Independent Evaluation of California's Race to the Top–Early Learning Challenge Quality Rating and Improvement System

Half-Term Report

Submitted to:

California Department of Education
Early Education and Support Division

Submitted by:

American Institutes for Research
RAND Corporation

August 2015

Independent Evaluation of California's Race to the Top–Early Learning Challenge Quality Rating and Improvement System:

Half-Term Report

August 2015

Project Leadership:

Heather E. Quick, *Project Manager*
Aleksandra Holod, *Deputy Project Manager*
Susan Muenchow, *Senior Advisor*
Deborah Parrish, *Senior Advisor*
Laura E. Hawkinson, *Analysis Lead*
Jill S. Cannon, *RAND Project Manager*
Susannah Faxon-Mills, *RAND Deputy Project Manager*
Lynn A. Karoly, RAND *Senior Advisor*
Gail L. Zellman, RAND *Senior Advisor*

Report Authors:

AIR team: Laura E. Hawkinson, Heather E. Quick, Susan Muenchow, Jennifer Anthony, Emily Weinberg, Aleksandra Holod, Deborah Parrish, John Meakin, Dong Hoon Lee, and Kate Tarrant

RAND team: Jill S. Cannon, Gail L. Zellman, and Lynn A. Karoly

AMERICAN INSTITUTES FOR RESEARCH®

2800 Campus Drive, Suite 200
San Mateo, CA 94403
650.843.8100 | TTY 877.334.3499

www.air.org

RAND CORPORATION

1776 Main Street
Santa Monica, CA 90401-3208
310.393.0411

www.rand.org

1950_041950_04/15

Contents

Executive Summary i
Introduction i
Overview of the Study iii
What Is the Status of RTT-ELC Implementation? v
How Well Does the QRIS Define Quality? vii
How Well Does the QRIS Perform As a Measure of Quality? viii
How Well Does the QRIS Differentiate Between Observed Quality of Programs? ix
How Do Alternative Rating Approaches Affect the Distribution and Validity of Ratings? x
Policy Options for Consideration xii
Other Considerations Relevant for Further Expansion of the System and Its Validation xiv
Acknowledgments xvi
Chapter 1. Introduction 1
Background on QRISs 1
California's RTT-ELC QRIS 5
The Evaluation of California's RTT-ELC QRIS 5
Organization of this Report 10
Chapter 2. Snapshot of RTT-ELC QRIS Implementation 12
Status of Implementation 12
Classroom Observations 15
Local Variation in the Implementation of the Hybrid Rating Matrix 21
Summary 24
Chapter 3. Content Analysis 26
The Research Base for the Elements in the Hybrid Rating Matrix and Other Indicators of Quality 26
The Research Base for the Program Quality Assessment Tools in the Hybrid Rating Matrix and Other Assessment Tools 43
Rating Structure Analysis 53
Summary and Possible Policy Considerations 55
Chapter 4. Measurement Properties of QRIS Ratings 57
Distribution of Ratings and Element Scores 58
Characteristics of Programs That Predict QRIS Ratings 64
Internal Consistency 69
How Element Scores Relate to Each Other and the Overall Rating 70

Summary: How Well Does the QRIS Perform As a Measure of Quality? 73
Chapter 5. Concurrent Validity 75
Concurrent Validity of California QRIS Ratings 75
Concurrent Validity of Element Scores 81
Summary: How Well Does the QRIS Differentiate Between Observed Quality of Programs? 86
Chapter 6. Alternative Rating Approaches 88
Sensitivity Analyses Comparing Distributions of California QRIS Ratings and Alternative Rating Approaches 89
Concurrent Validity of Alternative Rating Approaches 92
Percentage of Classrooms Observed 95
Summary: How Do Alternative Rating Approaches Affect the Distribution and Validity of Ratings? 98
Chapter 7. Discussion and Preliminary Conclusions 99
Summary of Findings 99
Preliminary Conclusions and Limitations 105
Policy Options for Consideration 107
Other Considerations Relevant for Further Expansion of the System and Its Validation 108
References 111
Appendix A. Literature Review 124
Evaluation Evidence for QRISs 124
Conclusions and Implications for California 132
Appendix A1. Summary Tables of Studies Reviewed and Their Findings 134
Bibliography for Literature Review 140
Appendix B. Validation Study Methods 149
Study Samples 149
Measures 160
Summary of Data Challenges and Limitations 171
Analysis Methods 172
Appendix C. Descriptive Comparisons of Classroom Observation Scores in Small Samples 175
Descriptive Average Scores by Rating Level, FCCHs 175
Descriptive Average Scores by Rating Level, Toddler Classrooms in Centers 176
Appendix D. Element Score Analysis Results 177
Internal Consistency Results 177
Element Score Descriptive Results 178
Element Score Concurrent Validity Results 179

Appendix E. Alternative Rating Approach Analysis Results 190
Cross-Tabulations of California QRIS Ratings and Alternative Rating Approaches 190
Alternative Rating Approach Concurrent Validity Results.....193

Executive Summary

A mid-term report on a study conducted by American Institutes for Research (AIR) and its partners at the RAND Corporation, Survey Research Management, and Allen, Shea & Associates on California's quality rating and improvement system (QRIS) examines the validity of California's QRIS by assessing the extent to which the quality elements measured in the QRIS relate to each other and how well the QRIS ratings align with independent observations of quality. This report provides preliminary findings on the validity of the QRIS ratings:

- California's QRIS captures important aspects of quality.
- The quality elements in California's QRIS are not redundant; each element measures a distinct aspect of program quality.
- Ratings function differently for centers and family child care homes.
- Variation in ratings is limited both for centers and for family child care homes.
- There is some evidence that the ratings capture meaningful differences in quality: Higher rated programs were found to be of higher quality on some—but not all—independent measures of observed quality.
- Calculating ratings by taking an average of all element scores improves the validity of the ratings.

Additional study analyses shed light on possible ways to strengthen or simplify the way that ratings are calculated, such as taking an average score across elements to improve concurrent validity results. Although there is some evidence for the validity of California's QRIS ratings, it is still early in the system's implementation to draw firm conclusions. Further, most participating centers at the time of the study were rated at Tiers 3 or 4, and most participating homes were rated at Tiers 2 or 3, which limited our ability to find effects. The findings may differ with a more diverse group of participating programs. The final report in January 2016 also will include a child outcomes study, providing further evidence regarding the validity and reliability of the Race to the Top–Early Learning Challenge (RTT-ELC) QRIS in California.

Introduction

In 2011, California won a federal RTT-ELC grant to develop a locally driven quality rating and improvement system (QRIS) or set of systems. A QRIS is a uniform set of ratings, graduated by level of quality, used to assess and improve early learning and care programs. In January 2013, a network of 17 Early Learning Challenge Regional Leadership Consortia began implementing QRISs to expand and strengthen preexisting quality initiatives in 16 counties.

California's locally based approach sets common goals for workforce development, program assessment, and child assessment for school readiness but allows for some flexibility in quality benchmarks. The participating Consortia worked with the California Department of Education (CDE) and First 5 California to develop and implement the Hybrid Rating Matrix, which specifies the criteria for five QRIS rating levels. This matrix includes criteria for seven different aspects of quality, referred to as elements: Child Observation, Developmental and Health

Screenings, Minimum Qualifications for Lead Teacher/Family Child Care Home (FCCH), Effective Teacher-Child Interactions, Ratios and Group Size, Program Environment Rating Scales, and Director Qualifications.

