quality

Study / Location / QRIS	Settings / Sample	Measure of Quality	Key Findings
Zellman et al. (2008) / Colorado / Qualistar	• 65 centers (Wave 1) • 38 FCCHs (Wave 1)	• ERS (ITERS, ECERS-R) • CIS* • Pre-K Snapshot subscales*	• QRIS ratings for centers were significantly positively related to two of the four CIS subscales (detachment and positive relationship) but not to any of the Pre-K subscales (Wave 1 data only) • QRIS ratings for FCCHs were not significantly related to the CIS or the Pre-K Snapshot subscales (Wave 1 data only)
Malone et al. (2011) / Florida (Miami-Dade County) / Quality Counts	• 253 licensed centers	• ERS (ITERS-R, ECERS-R)	• QRIS ratings were positively correlated with ERS
Elicker et al. (2011) / Indiana / Paths to Quality (PTQ)	• 135 classrooms in 95 licensed centers • 169 licensed FCCHs • 12 unlicensed registered child care ministries	• ERS* (ITERS-R, ECERS-R, FCCERS-R) • CIS*	• QRIS ratings were significantly positively associated with CIS and ERS scores—as scores increased, so did ratings • CIS and ERS overall and subscale scores for lowest rated providers (Level 1) were significantly different for the highest-rated providers (Level 4) • ERS scores were highly variable within each rating level for all QRIS levels and all types of care
Lahti et al. (2011) / Maine / Quality for ME	• 194 classrooms in 142 centers • 113 FCCHs	• ERS (ITERS-R, ECERS-R, SACERS, FCCERS-R)	• QRIS ratings were significantly positively correlated with ERS
Tout et al. (2010) / Minnesota (see Table 2 for sites) / Parent Aware	• 155 centers (ECERS-R, ECERS-E, CLASS) • 88 centers (ITERS-R) • 113 FCCHs	• ERS (ITERS-R, ECERS-R, ECERS-E*, FCCERS-R) • CLASS (centers only)	• Programs could receive a 4-star rating even with scores in the minimal range on the ERS and CLASS • There was some evidence that, at the 4-star level, programs tended to score better on observed quality measures than programs at other levels
Tout et al. (2011) / Minnesota (see Table 2 for sites) / Parent Aware	• 120 centers (ECERS-R, ECERS-E, CLASS) • 83 centers (ITERS-R) • 114 FCCHs	• ERS (ITERS-R, ECERS-R, ECERS-E*, FCCERS-R) • CLASS (centers only)	• ECERS-R scores for the 3- and 4-star fully-rated programs were significantly higher than those in 2-star programs • In all other cases, the scores across rating levels were not significantly different
Bryant et al. (2001) / North Carolina / n.a.	• 84 centers	• ERS (ECERS-R) • Teacher quality measures (education, wages, turnover)	• QRIS ratings were significantly positively correlated with ERS • The average teacher education and the average hourly wage were higher at centers with higher star levels; average annual turnover of teaching staff was lower at higher star levels

Table 3. Evaluations of QRIS Ratings and Program Quality, Continued

Study / Location / QRIS	Settings / Sample	Measure of Quality	Key Findings
Norris, Dunn and Eckert (2003) / Oklahoma / Reaching for the Stars	• 336 centers with at least one preschool room • Assessments for 279 I/T rooms, 336 preschool rooms, and 152 school-age rooms	• ERS* (ITERS, ECERS-R, SACERS) • CIS* • Other quality measures (director education, teacher education, turnover, salaries)	• Classroom ERS scores improved with each rating tier (4 total) with statistically significant differences in all pairwise tier comparisons made for ECERS-R and in 4 of 5 comparisons made for ITERS and SACERS • Classroom CIS scores improved with each rating tier but differences were statistically significant only for infant/toddler rooms • Most other quality measures (director education, teacher education, salaries) generally increased with each rating tier, with some pairwise tier comparisons being statistically significant • Turnover rates showed no relationship to rating tier
Norris and Dunn (2004) / Oklahoma / Reaching for the Stars	• 189 FCCHs	• ERS (FDCRS) • CIS*	• Two-Star FCCH providers had a higher ERS on average than either One-Star or One-Star Plus providers • Two-Star FCCH providers were more sensitive in their interactions with children than One-Star providers as measured by the CIS • Sample sizes were too small to analyze three-star (highest category) providers
Barnard et al. (2006) / Pennsylvania / Keystone-STARS	• 365 centers • 81 group child day care homes • 136 family child day care homes	• ERS (ECERS-R, FDCRS) • Other quality measures (teacher education, curriculum)	• QRIS ratings were positively correlated with ERS (significance not reported) • QRIS ratings for both centers and FCCHs were higher in those sites that used a defined curriculum and where teachers/caregivers had an associate's degree or higher
Sirinides (2010) / Pennsylvania / Keystone STARS	• Sample of QRIS-rated providers at all STAR levels for ERS administration (N not given) • Sample of 88 classrooms in STAR3 or STAR4 centers for CLASS administration	• ERS (ITERS–R, ECERS–R, SACERS, FCCERS–R) • CLASS*	• QRIS ratings were positively correlated with ERS (significance not reported) • Scores were higher for STAR 4 classrooms compared with STAR 3 classrooms on all CLASS subscales
Malone et al. (2011) / Tennessee / Star-Quality	• 1,369 licensed centers	• ERS (ITERS-R, ECERS-R)	• QRIS ratings were positively correlated with ERS

SOURCE: Cited studies.
NOTE: An asterisk denotes an independent measure of quality that was not included in the state or local QRIS rating scale.

programs sampled were already in the QRIS and had a rating designation that could be compared with the independent quality measures. For studies that performed original assessments, the combined samples across provider types ranged from just over 100 providers in Colorado (centers and FCCHs) to more than 300 providers in Oklahoma (centers only). Studies that relied solely on administrative data collected for the QRIS had access to larger samples (e.g., nearly 1,400 centers for Tennessee).