Consortia agreed to adopt the rating criteria in the Hybrid Rating Matrix as part of a Quality Continuum Framework, with the option to make some local adaptations to tier requirements, assessment protocols, and the supports and incentives for quality improvement. Consortia may make specific types of local adaptations to Tiers 2 and 5 while maintaining three common tiers (Tiers 1, 3, and 4). The California QRIS is referred to as a hybrid rating approach because ratings are determined using a combination of points earned by meeting standards for different quality elements and "blocks" that require programs to meet minimum criteria across elements for a given rating level. The hybrid rating matrix has block requirements for Tier 1 and offers point ranges for Tiers 2, 3, 4, and 5. However, Consortia have the local option to treat Tiers 2 and 5 as blocks. Other local adaptations to Tiers 2 and 5 include adding supplemental criteria to reach the tier in addition to the blocks or point ranges specified in the Hybrid Rating Matrix. The Hybrid Rating Matrix is included and described in more detail in Chapter 3.

The Quality Continuum Framework, in addition to the Hybrid Rating Matrix, includes an accompanying document, the Continuous Quality Improvement Pathways Core Tools and Resources, adopted by the Consortia in October 2013. These are the tools and resources listed in the Federal application that the Consortia are required to include in their Quality Improvement Plans, and data are to be gathered on how the Consortia use the tools and resources. Although the Hybrid Rating Matrix requires the use of some of the tools and resources, such as the *California Preschool Learning Foundations* and the *California Infant/Toddler Learning and Development Foundations* (the *Foundations*), the *California Preschool Curriculum Frameworks* and the *California Infant/Toddler Curriculum Frameworks* (the *Frameworks*), and the *Desired Results Developmental Profile,* others are not included in the rating of participating programs.

The federal RTT-ELC grant requires an independent evaluation of each state's QRIS system, including a validation of the rating system and an assessment of program quality improvements associated with participation in the system. Validation studies of existing QRISs are important as they assess the extent to which ratings within the systems are meaningful and accurate and successfully differentiate low-quality programs from high-quality ones. Such studies also assess the degree to which ratings predict children's learning and development outcomes. Evaluation studies of QRISs are also needed to demonstrate that the system is successful at promoting quality improvement in early learning programs.

In January 2014, the CDE contracted with American Institutes for Research (AIR) and its partners at the RAND Corporation, Survey Research Management, and Allen, Shea & Associates to conduct the evaluation. The purpose of this interim report is to present our first year's findings, focusing on the Validity and Reliability Study of California's RTT-ELC QRIS. The final report, to be completed at the end of 2015, will include our findings on outcomes associated with QRIS participation.

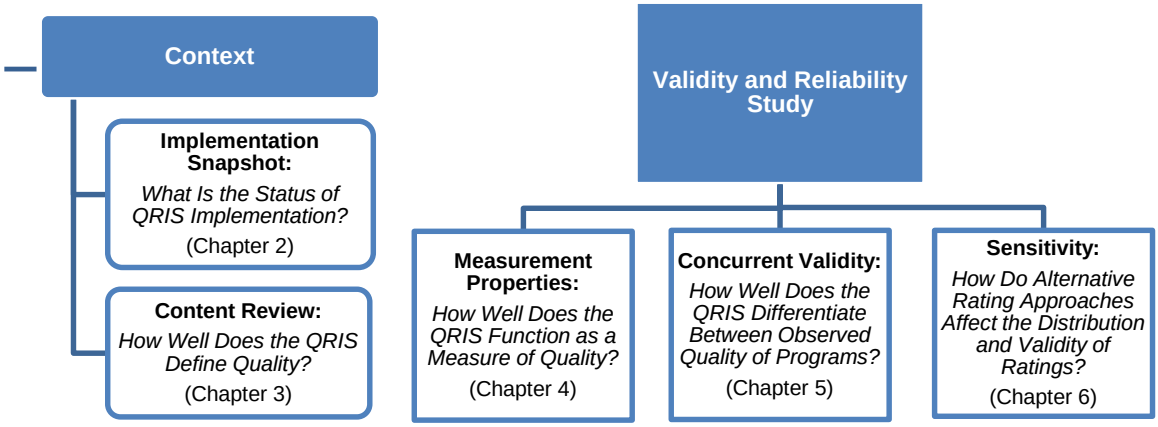
Overview of the Study

As a first step in the Validity and Reliability Study, we summarize the history and purpose of QRISs, review findings from other QRIS evaluation studies, and describe our own approach to validating the system in California. The majority of this report focuses on providing context for the California QRIS and assessing the validity and reliability of the system. The structure of the study, including the thematic questions that organize the report, is presented in Exhibit 1.

To provide context for the study, we offer a snapshot of the status of the implementation of the QRIS using the Hybrid Rating Matrix and the issues and challenges faced by Consortia. To understand how well the QRIS defines quality, we review the content of the Hybrid Rating Matrix, including an examination of the research base for the quality elements and rating approaches and common practices related to them across other QRISs.

We then conducted three activities to examine QRIS validity and reliability. To assess how well the QRIS functions as a measure of quality, we examine the distribution of the ratings obtained from the Common Data Elements for the RTT-ELC QRIS and assess the measurement properties of the rating. Drawing on our own classroom observations conducted for the study as well as data shared by the Consortia, we compare classroom observation scores for programs with different ratings to assess the degree to which the ratings capture quality differences as determined by an independent measure of quality. We also conduct sensitivity analyses, which compare the distribution and validity of ratings calculated using several different strategies including the local modifications to the Hybrid Rating Matrix that individual Consortia have adopted. Finally, the report summarizes the key preliminary findings of the validation study addressing six primary research questions and offers considerations for next steps.

Exhibit 1. Structure of the Validity and Reliability Study and Report



Study Design and Methods

The thematic questions and the underlying research questions are addressed by a series of analyses that draw on several main sources of data, described briefly in the following sections and in more detail in Chapter 1 and in Appendix B.

Interviews With QRIS Administrators

To provide a snapshot of the status of the implementation of the RTT-ELC QRIS in order to contextualize study findings, AIR and RAND staff interviewed the administrators of each of the 17 QRISs in the Regional Leadership Consortia in the spring of 2014. Depending on the structure of the Consortium, interviews included staff of local First 5 offices, county offices of education, and key partners. Using qualitative data analysis techniques, the study team analyzed the interview transcripts to gain an understanding of the work of each Consortium and to identify differences and common themes across them.

Content Review

To situate the RTT-ELC rating system and Hybrid Rating Matrix in the broader context and to assess the extent to which the ratings that are the product of the matrix are effectively *defining* quality, we conducted a content review of the Hybrid Rating Matrix. This review included an examination of the research base for each of the elements included in the Hybrid Rating Matrix, a review of the common indicators included in other states' QRISs, and an examination of California's hybrid rating method, and consideration of alternatives to this strategy for calculating ratings.

California's Common Data Elements

To assess the reliability and validity of the ratings, the study team collected extant data on the program characteristics and QRIS ratings of programs participating in the QRIS through each Consortium. These "Common Data Elements" include element scores, the sum of the element scores, the QRIS rating, the program average Classroom Assessment Scoring System (CLASS) scores used to calculate the CLASS element scores, and some program characteristics. Data were available for 1,272 programs, though only 472 had full ratings.[1] The remaining 800 did not have ratings or had ratings that were considered "provisional," in many cases because sites had not yet received their CLASS or Environment Rating Scales (ERS) observation. In these cases, the sites automatically received a score of 1 for these elements, and thus the rating was not able to capture their "true" level of quality on these elements. For this reason, only programs will full ratings were included in the validation analyses presented in this report.

Independent Classroom Observations

The study team collected classroom observation data from study sites in order to compare QRIS ratings and element scores against an independent measure of quality to assess concurrent validity. To measure classroom quality, we conducted classroom observations using three tools: CLASS, the Program Quality Assessment (PQA), and ERS. At the request of several of the Consortia and the CDE, we accepted some extant data from Consortia in lieu of conducting

[1] Programs with full ratings are defined as those with QRIS data that are complete and nonprovisional. Having complete and nonprovisional data is determined by having QRIS ratings and scores on each applicable rating element that are based on complete data for each element, and excludes programs awaiting classroom observations and thus without a finalized score on the associated elements. For the study, each Consortium identified the programs within their local QRIS that had provisional ratings as of January 2014 and these programs were not included in the validation analyses for this interim report.

direct observations of classrooms if the data had been collected within nine months of the study's data collection period. Complete data on the CLASS or PQA were obtained for 175 sites; these sites compose the concurrent validity sample. ERS data were also collected in a subset of sites for sensitivity analyses.

Study Limitations

Several limitations to the study are important to highlight. First, just over a third of programs across California that are participating in the RTT-ELC QRIS had a full, nonprovisional rating and were thus eligible for inclusion in the study. The study was launched while the RTT-ELC QRIS was still in the early stages of implementation, and many programs participating in the QRIS had not yet received a full rating at the start of the study because they did not have finalized scores on all of the rating elements. Furthermore, the fully rated programs have limited variability in rating levels and differ from programs without full ratings in several ways.

Second, the sample of programs that participated in data collection for the concurrent validity analyses had an insufficient number of family child care homes to permit statistical analysis for that subgroup, and the sample of centers in the concurrent validity sample had limited variability in QRIS ratings and was somewhat smaller than the anticipated sample size, in part because fewer programs were eligible for the study than anticipated and also because of delays in the start of recruitment for the study due to extended negotiations with the Consortia. The concurrent validity analyses cannot be considered conclusive because the small sample size and lack of variability in ratings among centers limits our ability to detect differences between each rating level. An additional implication of the limited samples is that the validation study results may not be generalizable to all programs participating in the RTT-ELC QRIS.

Third, there are some limitations to validation research conducted while the RTT-ELC QRIS is relatively new and not fully implemented. Although examining the system and how it is performing at this early stage has value and can help the state consider possible revisions to the QRIS, results presented in this report should be interpreted within the context of the system's stage of development and current participants, and conclusions should be considered preliminary. Furthermore, additional results related to predictive validity will be presented in the final report; this aspect of validation is also important to consider when evaluating the system as a whole.

What Is the Status of RTT-ELC Implementation?

To understand the status of the RTT-ELC QRIS implementation, it is important to stress a critical point at the outset: The system is still in its infancy, under development, and being continuously refined (see Chapter 2). Although some Consortia had longstanding, integrated systems in place prior to the receipt of the RTT-ELC grant, others had minimal experience with key components of QRIS implementation, such as conducting valid, reliable and independent CLASS and ERS observations. The differences in the prior history of the Consortia have implications not only for the status of the system implementation itself, but also for how many and for which types of programs are eligible to participate in this study.

Interviews conducted in spring 2014 revealed that Consortia had made progress, and most were on target with respect to their implementation plans. However, few had implemented all of their planned activities for the grant. Similarly, several Consortia had reached their targets for provider participation, but most Consortia were adhering to the phased-in enrollment projected in their initial plans. Many Consortia focused their efforts on conducting assessments in order to complete ratings for programs.

Although the majority of the Consortia were meeting the RTT-ELC requirements for frequency and sampling of classroom assessments (and several went beyond the minimum requirement), classroom observations and ratings were the biggest barrier to full implementation. Several Consortia voiced challenges around finding, training, and retaining qualified classroom assessors, particularly for the ERS. Others cited difficulties in aligning observations associated with the RTT-ELC with those for other quality initiative programs in the area, such as the Child Signature Program (CSP). QRIS administrators voiced particular challenges with finding sufficient ERS assessors because of the required training by an author or nationally certified anchor and the ongoing obligation to have frequent reliability checks. The costs inherent in getting staff or consultants trained to reliability were also identified as a barrier. To make observations more affordable, sustainable, and manageable, Consortia had implemented a number of strategies, including recruiting a large pool of classroom assessors prior to training and partnering with other Consortia or agencies in the area to share assessors.

In response to a federal RTT-ELC federal grant requirement, California's application included plans to provide objective ratings of early learning and development programs to families in an accessible, clear, and easy to understand format. However, perhaps in recognition of the time needed to implement a valid rating process, a quality continuum framework, and recruitment of programs to participate in the QRIS, publication of ratings was never viewed as the first step in system development.[2] As of July 2014, only one of the 17 Consortia had made the ratings available to the public through a searchable, online database. In contrast, most of the Consortia were focusing their time and energy on other grant requirements and were holding off on developing a clear plan to roll out the ratings. Almost half of the Consortia expressed concerns about publicizing ratings based on a tool, the Hybrid Rating Matrix, that was not yet validated.

In terms of validation, it is important to note that California's QRIS is not one uniform system. Although all of the Consortia use a common, five-tiered Hybrid Rating Matrix, each Consortium is allowed to make local modifications to elements within Tiers 2 and 5. However, as of June 2014, only two of the 17 Consortia had made changes to Tier 2; a few more had made changes to Tier 5 or were in the process of doing so. Overall, many Consortia cited the desire "to keep it simple" and not add additional costly quality elements to monitor. Consistent with findings shared in the *Descriptive Study* (AIR and RAND, 2013), most Consortia seemed to continue to prefer the hybrid system as opposed to a block system, which they thought might hinder participation by family child care homes (FCCHs) and private providers.

[2] California RTT-ELC Federal Grant Application, p. 94

How Well Does the QRIS Define Quality?

To examine how well the QRIS defines quality, we conducted a content review of the rating matrix (Chapter 3). California's RTT-ELC Hybrid Rating Matrix includes three core domains and seven quality elements for centers and five quality elements for FCCHs (see Exhibit 2).

Exhibit 2. Quality Elements Comprising the RTT-ELC Hybrid Rating Matrix

	Centers	FCCHs
CORE I: Child Development and School Readiness		
Child Observation	●	●
Developmental and Health Screenings	●	●
CORE II: Teachers and Teaching		
Minimum Qualifications for Lead Teacher or FCCH	●	●
Effective Teacher-Child Interactions: CLASS Assessments	●	●
CORE III: Program and Environment—Administration and Leadership		
Ratios and Group Size	●	
Program Environment Rating Scale(s)	●	●
Director Qualifications	●	

Program Quality Elements Within the Hybrid Rating Matrix

Overall, California's RTT-ELC Hybrid Rating Matrix includes three of the five most common quality elements found in QRISs across states—staff qualifications; environment; and program administration, management, and leadership (QRIS Online Compendium 2014). The Hybrid Rating Matrix also includes a key feature of the newer QRIS systems—child assessment. A review of the literature suggests that the elements in California's RTT-ELC Hybrid Rating Matrix have a research base (e.g., as predictors or correlates of program quality), with the strongest evidence for teacher-child interactions and program environment.

However, California's matrix does not match the field in a few areas. Unlike more than three fourths of the 38 state QRISs included in the QRIS Online Compendium (2014), the Hybrid Rating Matrix does not include a separate element for curriculum or alignment with state early learning foundations. Also, unlike 93 percent of other state QRISs, the Hybrid Rating Matrix does not have a separate indicator for family partnership (QRIS Online Compendium 2014), although this topic is included in a minor way in the Program Environment Rating Scales element by virtue of its being one of the subscales of the ERS. Other elements, such as Cultural and Linguistic Diversity, Dual Language Learning, and inclusion of children with special needs, are also elements considered by many to be important and are included in other systems but not in the California's Hybrid Rating Matrix.