Measures of quality. For all studies, the independent quality assessments are performed at the program level in family child care homes (FCCHs) or at the classroom level in center settings. As evidenced in Table 3, every study collected the environment rating scale (ERS) corresponding to the program type (home or center) and, in the case of center-based programs, the age range of the children in the assessed classroom. Most used the most recent version of the Harms family of ERSs: Family Child Care Environment Rating Scale–Revised (FCCERS–R) (Harms, Cryer, and Clifford, 2007); Infant/Toddler Environment Rating Scale–Revised (ITERS–R) (Harms, Cryer, and Clifford, 2006); Early Childhood Environment Rating Scale–Revised (ECERS–R) (Harms, Clifford, and Cryer, 2005); and School-Age Care Environment Rating Scale (SACERS) (Harms, Jacobs, and White, 1995). Several earlier studies used the precursor version to the ITERS–R (the ITERS) or the FCCERS–R (the Family Day Care Rating Scale or FDCRS). One study also used the Early Childhood Environment Rating Scale–Extension (ECERS–E) (Sylva, Siraj-Blatchfor, and Taggart, 2010).

A common problem with the studies in Table 3 is that the ERS measure is typically one of the components in the rating system and often one that receives considerable weight. Thus, a positive correlation between the ERS and the rating tier is effectively an artifact of the rating system design. For this reason, most of the studies also assessed center classrooms or FCCHs using at least one other quality tool that was not embedded in the rating scale (measures marked with an asterisk in Table 3). In the case of the Indiana validation study, since the ERS is not used to determine program ratings, it served as an independent measure of program quality. For that study, the Caregiver Interaction Scale (CIS) (Arnett, 1989), a process-oriented measure of quality, was also used. The CIS was used for the Colorado and Oklahoma studies, as well. Other process measures used in other studies include the Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, and Hamre, 2008) employed in the Pennsylvania study; the ECERS–E measured in the Minnesota study; and the Pre-Kindergarten Snapshot (Ritchie et al., 2001) collected in the Colorado study. In addition, the studies for Oklahoma and Pennsylvania also examined the relationship between rating tiers and other structural measures of ECE program quality, such as director or teacher education level, staff turnover, staff salaries, and program curriculum, although some of these quality components are also used to generate program ratings.

Findings. Given the use of ERSs already embedded in all but one of the QRIS systems, it is not surprising that of the 11 studies that examined the relationship between the ERS and rating tiers, all but one found that they were positively correlated, although the relationship was not

always statistically significant. Minnesota was the exception, and the lack of a positive relationship between the ERS and ratings stemmed from the fact that some programs were automatically rated in the top tier without a formal quality assessment (Tout et al., 2010, 2011). Several studies noted that there was considerable variation in ERS scores within the tiers (e.g., Indiana, Minnesota). In other words, programs rated in the top tier could have a score on the ERS in the low range (scores under 3 on the 7-point scale). For those programs that examined independent process measures of quality such as the CIS and CLASS, positive correlations were also found, but again they were not always significant. Two studies, Colorado and Indiana, focused exclusively on the relationship between independent process measures and QRIS ratings. For the Colorado validation study, ratings were related to two of the four CIS subscales (detachment and positive relationship), but not to any of the Pre-Kindergarten Snapshot (Pre-K) subscales (Zellman et al., 2008). In the case of the Indiana study, there was a positive and significant correlation between the QRIS rating tiers and the ERS and CIS, and the contrast between level 1 and level 4 providers were often statistically significant (Elicker et al., 2010).

In sum, the studies listed in Table 3 provide evidence that the ratings in the QRISs examined are capturing meaningful differences in program quality, although the strength of the relationships were not as robust as QRIS designers would hope for. Moreover, the reliance in most studies on a measure of quality, namely the ERS, which is embedded in the rating scale, limits the weight that can be placed on the findings.

Evaluations of Changes in Program Ratings or Quality Indicators

The second validation question is addressed by the six studies (reported in seven publications) listed in Table 4 covering six QRISs. As a group, these studies ask whether ECE program ratings or other measures of program quality improve over time among those providers participating in the QRIS. Thus, providers outside the QRIS are not included in the analyses.

Settings and samples. With two exceptions, the studies examine changes over time in ratings for both home- and center-based programs. The sample sizes are determined in part by the number of programs participating in the QRIS and whether the analysis relies on QRIS administrative data or original data collection. The analysis for Pennsylvania, for example, is based on assessed centers and FCCHs in the system from 2004 to 2009, whereas the study for Oklahoma is based on 38 centers with ERS scores measured in 2002 and in an earlier validation study in 1999. Most of the other studies involve at most a few hundred providers.

Methods. Five of the studies use a longitudinal design to track the same programs through time (up to four points in time, typically one year apart) and examine the movement over time for individual programs, either in terms of program ratings (two studies) or ERS scores (three studies). The study for Indiana asked providers to self-report if they had a change in their rating in the last six months and the share with a change was reported for the cross-section of programs. The Pennsylvania analysis examined repeated cross-sections of ERS scores from administrative data for all rated providers to determine if program quality was increasing in aggregate. In this