Program Quality Assessment Measures Within the Hybrid Rating Matrix

The RTT-ELC Hybrid Rating Matrix includes the two most commonly used classroom quality measures: the CLASS and the ERS. Overall, the percentage of QRISs relying on the ERS alone

has declined from 67 percent in 2010 to 40 percent in 2014. At the same time, the percentage of systems using both the ERS and CLASS (or a third instrument) has increased from 7 percent in 2010 to 30 percent in 2014 (QRIS Online Compendium 2014). The literature shows that the CLASS and the Early Childhood Environment Rating Scale-Revised (ECERS-R) tools can be administered reliably and are modestly predictive of child outcomes. However, to use them for deriving program ratings, it is critical to set research-informed thresholds that programs need to attain in order to reach the highest quality levels. A review of the evidence base for other program quality assessment tools does not provide strong support for the addition of any other specific observations tools currently available to the Hybrid Rating Matrix.

Rating Structure Analysis

The building block approach remains the most common rating structure among QRISs; however, hybrid and points approaches gained in popularity between 2010 and 2014 (QRIS Online Compendium 2014). One recent study using simulated ratings found that whereas less than one fifth of programs achieved a rating level of 3 or 4 in a block structure, more than 70 percent achieve such a level in a points and hybrid structure (Tout and others 2014). These research findings indicate that the rating structure can significantly influence the distribution of programs that reach each level in a QRIS.

How Well Does the QRIS Perform As a Measure of Quality?

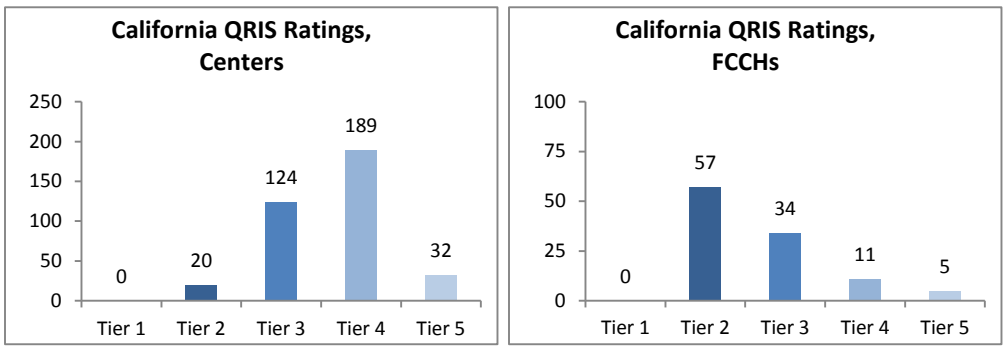
To determine how well California's QRIS ratings function as a measure of program quality, we analyzed QRIS rating data, including rating levels and element scores, from programs across the state with full QRIS ratings as of January 2014 (Chapter 4). The analysis included an examination of the distribution of ratings and element scores, the characteristics of programs that predict QRIS ratings, the internal consistency of the ratings, and how element scores relate to each other and to the overall rating.

Results indicate that, first, the distribution of ratings in the limited sample of fully rated programs is truncated; it does not span all five possible QRIS rating levels. Among the sample of 472 programs with full ratings, no programs were rated at Tier 1 using California's QRIS criteria. In contrast, 19 percent of programs with provisional ratings were rated at Tier 1, perhaps suggesting that programs do not complete a full rating until they are able to earn enough points to achieve Tier 2. Ratings of fully rated programs were generally high, with half of all sites rated at Tier 4 or higher. This may be due to the population of programs participating in the system: as a voluntary system, programs that might score lower have little motivation to become involved. In fact, many of the fully rated programs are State Preschool or CSP sites—programs with specific quality requirements—and many have been participating in quality improvement efforts for many years (prior to RTT-ELC funding and the development of the Hybrid Rating Matrix) and have thus had the benefit of significant professional development and quality improvement resources. However, relatively few programs (8 percent) were rated at Tier 5, indicating that the Tier 5 criteria set a high bar.

In addition, the distribution of ratings differs markedly for fully rated centers and FCCHs (see Exhibit 3). Although the most common rating for centers is Tier 4, and 86 percent of centers were rated at Tiers 3 or 4, the most common rating for FCCH is Tier 2, and 85 percent of FCCHs

were rated at Tiers 2 or 3. More than half of the FCCHs scored below a Tier 3 because they did not participate in an ERS observation, which may reflect the challenges that Consortia face in obtaining qualified assessors, as discussed previously. Once again, however, the extent of differences in ratings of centers and FCCHs is specific to the sample of programs with full ratings in January 2014. Differences in ratings between centers and FCCHs may be partially explained by differences in the percentage of centers (95.8 percent), and FCCHs (42.7 percent) that are required to meet high quality standards for State Preschool, Child Signature Program, or Head Start funding. It is not known if these patterns will be similar when a larger and more diverse sample of centers and FCCHs participating in the QRIS.

Exhibit 3. Distribution of California QRIS Ratings for Centers and FCCHs With Full Ratings in January 2014



Examining how elements relate to each other, we found that none of the element scores were redundant, indicating that the elements capture different aspects of program quality. However, as might be expected with a multidimensional scale like the QRIS rating, we found low internal consistency of the various elements of the overall rating, and some pairs of elements have very low correlations—this is true among centers and FCCHs alike. Low internal consistency does not suggest that the rating is flawed, but rather that the five to seven unique domains of quality measured for the QRIS are not always closely related to each other. These findings suggest that the California QRIS rating represents a multidimensional construct of program quality made up of unique elements.

Elements with greater variability in scores (such as Minimum Qualifications for Lead Teachers in centers and Program Environment Rating Scales and Developmental and Health Screenings for FCCHs) are more highly correlated with the overall QRIS rating, while others (such as Ratios and Group Sizes for centers and Effective Teacher-Child Interactions [CLASS] for FCCHs) are poorly correlated.

How Well Does the QRIS Differentiate Between Observed Quality of Programs?

To determine how effective the RTT-ELC rating structure is at defining and measuring quality in early learning settings, we evaluated the concurrent validity of the ratings (Chapter 5). Evaluating the concurrent validity involves comparing the ratings assigned by the Consortia to independent measures of quality to see how closely they align. Our analyses compare QRIS

rating levels and element scores from 175 fully rated centers with their scores on the independently observed measures of quality, the CLASS and the PQA.

Results from the concurrent validity analyses find some evidence that the California QRIS ratings differentiate between observed quality of programs. In particular, California QRIS ratings positively and significantly predict CLASS total scores, Preschool CLASS Instructional Support scores, and Preschool PQA Adult-Child Interaction scores. Sample sizes for toddler classrooms and FCCHs were not sufficient to produce reliable conclusions for these settings.

Among the concurrent validity analyses using element scores, only the element scores based on the CLASS and ERS consistently and significantly predict observation scores. The other element scores may be thought of as indicators of structural quality, and some previous studies have found that structural quality measures predict classroom observation scores (Burchinal and others 2002; Goelman and others 2006; NICHD Early Child Care Research Network 2002; Phillips and others 2000). However, in this study of the California QRIS, none of the structural element scores positively predict the PQA Form B, which is an independent measure of program structural quality. This lack of a relationship between structural element scores and the independent structural quality measure suggests that the element scores could be improved to ensure more variability.

Expanding QRIS participation to a more diverse group of programs may have the effect of increasing variability in the structural element scores and improving the relationship with the independent measures of structure quality. As with our findings on the distribution of California's QRIS ratings, it is important to stress that our sample of fully rated programs was small and of higher quality than programs participating in the QRIS in 2013 that had provisional ratings. Thus, results of the concurrent validity analyses may not apply to the broader range of programs participating in the QRIS.

How Do Alternative Rating Approaches Affect the Distribution and Validity of Ratings?

Ratings for a QRIS can be calculated many different ways; California's hybrid method is one approach. To explore how ratings and validity would change under different rating approaches, we tested six alternative rating approaches using the same element scores collected for the California QRIS ratings (shown in Exhibit 4 and discussed in more detail in Chapter 6). The state currently uses two of the alternative approaches as local adaptations to the statewide rating approach, and four are not currently used in the state.

Exhibit 4. Alternative Rating Approaches Examined in This Study

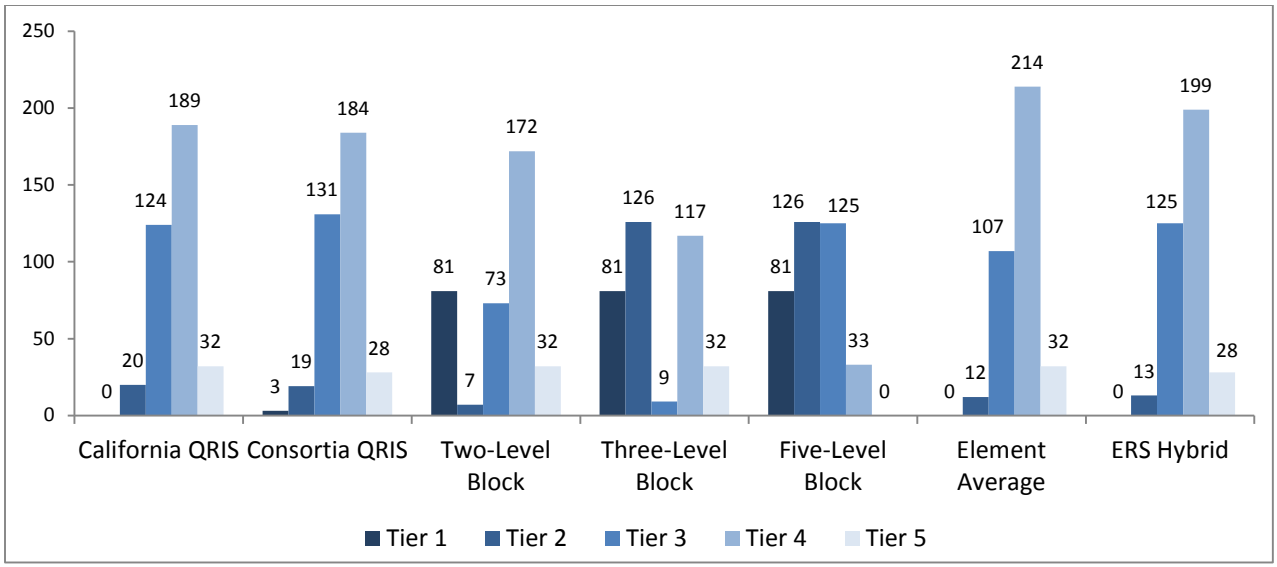
Rating Type	Rating Definition
Rating Approaches Currently Used in California	
California QRIS	Tier 1 is blocked; Tiers 2–5 are point-based for programs meeting block criteria for Tier 1: Rating is determined by total points earned across elements. This is California's rating approach without local adaptations to the way the ratings are calculated using the element scores.
Two-Level Block	Tiers 1 and 2 are blocked, and Tiers 3–5 are point-based for programs meeting block criteria for Tier 2. This approach is used as a local adaptation to California's rating approach in some counties.
Consortia QRIS	This is not a single rating approach but instead refers to the ratings assigned by Consortia, using local adaptations to the California QRIS ratings.
Rating Approaches Under Consideration, But Not Currently Used in California	
Three-Level Block	Tiers 1–3 are blocked, and Tiers 4–5 are point-based for programs meeting block criteria for Tiers 3.
Five-Level Block	Tiers 1–5 are blocked.
Element Average	Scores are determined by taking the average of all applicable rating elements. Averages are rounded to whole numbers (round up for 0.5 and greater, round down for 0.5 and less).
ERS Hybrid	Tier 1 is blocked, Tiers 2–4 are point-based for programs meeting block criteria for Tier 1, and Tier 5 is based on points for programs meeting block criteria for Tier 1 and a Tier 5 block in the program Environment Rating Scale element. Point ranges for Tiers 2–5 are adjusted to exclude the Program Environment Rating Scale element from the total points.

First, we found that the distribution of rating levels varies by rating approach (see Exhibit 5). The largest changes occur in rating approaches using blocks; 63 percent of programs have lower ratings when only Tier 2 is blocked, while 94 percent of programs have lower ratings when all five tiers are blocked.

Second, we found that element average ratings are more effective than California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores. Although ratings using blocks are less effective than California QRIS ratings at differentiating centers by CLASS scores, five-level blocks are more effective at differentiating centers according to the PQA observation scores. Both the ERS hybrid rating approach (removing the element score based on the ERS from the points-based part of the rating, but including a block requirement for an ERS score of 5.5 or higher for a Tier 5 rating) and the Consortia QRIS ratings (using the Consortia applied local options for modifying the Hybrid Rating Matrix) are similar to the California QRIS ratings in their patterns of relationships with classroom observation scores.

It is important to remember when interpreting these concurrent validity analyses using alternative rating approaches that they are specific to the sample of centers included in the study, and that the relationships between alternative rating approaches and observed quality scores may differ for other programs in California.

Exhibit 5. Distribution of Ratings Using Alternative Rating Approaches, Centers



Finally, we found that when we examined different samples of classrooms for observations, CLASS element scores were affected by the combination of classrooms selected in about one third of the programs in the small sample of centers included in this analysis. The Program Environment Rating Scale element score, on the other hand, was fairly stable, regardless of how many or which combination of classrooms' ERS scores were included in the element score calculation. In addition, although CLASS element scores varied depending on how many classrooms were observed, QRIS ratings among centers in our sample were rarely affected by these differences. Therefore, although the evidence is limited, this result suggests that although CLASS element scores might fluctuate depending on the sampling approach taken, the overall ratings are fairly robust.

Policy Options for Consideration

Although study results cannot be considered conclusive at this stage for the reasons previously described, our analyses do suggest some directions that may be worth consideration by the state, at least in a preliminary way. In this section, we offer some suggestions for modifications to the system that the state might want to consider in light of the evidence available to date and other contextual factors (see also Chapter 7).

Consider Ways to Increase Attention to Curriculum, Family Engagement, and Special Populations in the Hybrid Rating Matrix

As noted previously, the state may wish to consider adding to the Hybrid Rating Matrix alignment of curricula to the *Foundations* and *Frameworks*, as was recommended by the California Early Learning Quality Improvement System (CAELQIS) Advisory Committee. This alignment can provide greater assurance that teachers and providers are aware of the *Foundations* and *Frameworks* and are learning how to incorporate them into their instruction; that children in participating Consortia are receiving instruction consistent with the frameworks developed for the state; and that PK–3 alignment for teachers and students in California is being

supported by the QRIS. Adding this element would require Consortia to review program policies for compliance with the provision. Some states provide a list of curricula determined to be aligned with educational standards; employing one of these curricula meets requirements. Participating programs that prefer to use a different curriculum are required to demonstrate how their curriculum is aligned.