Table 4. Evaluations of Program Ratings or Quality Indicators Over Time

Study / Location / QRIS	Settings / Sample	Methods	Key Findings
Zellman et al. (2008) / Colorado / Qualistar	43 centers with data for Waves 1, 2, and 3	Longitudinal ERS measures at three points in time (~ annually) for QRIS-rated providers	• Panel: Program quality, primarily the ECERS-R, increased over time for providers who were retained in the study
Shen, Tackett, and Ma (2009) / Florida (Palm Beach County) / n.a.	• For ITERS–R: 15 centers at baseline and T1; 7 centers at T1 and T2; • For ECERS–R: 40 centers at baseline and T1; 32 centers at T1 and T2; 19 centers at T2 and T3 • For FDCRS: 19 FCCHs at baseline and T1	Longitudinal ERS measures for up to four points in time for QRIS-rated providers	• Panel: ITERS-R scores improved from baseline to 13 months (all subscales), but not from 13 to 26 months (no 39-month follow-up) • Panel: ECERS-R scores improved from baseline to 13 months (all subscales) and from 13 to 26 months (4 of 7 subscales), but not from 26 to 39 months (no subscales) • Panel: FDCRS scores improved from baseline to 13 months (subscales not available and no later follow-up)
Elicker et al. (2011) / Indiana / Paths to Quality (PTQ)	• 90 licensed centers • 164 licensed FCCHs • 11 unlicensed registered child care ministries	Provider self-reports of QRIS rating change in past six months	• Cross-section: 24% of providers reported a change in the rating level in the past six months (22% advanced one or more levels, 2% dropped a level), while 71% of providers had remained at the same level (the other 5% had moved or closed)
Tout et al. (2010) / Minnesota (see Table 2 for sites) / Parent Aware	• 22 centers • 26 FCCHs	Longitudinal QRIS ratings at two points in time (~ annually) for QRIS-rated providers	• Panel: 65% of providers increased their ratings by at least one star between their first and second ratings
Tout et al. (2011) / Minnesota (see Table 2 for sites) / Parent Aware	• 40 centers • 57 FCCHs	Longitudinal QRIS ratings at two points in time (~ annually) for QRIS-rated providers	• Panel: 60% of centers and 70% of FCC providers increased their ratings by at least one star between their first and second ratings
Norris, Dunn, and Eckert (2003) / Oklahoma / Reaching for the Stars	38 centers with preschool classroom ERS scores in 1999 (earlier study) and 2002	Longitudinal ERS measures at two points in time for QRIS-rated providers	• Cross-section: ECERS-R scores were significantly higher in 2002 (6.2) than in 1999 (5.6) • Panel: Of the 20 programs not already at the top tier in 1999, 12 (60%) moved up at least one rating tier
Sirinides (2010) / Pennsylvania / Keystone STARS	All assessed centers and FCCHs from 2004 through 2009	ERS measures for six points in time for all QRIS-rated providers	• Cross-section: ERS scores (ITERS–R, ECERS–R, SACERS, FCCERS–R) have been steadily increasing from 2004 to 2010

SOURCE: Cited studies.

case, the movement of individual providers through the rating tiers was not tracked. One issue with the cross-sectional design is that it is not possible to tell if an observed increase in quality over time results from quality improvements for programs within the system or from a shift in the composition of providers toward those with higher quality as a result of providers entering and leaving the system over time.

Findings. Across the six studies in Table 4, a consistent finding is that quality—as measured by the summary rating accounting for all quality components or based solely on the ERS, typically a heavily-weighted component in the QRIS—increased over time among participating providers. For the providers in Indiana, about one in five had advanced at least one level in the prior six months, while a few had dropped to a lower level (Elicker et al., 2011). In Minnesota, 60 to 65 percent of providers had advanced at least one level between ratings, typically about one year apart (Tout et al., 2010, 2011). Over the three-year horizon examined for Oklahoma, about 60 percent of providers not already in the top tier were able to advance one or more levels (Norris, Dunn, and Eckert, 2003). At the same time, the analysis for the centers and FCCHs in Palm Beach County, Florida suggests that providers may improve quality in the first or second year after their initial rating but then further movement may taper off (Shen, Tackett, and Ma, 2009).

It is important to acknowledge that these studies are not able to evaluate the impact of the QRIS on ECE program quality or ratings. In the absence of a control or comparison group, it is not possible to determine what would have happened to program quality in the absence of participating in the QRIS. Thus, the studies should not be interpreted as demonstrating that a given QRIS as a whole or its specific components, such as TA activities, produced the observed changes in quality. Another challenge with the studies that track providers over time is the impact of attrition on the inferences made. For example, in the Colorado evaluation, the finding of improved quality over time is potentially compromised by the fact that lower performing centers were more likely to leave the study before the final wave of data collection, so the reported correlations are based on the higher-quality providers that were retained in the sample (Zellman et al., 2008).

4. Prior Studies Validating the Relationship Between QRIS Ratings and Child Developmental Outcomes

Although a central tenant in the QRIS logic model is that program ratings capture meaningful differences in quality that are predictive of child developmental outcomes (Zellman and Perlman, 2008), there have been only a few studies that empirically test that assumption. Table 5 lists the seven studies to date (reported in eight publications) covering seven QRISs that have attempted to address this validation question. As illustrated in the table, these studies differ considerably in terms of the care settings examined, the size of the analysis samples, the analytic methods