Adding specific reference to the *Foundations* and *Frameworks* in the Hybrid Rating Matrix would also help to address some of the other domains not fully represented in the Hybrid Rating Matrix, such as Cultural and Linguistic Diversity, Dual Language Learning, Cultural Competency, Special Needs, and Health Practices. These program elements have some theoretical foundation and are important, but either the research does not exist yet to show whether they do or do not link to improvements in teaching practices or children's outcomes, or there is little agreement on how to measure the elements. Adding alignment of curricula with the *Foundations* and *Frameworks* to the Hybrid Rating Matrix would offer a modest approach to addressing these issues.

In terms of family engagement, although the literature supports the importance of family engagement in early childhood programs, there is little consensus on how best to measure it (AIR and RAND 2013). The ERS, already in use, includes a subscale on *Parents and Staff,* although it is not a comprehensive measure of family engagement. A new measure—the *Family and Provider/Teacher Relationship Quality* measure (Kim and others 2014)—which has recently been developed, assesses site staff's knowledge, practices, and attitudes around family engagement. Although the measure has not yet been validated, the state might wish to explore the use of the tool as one option for addressing family engagement.

Consider Alternative Rating Strategies to Strengthen Validity or Simplify Implementation

Although some evidence supports the validity of the Hybrid Rating Matrix in its current form, our analyses shed light on ways to strengthen or simplify the rating approach that the state might consider. For example, ratings calculated by taking an average score across elements are more effective than the California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores. California's decision makers may wish to consider this as a simple alternative to the current rating strategy. However, it is important to remember that these concurrent validity results might change when programs with a wider distribution of ratings are included in the analytic sample.

The state might also consider modifying the ERS element in light of implementation challenges consistently experienced across Consortia. The ERS is a difficult and costly instrument on which to train and maintain a cadre of observers, and reducing this burden for Consortia likely will result in more fully rated programs. We explored one option for doing so in our analyses: limiting the requirement for the ERS to Tier 5 and blocking at that level. Results suggest that this change would have minimal impact on ratings. Although the validation study results did not test whether eliminating the ERS element would improve validity, they do indicate that it might be possible to reduce its use dramatically without affecting ratings. Given the implementation challenges associated with using the ERS, it may be wise to consider ways to reduce its use, especially if the system is to be sustained long term.

Consider Options for the Presentation of Ratings Information to Parents

Given the multidimensional nature of the Hybrid Rating Matrix, the positive results for the CLASS element, and the potential value of providing parents with more specific information that they can use in making care decisions, the state might consider presenting element scores or subratings along with summary ratings once the ratings become publicly available. This would enable parents to make finer distinctions between programs that might share the same or similar QRIS rating. The multidimensional nature of the rating and the fact that different rating elements measure different program components means that two programs with the same overall rating may actually have important underlying differences, reflecting varying strengths and weaknesses on different elements. Although the original intent of a hybrid rating system was to provide programs some flexibility in how they could reach certain levels of quality, in practice it makes comparing programs with the same ratings problematic. Moreover, parents may value some rating elements more than others; element scores would enable parents to focus their search on programs that rate highest on the elements about which they may care most.

Other Considerations Relevant for Further Expansion of the System and Its Validation

In addition, though not directly arising from the validation study results, the state may want to explore other considerations relevant to further research and validation. To support continuous quality improvement in the QRIS, the state may want to consider ways to expand the system and may also want to consider supporting another validation phase when the QRIS is more mature and more programs are participating in the system.

Consider Ways to Encourage or Require More Providers to Participate in the System

One issue for the state to consider is whether ratings should be voluntary or required and, if required, for which programs. Although the validation analyses do not directly address this, we note frequently in this report that one of the major limitations of this research has been the relative lack of variation in the sample of programs participating in the study. The majority of programs and providers participating have been at Tiers 3 or 4, with no programs from Tier 1 and only a few at Tiers 2 and 5. Moreover, the sample is heavily skewed toward state and federally contracted programs that were already held to a set of contract standards intended to focus on quality before the implementation of the RTT-ELC QRIS. The lack of variation is not just a problem for researchers attempting to gauge the effectiveness of the system in rating quality; the narrow range of programs participating also limits the potential impact of the QRIS in providing information to families choosing care for their young children. It also forgoes an opportunity to assess the quality of the large group of private programs receiving some public funds in the form of vouchers, and makes it difficult for the public or policymakers to determine how best to direct limited resources for quality improvement.

The state might, therefore, want to consider piloting a system in one or more counties that requires all centers and FCCHs receiving state and federal subsidies to participate in the QRIS. At least nine states require programs receiving subsidies from the federal Child Care and

Development Fund to participate in their QRIS, and several states, such as Illinois and Washington, make participation mandatory for school-operated early care and education programs. Finally, such a pilot would provide a more complete picture of the extent to which the rating system captures the distinctions between all five tiers in the Hybrid Rating Matrix.

Of course, if participation were mandatory, it would be important to ensure that programs had access to program quality assessments so that all sites could be assessed and receive a full—as opposed to provisional—rating. Establishing a process that would ensure such access to newly mandated programs would be an important part of a mandatory participation pilot before statewide implementation could be considered.

Consider Another Validation Phase Once the System Is Further Developed

As noted throughout this report, data limitations due in part to the QRIS's stage of development constrain the analyses and limit the generalizability of the results. To address this constraint, the state might consider revisiting system validation once refinements currently under discussion are made and once the system is expanded to include a more diverse array of programs. If further analyses are to be conducted, it would be essential for Consortia to collect, maintain, and share with the state additional classroom- and site-level data. Such data would enable additional analyses and suggest evidence-based refinements; this work would not be possible without these more detailed data. In particular, it would be helpful to have raw element-level data (e.g., ratios, ERS scores). In addition to being useful for accountability purposes, retaining these data would permit the examination of element score cut points and the simulation of ratings based on modified cut points in order to refine the element scoring criteria. Such refinements would strengthen the reliability and validity of the ratings, making the QRIS a more meaningful signal of quality for parents and a more effective tool for targeting quality improvement resources.

Acknowledgments

The American Institutes for Research and RAND Corporation study team extends our deep appreciation to the representatives from the 17 RTT-ELC Leadership Consortia that contributed to this research. In particular, we thank the county administrators who participated in interviews, provided data, and encouraged their early learning and development sites to participate in the study. We also would like to thank the site administrators, providers, and teachers who opened their doors to the study team and allowed us to observe their practice.

We also wish to acknowledge the invaluable assistance of data analysts and research support staff who contributed to this research, including Alejandra Martin, Debbie Davidson-Gibbs, Kiana Abram, Susannah Faxon-Mills, Raquel González, Carmen Martínez, Nicol Christie, Christine McGuigan, Martha Ramirez, Erik Loewen, Shaheen Khan, John Mezzanotte, and Megan Brown. We also thank Anja Kurki and Bokhee Yoon for their technical review and guidance.

Thanks also go to our partners at Survey Research Management – Linda Kuhn, Tony Lavender, Betsy Quicksall, Aimee Elsey, Lyn Bopp, Ashley Bronzan, Daniel Mackin, and all of the field staff who collected data; and at Allen, Shea & Associates – William Allen, Mony Flores-Bauer, and Mechele Small Haggard for their contributions to research support activities.

Last, we wish to acknowledge the guidance and input provided by Cecelia Fisher-Dahms, Channa Hewawickrama, Gretchen Williams, and other staff of the California Department of Education, Early Education and Support Division, and First 5 California.

Chapter 1. Introduction

In 2011, California successfully submitted a Race to the Top–Early Learning Challenge (RTT-ELC) grant application that would move the state toward a locally driven Quality Rating and Improvement System (QRIS) or set of systems. The state proposed building a network of 17 Early Learning Challenge Regional Leadership Consortia that had already established—or were in the process of developing—QRIS initiatives in 16 counties. These Consortia, composed of local First 5 commissions, county offices of education, and other key stakeholders, represent counties which together have more than 1.8 million children ages birth to five. This locally based approach sets some common goals for workforce development, program assessment, and child assessment for school readiness but allows for some flexibility in quality benchmarks. The counties participating in the RTT-ELC Regional Leadership Consortia have voluntarily adopted a Hybrid Rating Matrix that allows considerable variability in some of the local tier requirements, the local rating protocol, and the supports and incentives for quality improvement.

The RTT-ELC grant included an independent evaluation of the system; that evaluation consists of a Year 1 validation of the rating and a Year 2 assessment of outcomes associated with participation in the system, including an examination of children's outcomes, to assess predictive validity. In January 2014, the California Department of Education (CDE) contracted with American Institutes for Research (AIR) and its partners at the RAND Corporation, Survey Research Management, and Allen, Shea & Associates to conduct the evaluation. The first year's findings, focusing on QRIS validation, are presented in this interim report.

In this introductory chapter, we present a brief summary of the history and purpose of QRISs as well as a review of what other QRIS evaluation studies have found. We provide an overview of the goals and approach to be used in the evaluation of California's RTT-ELC QRIS, including the study questions and methods that drive the Year 1 validation component of the study. The chapter concludes with an overview of the report, its structure, and contents.

Background on QRISs

Research findings highlight the importance of the period from birth to school entry for child development and focus attention on the quality of care and early learning experiences that young children receive (Center on the Developing Child at Harvard University 2007; National Research Council 2001; Shonkoff and Phillips 2000; Vandell and Wolfe 2000). Numerous studies have demonstrated that higher quality care, defined in various ways, is related to positive developmental outcomes for children, including improved language development, cognitive functioning, social competence, and emotional adjustment (e.g., Burchinal and others 1996; Clarke-Stewart and others 2002; Howes 1988; Mashburn 2008; National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network [ECCRN] 2000; Peisner-Feinberg and others 2001; Weiland and others 2013), although the benefits tend to be largest for children from disadvantaged backgrounds (e.g., Gormley and Gayer 2005; Gormley and others 2005; Karoly 2009; Pianta and others 2009). More recent studies that examine the effects of dosage (how long a child has been attending a program, as well as cumulative participation in specified programs, e.g., Burchinal, Kainz, and Cai 2011) and quality thresholds (whether a particular quality level must be achieved to demonstrate effects on children,

summarized in Zaslow and others 2010) underscore the importance of high-quality care in improving child outcomes.

Research also suggests that, when faced with choices in early care for children, parents are not always accurate in rating the quality of care provided to their children (e.g., Helburn, Morris, and Modigliani 2002). Parents tend to rate child care providers very positively (e.g., Barraclough and Smith 1996; Cryer and Burchinal 1997; Helburn 1995; Wolfe and Scrivner 2004), and their ratings do not correlate with observer quality ratings (e.g., Barraclough and Smith 1996; Cryer and Burchinal 1997; Cryer, Tietze, and Wessels 2002). Many parents (inaccurately) believe that licensing includes scrutiny of program quality and that licensure indicates that a program is of high quality (National Association of Child Care Resource and Referral Agencies 2011).

These findings highlight the need for systematic, reliable, and valid information about the quality of the care and early learning environments for young children—such as that provided through a QRIS—to be publicly available. Thus, quality rating and improvement systems aim to (1) provide quality information to parents to inform their choice of early learning and development programs for their children and (2) expand meaningful parental choice by supporting program quality improvement.

QRISs were first introduced a little more than 15 years ago and were operating in 22 states and the District of Columbia by 2010 (Tout and others 2010a). To date, all but one state currently implement or plan to implement some form of QRIS (QRIS National Learning Network, 2013). QRISs have recently garnered national attention through the U.S. Department of Education's RTT-ELC grant program. In the RTT-ELC request for applications (RFA), the Department of Education (ED) encouraged each state to design and implement a tiered quality rating and improvement system that was standards based and that provided "meaningful" ratings for the quality of each program. ED also encouraged broad participation in the QRIS across program types, with a priority toward including all licensed or state-regulated early learning and development programs in the system. In addition, ED emphasized a focus on continuous program improvement and a dissemination plan for ratings that would allow families to make informed decisions about which programs could best serve the needs of their children. Also required as part of RTT-ELC funding was a rigorous evaluation and validation of the QRIS (U.S Department of Education, 2011, p. 8).[3]

In California, the movement to create a QRIS pre-dates the federal focus on QRIS development. Beginning in 2004, First 5 California funded Power of Preschool initiatives featuring many of the typical elements of a QRIS: quality standards, provider support, program quality assessments, ratings to determine the level of payment, and financial incentives. A number of counties established their own initiatives designed to use publicly disseminated ratings as the major impetus for quality improvement.

In 2008, Senate Bill 1629 established a California Early Learning Quality Improvement System (CAEL QIS) Advisory Committee to design a QRIS for California. The committee produced a

[3] RTT-ELC application information is available at http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/applicant.html.

report in December 2010 that detailed a design for a QRIS with a block system, (where all elements in one tier must be achieved before advancing to the next tier) that included five quality elements for the rating structure. The CAEL QIS Advisory Committee proposed piloting the system over three years before implementing it on a statewide basis and advised that the system should be phased in over five years or more, after the completion of the pilot. In 2011, before the piloting of the proposed system had begun, the State of California—citing serious budget concerns as well as the challenge of implementing a one-size-fits-all program in such a large and diverse state—successfully submitted an RTT-ELC application that moved toward a more locally driven QRIS approach. The state proposed building a network of 17 ELC Regional Leadership Consortia across 16 counties that already had established, or were in the process of developing, QRIS initiatives. Key participants in the Consortia include local First 5 commissions and county offices of education as well as other key stakeholders.

Why Evaluation, and Validation Studies in Particular, Are Important

The investment of considerable federal and state funds to improve the quality of early learning and development programs using QRIS initiatives has increased the need for informative and rigorous evaluations of QRISs across states. A major component of QRIS evaluations are validation studies. As a tool, QRISs have tremendous potential to transform the early childhood landscape; however, understanding the validity of the ratings they provide is critical to determining their utility. Validation studies of existing QRISs are needed to demonstrate that ratings within the systems are meaningful and accurate and that they successfully differentiate low-quality programs from high-quality programs. When conducted with rigor, validation studies of QRISs assess whether the ratings developed in the system can be accurate indicators of program quality and whether they predict learning and development outcomes for children. In addition to the validation of the rating itself, evaluations of QRISs are also needed to demonstrate that the system is successful at promoting quality improvement in early learning programs.

Literature Review: What Other Evaluations Have Found

In a literature review for the *Local Quality Improvement Efforts and Outcomes Descriptive Study* (AIR and RAND 2013) and updated for this report, the AIR/RAND study team found that although QRISs are being designed and implemented in most states, evaluation evidence for QRISs comes from just 12 states or substate areas. In California, where local QRISs and quality improvement systems have been developing for many years, most of these efforts have incorporated evaluation in the process of program design and implementation. For the *Descriptive Study*, the AIR/RAND team reviewed 30 local evaluations that provided some initial evidence to support the validity of the quality improvement initiatives by demonstrating associations between participation in the systems and program quality improvements. The current RTT-ELC funding, however, requires a more rigorous evaluation and validation approach moving forward.