Table 5. Evaluations of QRIS Ratings and Child Developmental Outcomes

Study / Location / QRIS	Settings / Sample	Methods	Measures of Child Development	Key Findings
a. Cross-Sectional Designs				
Shen, Tackett, and Ma (2009) / Florida (Palm Beach County) / n.a.	87 QRIS-rated sites in Florida Voluntary Pre-Kindergarten (VPK) program and 88 non-QRIS sites in VPK	• Cross-sectional • Site level (aggregated from child-level records) • Controls for site characteristics • Comparison to non-QRIS rated sites • Administrative data	Teacher administered Florida Kindergarten Readiness Screener which includes: • 19 items from Early Childhood Observation System (ECHOS) • Letter Naming Fluency (LNF) from Dynamic Indicators of Basic Early Literacy Skills (DIBELS) • Initial Sound Fluency (ISF) from Dynamic Indicators of Basic Early Literacy Skills (DIBELS)	• QRIS ratings were found to be positively and significantly associated with the school readiness assessment and its components • Over time, the rate of growth of school readiness rates was higher for QRIS sites compared with non-QRIS sites, but not significantly so
Elicker et al. (2011) / Indiana / Paths to Quality (PTQ)	557 children in QRIS-rated centers and FCCHs (2 per classroom or home) • 249 children ages 6 to 35 months • 308 children ages 36 to 60 months	• Cross-sectional • Child level • Family background controls (parent survey) • Primary data	Independent assessment • I/T: Mullen Scales of Early Learning • I/T: Brief Infant Toddler Social and Emotional Assessment • P: PPVT–4 • P: WJ–III Letter Word Identification • P: WJ–III Applied Problems • P: Social Competence and Behavior Evaluation	• Infant-toddler developmental assessments were not significantly related to QRIS tier or type of care, even when controlling for parental education and household income; associations were in the expected direction • With the exception of anxiety/withdrawal behaviors, developmental assessments for preschool-age children were not significantly related to QRIS tier, even when controlling for parental education and household income • There was a positive correlation between some of the subscales of the independent quality measures (ERS and CIS) for some of the measures of infant-toddler and preschool development
Sirinides (2010) / Pennsylvania / Keystone STARS	8,464 (fall) and 9,268 (spring) preschool-age children enrolled in STAR 3 and STAR 4 centers	• Repeated cross-section (fall and spring) • Aggregated child-level data • Administrative data	Teacher assessment of child development (not yet, in process, proficient) • Work Sampling System • Ounce Scale System	• The percentage of children scoring “proficient” according to teacher ratings was significantly higher in the spring than in the fall in seven developmental domains: Personal and Social Development, Language and Literacy, Mathematical Thinking, Scientific Thinking, Social Studies, The Arts, and Physical Development and Health • The percentage of “proficient” children was greater for STAR 4 participants than STAR 3 participants in the spring on all of the above measures (statistical significance not reported, change scores not reported)

Table 5. Evaluations of QRIS Ratings and Child Developmental Outcomes, Continued

Study / Location / QRIS	Settings / Sample	Methods	Measures of Child Development	Key Findings
b. Longitudinal Designs				
Zellman et al. (2008) / Colorado / Qualistar	1,368 preschool-age children enrolled in QRIS-rated centers or FCCHs in Wave 1; 829 children in Wave 2; 619 children in Wave 3	• Longitudinal (3 points in time) • Family background controls (parent survey) • Primary data	Independent assessment • PPVT–4 • WJ–III Letter Word Identification • WJ–III Passage Comprehension • WJ–III Applied Problems Teacher assessment • Child Behavior Inventory (CBI) Parent assessment • Strength and Difficulties Questionnaire (SDQ) (Wave 3 only)	• QRIS ratings were not associated with improvement in child outcomes for either centers or FCCHs • Individual components of the QRIS ratings (e.g., average class ratio, parent survey, head teacher educational attainment) were not associated with any improvement in child outcomes • Subgroup analyses did not show that low-income children were more likely to benefit from highly rated centers
Tout et al. (2010) / Minnesota (see Table 2 for sites) / Parent Aware	421 preschool-age children in two cohorts (2008-2009 and 2009-2010) enrolled in 84 QRIS-rated centers or FCCHs	• Longitudinal (fall to spring) • Child level • Family background controls (parent survey) • Primary data	Independent assessment • PPVT–4 • Individual Growth and Development Indicators (IGDI) Picture Naming • Test of Preschool Early Literacy (TOPEL) Phonological Awareness and Print Knowledge • WJ–III Applied Problems • WJ–III Quantitative Concepts Teacher assessment • Social Competence and Behavior Evaluation short form (SCBE-30) • Preschool Learning and Behavior Scale (PLBS) Persistence subscale	• There were no definitive patterns of linkages between quality rating categories and children's developmental gains • Only two statistically significant effects in the expected direction were found for components of the QRIS (Parent Aware): Tracking Learning predicted PPVT change scores and Teacher Training and Education predicted WJ–III Quantitative Concepts change scores • For some measures, Parent Aware subscale scores negatively predicted child outcomes

Table 5. Evaluations of QRIS Ratings and Child Developmental Outcomes, Continued

Study / Location / QRIS	Settings / Sample	Methods	Measures of Child Development	Key Findings
Tout et al. (2011) / Minnesota (see Table 2 for sites) / Parent Aware	701 preschool-age children in three cohorts (2008-2009, 2009-2010, and 2010-2011) enrolled in 138 QRIS-rated centers or FCCHs	• Longitudinal (fall to spring) • Child level • Family background controls (parent survey) • Primary data	Independent assessment • PPVT–4 • Individual Growth and Development Indicators (IGDI) Picture Naming • Test of Preschool Early Literacy (TOPEL) Phonological Awareness and Print Knowledge • WJ–III Applied Problems • WJ–III Quantitative Concepts Teacher assessment • Social Competence and Behavior Evaluation short form (SCBE-30) • Preschool Learning and Behavior Scale (PLBS) Persistence subscal	• Children overall and children in poverty in programs at different quality rating levels did not differ systematically from each other in their developmental gains from fall to spring • There was some evidence for differences in children's receptive vocabulary (PPVT) across star levels, but these findings were not robust to variations in models
Thornburg et al. (2009) / Missouri (see Table 2 for sites) / Missouri Quality Rating System	350 preschool-age children in 66 classrooms enrolled full-time (25+ hours) in 32 licensed centers and 6 licensed FCC homes (excluded non-English speakers and those with severe disabilities)	• Longitudinal (fall to spring) • Child level • Family background controls (parent survey) • Primary data	Independent assessment • PPVT–4 • TERA–3 Reading Quotient • TERA–3 Alphabet subtest • TERA–3 Conventions subtest • TERA–3 Meaning subtest • WJ–III Applied Problems • Shape identification • Color identification • Uppercase alphabet • Fine motor • Gross motor • DECA Total Protective Factors • DECA Initiative scale • DECA Self-control scale • DECA Attachment scale • DECA Behavioral Concerns	For all children by rating tier, statistically significant greater gains were found for the following outcomes (effect sizes in parentheses): • High (4-5 stars) versus low (1-2 stars): overall social and behavioral skills (0.80), motivation (0.79), self-control (0.65), and positive adult relationships (0.45) • Medium (3 stars) versus low (1-2 stars): overall social and behavioral skills (0.36) and motivation (0.43) For children in poverty by rating tier, statistically significant greater gains were found for the following: • High versus low: overall social and behavioral skills (0.79), motivation (0.78), and vocabulary (0.74) • Medium versus low: vocabulary (0.64) • High versus medium: self-control (0.61) For children not in poverty by rating tier, statistically significant greater gains were found for the following: • High versus low: overall social and behavioral skills (0.79), motivation (0.79), and self-control (0.66) • Medium versus low: overall social and behavioral skills (0.49), motivation (0.57), and positive adult relationships (0.33)

Table 5. Evaluations of QRIS Ratings and Child Developmental Outcomes, Continued

Study / Location / QRIS	Settings / Sample	Methods	Measures of Child Development	Key Findings
Sabol and Pianta (2012) / Virginia / Virginia Star Quality Initiative	2,805 preschool-age children in 71 QRIS rated state-funded pre-kindergarten programs	• Longitudinal (fall to spring in preK and K year) • Child level • Family background controls (child record) • Center and community characteristics controls (or fixed effects) • Primary data	PreK and K teacher assessment of pre-literacy skills • Phonological Awareness Literacy Screening (PALS) Pre-K (seven subtests; used to derive two factors: Alphabet Knowledge and Phonological Awareness) • Phonological Awareness Literacy Screening (PALS) K (seven subtests; used to derive two factors: Alphabet Knowledge and Phonological Awareness)	• There was no correlation between preK star levels and fall K pre-literacy skills after controlling for preK fall pre-literacy skills, family background, center characteristics, and community characteristics • Using the same controls, the growth in Alphabet Knowledge during the preK year was significantly higher for children in 3-star programs versus 2-star programs (effect size of 0.43) and in 4-star programs versus 2-star programs (0.40); the growth in Phonological Awareness in the preK year was significantly higher only for children in 3-star programs versus 2-star programs (0.37) • Using the same controls, compared to 2-star programs, children in 3-star and 4-star programs had significantly higher declines in Alphabet Knowledge between the spring preK and fall K assessments (effect sizes of -0.12 and -0.18, respectively) • Using the same controls, there was no difference in fall-spring growth during the K year by preK star rating

SOURCE: Cited studies.

employed, the child developmental measures assessed and method of assessment, and the number of time periods over which children were assessed. The table differentiates between two general study approaches: those studies employing a cross-sectional methodology (either a single cross-section or a repeated cross-section design) are listed first in panel (a) (three studies), while those studies using a longitudinal design are listed second in panel (b) (the remaining four studies).

Settings and Samples

As with the studies previously summarized in Tables 3 and 4, the studies that examine child outcomes typically analyzed outcomes for children in both home- and center-settings, although three studies (Florida, Pennsylvania, and Virginia) were limited to center-based providers. With two exceptions, the studies in Table 5 examine the relationship between quality ratings and child outcomes for children enrolled in programs that are already part of the QRIS. The Missouri study was a pilot study performed prior to statewide implementation of the QRIS, so the sample of providers were drawn from ECE programs in three communities that had rated programs. The Florida analysis used data aggregated to the site level and included both QRIS-rated sites and non-QRIS-rated sites for some analyses.

There is considerable variation in the size of the samples, largely driven by whether children were assessed by independent, trained assessors (four studies), as compared with developmental assessments exclusively performed by teachers (e.g., a preschool developmental assessment or kindergarten entry assessment performed by the child's teacher) (three studies). For example, the Pennsylvania study relied on teacher assessments of children's development performed three times per year to inform teaching practice, with assessment results recorded as part of a state online database system (Sirinides, 2010). With such administrative data, the samples ranged from about 8,500 to 9,300 children at the two assessment points. The Virginia study, which also relied on teacher assessments in the fall and spring of both preschool and kindergarten years, had a sample size of nearly 3,000 children (Sabol and Pianta, 2012). Among the four studies that performed the more costly direct child assessments, the largest sample occurred in the first wave of the Colorado study (more than 1,300 children), while the smallest was for the Missouri validation study (350 children).

Methods

To assess whether ECE program ratings are predictive of child development, the ideal study design would randomly assign children to programs of varying levels of rated quality (i.e., star 1 programs, star 2 programs, etc.) at the start of the program year and then measure domains of child development after a sufficient time has passed. If the program ratings capture meaningful differences in quality, we would expect to see higher levels of child development on average for programs at each successive rating tier. In the absence of such random assignment, if we simply

examine the average level of child development for programs in different rating tiers at a point in time, it is possible that the differences we observe are at least in part the result of selectivity, i.e., parents choose to enroll their children in programs based on their characteristics including dimensions of quality. If parents with more resources choose higher quality programs, then the higher levels of development for the children in those programs is likely attributable to some combination of family background factors and the impact of the program itself. Without accounting for the impact of selectivity when using observational data, estimates of the relationship between program quality and child outcomes will be biased.

Since no studies to date have had the option of random assignment, researchers have employed various research designs to try to mitigate the potential for selectivity bias. One option is to include child and family background covariates, measured through a parent survey or through administrative program data (e.g., eligibility for free or reduced-price lunch), to control for observed factors that may influence the selectivity of children into ECE programs. This approach is taken in five of the studies in Table 5 (one in panel (a) and all four in panel (b)). In one case, the Virginia analysis, controls for program and community characteristics are included as well (Sabol and Pianta, 2012). A second option is to use a pretest–posttest design, where gains in child development are calculated from a baseline (pretest) to a follow-up wave (posttest). The approach controls for differential levels of development at the baseline. This approach is adopted for the four studies in panel (b) of Table 5. Controls for family background in such a longitudinal analysis, as is done for all of the panel (b) studies, may further diminish any selectivity bias. For this reason, these four studies—Colorado, Minnesota (two publications), Missouri, and Virginia— have the strongest design for assessing the relationship between QRIS rating tier and child developmental outcomes.

Of the three cross-sectional analyses in panel (a), the study for Indiana uses a single cross-section, although it does include family background controls. The study for Pennsylvania has child assessment data in the fall and spring, but the data are analyzed as a repeated cross-section, rather than as a longitudinal design. Finally, the study for Florida is a cross-sectional comparison at the site-level based on aggregated child-level data. The study does compare child outcomes in QRIS rated and non-rated sites, an approach more relevant for potentially measuring the impact of participation in the QRIS.

Measures of Child Development

For the most part, the studies in Table 5 recognize that there are multiple dimensions of child development and therefore more than one assessment tool is required to capture cognitive, social, emotional, behavioral, and physical development. One basic difference is whether direct child assessments are performed by trained, reliable independent assessors or whether developmental ratings are made by parents or teachers who may vary in how they rate children against a given benchmark (i.e., low inter-rater reliability). Two of the studies in panel (a) rely on solely on teacher-administered assessments, namely the observational assessments performed throughout

the year by preschool teachers in the case of the Pennsylvania study and a kindergarten readiness screener performed by kindergarten teachers in the case of the Florida study. One of the panel (b) studies, the Virginia validation study also relies on teacher assessments in both the preschool and kindergarten years. All of the other studies incorporate at least some direct child assessments by trained observers. For the Colorado and two Minnesota studies, social and behavioral assessments performed by teachers or teachers and parents are used in addition to the independent assessments.

There are some commonalities among the studies in the direct assessments that are employed. For cognitive domains for preschool-age children, the following are used by two or more studies:

- Woodcock-Johnson III Tests of Achievement (WJ–III), subtests for Letter Word Identification, Passage Comprehension, Applied Problems, and Quantitative Concepts (Woodcock, McGrew & Mather, 2001); and
- Peabody Picture Vocabulary Test–Fourth Edition (PPVT–4) (Dunn & Dunn, 2007).

A range of other well-validated cognitive assessments are used by just one study listed in Table 5, either for measuring development in preschool-age children or development for infants and toddlers. For the social and behavioral measures, the Social Competence and Behavior Evaluation (SCBE or short form SCBE-30) (LaFreniere and Dumas, 1996a, 1996b) is the only measure used by more than one study.

Findings

As a whole, the studies in Table 8 provide only limited evidence that programs rated more highly in a given QRIS are associated with larger developmental gains for the enrolled children. Among the cross-sectional studies in panel (a), two of the three studies, those for Florida and Pennsylvania, found that child development was positively correlated with program ratings. But these studies had weaker designs that relied solely on teacher assessments rather than independent assessors. The study for Indiana, which relied on independent assessments, employed child-level data, and had family background controls, found associations in the expected direction but none that were statistically significant (with one exception that may possibly be due to chance).

Among the three longitudinal studies in panel (b) with the stronger designs in terms of pretest–posttest comparisons, controls for family background, and use of multiple well-validated independent assessments, just one of the studies found the hypothesized relationship between quality ratings and child outcomes. The study of Missouri's QRIS, conducted during the pilot phase, reported that more highly rated programs are associated with better developmental outcomes, although only for a limited set of the social-emotional measures (Thornburg et al. 2009). Notably, none of the nine measures of early reading or quantitative skills showed a relationship with rated program quality for the full sample. When the analysis was stratified by

child poverty status, the PPVT–4 showed the expected relationship with the quality tier for the subset of children in poverty. In contrast, the study for Minnesota's Parent Aware QRIS (Tout et al., 2010, 2011) and the study for the Colorado Qualistar QRIS (Zellman et al., 2008) showed that, as configured, those rating systems did not generate quality ratings that distinguished programs in terms of the developmental gains experienced by participating children.

The final longitudinal study in panel (b) for Virginia—while relying solely on a teacher-performed assessment of pre-literacy skills in two domains (Alphabet Knowledge and Phonological Awareness) for the set of state-funded prekindergarten programs in the QRIS—included a potentially richer set of control variables measured at the child, center, and community level. This study of Virginia's QRIS demonstrated significantly higher gains during the prekindergarten year for four-star versus two-star programs and three-star versus two-star programs for one or both of the pre-literacy measures. At the same time, there was no indication that program quality as rated by the QRIS was associated with subsequent performance on the literacy measures during the kindergarten year.

Here again, it is important to note that the studies summarized in Table 5 are not designed to test a causal link between participating in higher quality ECE programs and child developmental outcomes. In the absence of random assignment of children to programs with varying levels of quality, it is not possible to fully control for the potential selectivity bias given that families choose where to enroll their children. Likewise, without an experimental design or a rigorous quasi-experimental design, these studies do not provide evidence of a causal link between the implementation of a QRIS and child developmental outcomes. Such an impact study has not been published to date, to our knowledge.

At the same time, validation studies that address the third question in Table 1 (V3) are useful in demonstrating whether or not the quality ratings defined in a given QRIS distinguish between programs that are associated with smaller gains in child development (rated as low quality) from those that are associated with larger gains in child development (rated as high quality). The mixed findings reported in Table 5 suggest that it cannot be assumed that QRISs will automatically be successful in classifying programs in this way. Among the QRISs evaluated in Table 5 with a sample of both center- and home-based providers, only the Missouri QRIS as piloted (and which has not been subsequently implemented at scale) showed that the ratings generated could distinguish programs with smaller versus larger child developmental gains, and then only for a subset of developmental domains. The Virginia QRIS also showed some ability to distinguish higher-rated programs with larger gains in pre-literacy skills from lower-rated program with smaller gains, although this analysis was based on teacher assessments performed in a set of center-based pre-kindergarten programs. These mixed findings may also reflect the challenge of adequately controlling for unobserved family background factors that confound our ability to measure the relationship between program ratings and child outcomes.

5. Conclusions

QRIS validation is a complex endeavor that involves an array of potential research questions. Until recently, there has been little guidance regarding best practices for conducting such research (Lugo-Gil et al., 2011; Zellman and Fiene, 2012). This means that there has been considerable variation across studies to date in the samples of ECE programs analyzed, the measures of ECE program quality and child development collected, and the analytic methods employed. Our review and synthesis of the prior literature has produced a number of findings:

- Although QRISs are being designed or implemented in almost every state, the time, expense, and challenges of validation research means that, to date, published studies are available for just 12 state- or local-level QRISs.
- A natural starting point for QRIS validation research is to assess whether higher rated programs indeed have higher observed quality, ideally using one or more measures of quality that are independent of the measures incorporated in the rating scale. We identified 11 such studies. Most studies found that programs with higher ratings had higher ERS scores, but this is not surprising because the ERS is usually one of the rating elements. Independent measures of quality did not always show the expected positive relationship with quality.
- We reviewed six studies that assessed whether program ratings or other summary measures of quality like the ERS improved over time, a second type of validation study. These studies generally showed that programs participating in the QRIS did improve their quality or quality ratings over time. For one study, a closer look at the patterns over time showed that the quality gains for individual programs did not persist past the first year or two after entering the QRIS.
- A subset of seven studies examined the relationship between QRIS ratings and child development, a critical third type of validation exercise. These studies are the most challenging to implement and can be costly to conduct when independent child assessments are performed. Consequently, there has been considerable variation in methods to date across these studies. Among the four studies with the stronger designs, two found the expected relationship between QRIS ratings and child developmental gains. In one case, the significant effects were primarily for measures of social and behavioral development. In the other case, the results were for a teacher-performed assessment of pre-literacy skills. The lack of robust findings across these studies indicate that QRISs, as currently configured, do not necessarily capture differences in program quality that are predictive of gains in key developmental domains.

With the requirement for QRIS validation in the Race to the Top–Early Learning Challenge grants, there is an opportunity to advance the methods used to validate state QRISs and contribute not only to improvement of such systems in any given state, but also to add to the knowledge based about effective systems more generally. Advances over prior research would include the following:

- Where possible, incorporate analysis of both home- and center-based settings with sufficient sample sizes to ensure adequate statistical power when the data are pooled and when stratified by type of setting;
- Where possible, include ECE programs both participating in and not participating in the QRIS to ensure that the full range of program quality in the community is captured in the validation study;
- To test if programs with higher ratings have higher observed quality, use measures of ECE program quality that are not already embedded in the rating system and that have been causally linked to child developmental outcomes in prior research;
- To test the relationship between QRIS ratings and child developmental outcomes, at a minimum employ a pretest–posttest design and control for as many observed variables as possible at the child, family, program, or community level;
- When measuring child developmental outcomes, include assessments from a range of developmental domains, preferable those that can be reliably measured.

References

American Institutes for Research (AIR) and RAND Corporation. 2013. *Local Quality Improvement Efforts and Outcomes Descriptive Study*. San Mateo, CA and Santa Monica, CA: AIR and RAND Corporation.

Arnett, J. 1989. "Caregivers in Day-Care Centers: Does Training Matter?" *Journal of Applied Developmental Psychology* 10(4): 541–552.

Barnard, W., W. E. Smith, and others. 2006. *Evaluation of Pennsylvania's Keystone STARS Quality Rating System in Child Care Settings.* University of Pittsburgh Office of Child Development and the Pennsylvania State University Prevention Research Center. As of October 1, 2013: http://www.ocdelresearch.org/Reports/Keystone%20STARS/Keystone%20STARS%202010%20Evaluation%20Report.pdf

Boller, K., P. Del Grosso, and others. 2010. *The Seeds of Success Modified Field Test: Findings from the Impact and Implementation Studies*. Princeton, NJ: Mathematica Policy Research, Inc. As of October 1, 2013: http://www.mathematica-mpr.com/publications/PDFs/earlychildhood/seeds_to_success_mft.pdf

Bryant, D. M., K. Bernier, and others. 2001. *Validating North Carolina's 5-star Child Care Licensing System.* Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Center. As of October 1, 2013: http://fpg.unc.edu/resources/validating-north-carolinas-5-star-child-care-licensing-system

Dunn, L. M. and D. M. Dunn. 2007. *Peabody Picture Vocabulary Test-Fourth Edition*. Minneapolis, MN: Pearson.

Elicker, J., and K. Thornburg. 2011. *Evaluation of Quality Rating and Improvement Systems for Early Childhood Programs and School-age Care: Measuring Children's Development*. Research-to-Policy, Research-to-Practice Brief. Washington, DC: OPRE.

Elicker, J., C. C. Langill, and others. 2011. *Evaluation of Paths to QUALITY, Indiana's Child Care Quality Rating and Improvement System: Final Report*. West Lafayette, IN: Purdue University. As of October 1, 2013: http://www.cfs.purdue.edu/cff/documents/project_reports/PTQFinalReportRev11012.pdf

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

Harms, T., D. Cryer, and R. M. Clifford. 2006. *Infant/Toddler Environment Rating Scale–Revised Edition (ITERS–R), Updated*. New York: Teachers College Press.

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

Harms, T., E. V. Jacobs, and D. R. White. 1995. *School-Age Care Environment Rating Scale (SACERS*). New York: Teachers College Press.

Karoly, L. A. 2012. *The Use of Early Care and Education by California Families*. OP-356, Santa Monica, CA: RAND Corporation.

LaFreniere, P. J. and J. E. Dumas. 1996a. *Social Competence and Behavior Evaluation, Preschool Edition (SCBE)*. Torrance, CA: Western Psychological Services.

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

Lahti, M., C. Cobo-Lewis, A. and others. 2011. *Maine's Quality For Me – Child Care Quality Rating And Improvement System (QRIS): Final Evaluation Report.* Maine: Department of Health and Human Services. As of October 1, 2013: http://muskie.usm.maine.edu/maineroads/pdfs/QRISEVALRPRT_FINAL.pdf

Lahti, M., C. Sabol, T., and others. 2013. *Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States*. Research-to-Policy, Research-to- Practice Brief. OPRE 2013-06. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

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

Malone, L., G. Kirby, and others. 2011. *Measuring Quality Across Three Child Care Quality and Improvement Systems: Findings from Secondary Analyses.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. As of October 1, 2013: http://qrisnetwork.org/sites/all/files/resources/gscobb/2011-09-28%2014:15/Report.pdf

Norris, D. J., and L. Dunn. 2004. *Reaching for the Stars: Family Child Care Home Validation Study Final Report.* Stillwater, OK, and Norman, OK: Early Childhood Collaborative of Oklahoma. As of October 1, 2013: http://www.okdhs.org/NR/rdonlyres/11E190FA-7FF2-4166-8F6E-

19ED34FE4E0F/0/ReachingForTheStarsFamilyChildCareValidationStudyFinalReport_dcc_05092007.pdf

Norris, D. J., L. Dunn, and L. Eckert. 2003. *Reaching for the Stars: Center Validation Study Final Report,* Stillwater, OK, and Norman, OK: Early Childhood Collaborative of Oklahoma. As of October 1, 2013: http://www.okdhs.org/NR/rdonlyres/4C0EF19D-6FC4-40C6-8926-A3371B7F4130/0/ReachingForTheStarsCenterValidationStudyFinalReport_dcc_05212007.pdf

Pianta, R. C., K. M. La Paro, and B. K. Hamre. 2008. *Classroom Assessment Scoring System (CLASS) Manual, Pre-K*, Baltimore, MD: Paul H. Brookes Pub. Co.

Ritchie, S., B. Weiser, M. Kraft-Sayre, and C. Howes. 2001. Emergent Academics Snapshot Scale. Los Angeles: UCLA (unpublished instrument).

Sabol, T., and R. Pianta. 2012. *Improving Child Care Quality: A Validation Study of the Virginia Star Quality Initiative*. Charlottesville, VA: University of Virginia Curry School of Education.

Schilder, D. 2013. *Quality Rating and Improvement System (QRIS) Validation Study Designs.* CEELO FASTfacts, New Brunswick, NJ: Center on Enhancing Early Learning Outcomes.

Shen, J., W. Tackett, and X. Ma. 2009. *Second Evaluation Report For Palm Beach County Quality Improvement System*. Palm Beach, CA: Children's Services Council of Palm Beach County. As of October 1, 2013: http://cache.trustedpartner.com/docs/library/000238/Download%20Second%20Evaluation%20Report%20for%20Palm%20Beach%20County%20Quality%20Improvement%20Center.pdf

Sirinides, P. 2010. *Demonstrating Quality: Pennsylvania Keystone STARS: 2010 Program Report*. Harrisburg, PA: Office of Child Development and Early Learning. As of October 1, 2013: http://www.ocdelresearch.org/Reports/Keystone%20STARS/Keystone%20STARS%202010%20Evaluation%20Report.pdf

Sylva, K., I. Siraj-Blatchford, and B. Taggart. 2010. *ECERS–E: The Four Curricular Subscales Extension to the Early Childhood Environment Rating Scale (ECERS), Fourth Edition*. New York: Teachers College Press.

Thornburg, K. R., W. A. Mayfield, and others. 2009. *The Missouri Quality Rating System School Readiness Study.* Columbia, MO: Center for Family Policy & Research. As of October 1, 2013: http://mucenter.missouri.edu/MOQRSreport.pdf

Thornburg, K., D. Mauzy, and others. 2011. "Data-driven Decisionmaking in Preparation for Large Scale QRS Implementation," in *Quality measurement in early childhood settings*. Edited by M. Zaslow, and others. Baltimore: Paul H. Brookes Publishing Co.

Tout, K., R. Starr, and others. 2010. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Year 3 Evaluation Repor*t, Minneapolis, MN: Child Trends. As of October 1, 2013: https://s3.amazonaws.com/Omnera/VerV/s3finder/38/pdf/Parent_Aware_Year_3_Evaluation_Report_Nov_2010.pdf

Tout, K., R. Starr, and others. 2011. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Final Evaluation Report*. Minneapolis, MN: Child Trends. As of October 1, 2013: https://s3.amazonaws.com/Omnera/VerV/s3finder/38/pdf/Parent_Aware_Year_4_Final_Evaluation_Technical_Report_Dec_2011.pdf

Woodcock-Johnson III Tests of Achievement (WJ–III), subtests for Letter Word Identification, Passage Comprehension, Applied Problems, and Quantitative Concepts (Woodcock, McGrew & Mather, 2001);

Zellman, G.L., R.N. Brandon, and others. 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., and R. Fiene. 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. and L. A. Karoly. 2012. *Moving to Outcomes: Approaches to Incorporating Child Assessments into State Early Childhood Quality Rating and Improvement Systems*. OP-364. Santa Monica, CA: RAND Corporation.

Zellman, G. L., and M. Perlman. 2008. *Child-Care Quality Rating and Improvement Systems in Five Pioneer States. Implementation Issues and Lessons Learned*. Santa Monica, CA: RAND Corporation.

Zellman, G. L., M. Perlman, and others. 2008. *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child-Care Quality*. MG-650. Santa Monica, CA: RAND Corporation.