For this report, we reviewed empirical evaluations of existing QRISs to identify what is known from the published literature about effective system design and evidence of system impact. Our review of QRIS evaluation studies produced the following key points regarding validation and impact findings:

- The 14 evaluations (across 12 states or substate areas) we identified almost exclusively consist of validation studies that address one or more questions about the effectiveness of the QRIS design. Only one study provides any evidence of QRIS impact and only for a narrow question.
- Eleven studies examined the relationship between QRIS ratings and a measure of program quality. Ten of the 11 studies used the Environment Rating Scales (ERS) as an outcome measure. All but one found that the system ratings were correlated positively with observed quality, although the correlation was not always statistically significant. Moreover, the ERS was generally not an independent measure of quality, as it was used to determine the ratings that were being validated.
- Five studies aimed to determine whether program ratings or other program quality measures improve over time. These studies provide consistent evidence, given the way quality is defined, measured, and incentivized in the QRIS, that programs can raise their rating and improve their quality over time.
- Seven studies examined the relationship between QRIS ratings and child developmental outcomes. The findings from these studies are mixed, at best, indicating that there is little evidence to date to suggest that QRIS ratings, as currently configured, are predictive of child gains for key developmental domains.
- Two studies provide validation evidence about parents' knowledge and understanding of the QRIS ratings. These studies conclude that parents in rated programs know more about the rating system than the general public does and that knowledge of the system tends to increase over time. Even so, the extent of parental awareness of the examined QRISs did not exceed 20 percent for the general public and 40 percent for those using rated providers.
- Although QRIS designers may ultimately be interested in measuring the impact of implementing key elements of a QRIS, or a QRIS as a whole, on a range of system outcomes—provider mix, parental choice, teacher professional development, program quality, or child outcomes—making such causal inferences requires experimental or quasi-experimental designs that have rarely been implemented to date. The one available experimental study demonstrates the potential for using scientifically rigorous methods to extend our understanding of the causal impacts of QRIS implementation.

The complete literature review can be found in Appendix A.

California's RTT-ELC QRIS

As described above, California's RTT-ELC grant led to a new QRIS that was adopted by 17 Consortia representing 16 counties in 2013. These participating counties include a mix of small and large counties representing diverse areas of the state, and include some counties with no previous QRIS as well as other counties that had operated separate local QRISs for as long as a decade. The participating Consortia worked with the CDE to develop the Hybrid Rating Matrix, which specifies the criteria for five rating levels. Consortia agreed to adopt the rating criteria in the Hybrid Rating Matrix, with the option to make some local adaptations to Tiers 2 and 5 while maintaining three common tiers (Tiers 1, 3, and 4). The California QRIS is referred to as a hybrid rating approach because ratings are determined using a combination of points earned by meeting standards in different quality elements and "blocks" that require programs to meet minimum criteria across elements for a given rating level. The hybrid rating matrix has block requirements for Tier 1 and offers point ranges for Tiers 2, 3, 4, and 5. However, Consortia have the local option to treat Tiers 2 and 5 as blocks. Other local adaptations to Tiers 2 and 5 include adding supplemental criteria to reach the tier in addition to the blocks or point ranges specified in the Hybrid Rating Matrix. The Hybrid Rating Matrix is included and described in more detail in Chapter 3.

California QRIS Key Terms

Consortia: County-based agencies administering the QRIS locally

Tiers: California QRIS rating levels, ranging from 1 (lowest) to 5 (highest)

Elements: Aspects of quality measured in California's QRIS. Programs receive scores from 1 to 5 on as many as seven elements (the number of rated elements depends on the program type). The element scores are used to determine the program's Tier.

Hybrid Rating Matrix: The California QRIS document that outlines criteria for each element score, as well as criteria for each Tier. Consortia may make local adaptations to the criteria for Tier 2 and Tier 5.

Continuous Quality Improvement Pathways: The California QRIS document that outlines additional aspects of quality that are not measured for the QRIS but are prioritized as part of the state's Quality Continuum Framework.

Accompanying the Hybrid Rating Matrix as part of a Quality Continuum Framework is the Continuous Quality Improvement Pathways. The Pathways Core Tools and Resources includes the California *Foundations* and *Frameworks*, *Preschool English Learner Guide*, the Desired Results Developmental Profile Assessment, Ages and Stages, Center on the Social and Emotional Foundations for Early Learning (CSEFEL), Strengthening Families Five Protectors Factors Framework, and other resources listed in the federal application that the Consortia are required to include in their Quality Improvement Plan. Data are to be gathered regarding how these tools and resources are used by the Consortia. Although some of the resources also are listed in the Hybrid Rating Matrix, others are not included in the ratings.

The Evaluation of California's RTT-ELC QRIS

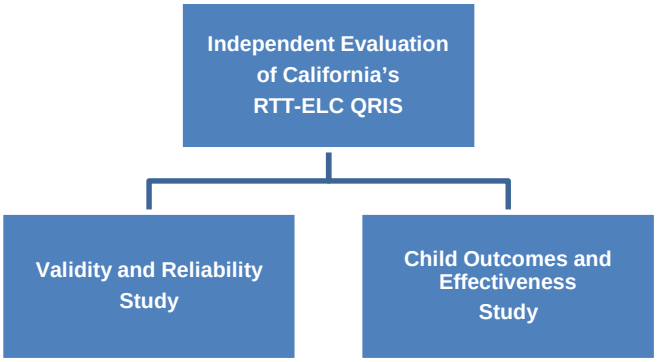
Adding to the body of literature described previously is this independent evaluation of California's RTT-ELC QRIS; this half-term report summarizes the results of the Year 1 validation component of the study. The two-year evaluation is intended to provide information about the validation of the system as well as program quality outcomes associated with

participation in the system with the hope of informing further refinements to the system and enhancing its effectiveness at assessing and supporting program quality.

Overview of the Evaluation

The evaluation is organized into two broad studies—the Validity and Reliability Study and the Child Outcomes and Effectiveness Study (see Exhibit 1.1).

Exhibit 1.1. Overview of Study Structure



Validation studies of QRIS ratings typically involve one or more approaches (Lugo-Gil and others 2011; Office of Planning, Research, and Evaluation 2012; Zellman and Fiene 2012):

1. **Content validity**—the extent to which the quality standards used for the QRIS ratings include the key domains of quality according to empirical research and expert opinion
2. **Concurrent validity**—the association between the quality rating and independent measures of quality
3. **Reliability**—the extent to which the ratings and rating components exhibit sound measurement properties and differentiate programs as expected
4. **Sensitivity**—the extent to which program ratings are affected by alternative rating calculation methods
5. **Predictive validity**—the association of quality rating levels with children's early learning and development outcomes

In addition, evaluation studies of QRISs may go beyond validation research, to examine the following:

6. **QRIS implementation—**documentation of how QRISs function and issues that arise during implementation or expansion
7. **Quality improvement—**how quality improvement supports and incentives embedded in the QRIS affect the quality of early learning programs or child outcomes

Evidence from validation research informs policymakers about the extent to which the QRIS ratings detect meaningful and reliable differences in program quality and, if needed, provides actionable recommendations on how specific changes to the QRIS rating approach would improve differentiation or the reliability of the ratings. Evidence from evaluation research

informs policymaking, to ensure that investments in program quality are cost-effective and linked to measurable improvements in program quality and child outcomes.

The independent evaluation of California's RTT-ELC QRIS addresses all seven aspects of validation and evaluation research described previously. The first four elements of validation studies are the main focus of the first year of the study (2014) and compose the Validity and Reliability Study, described in detail in the next section. The Outcomes and Effectiveness Study includes an examination of predictive validity as well as QRIS implementation and quality improvement and is the focus of 2015, to be discussed in detail in the Final Report.

The Validity and Reliability Study

The Validity and Reliability Study addresses six primary research questions, which can be classified according to the aspects of validation studies outlined previously: content validity, concurrent validity, reliability, and sensitivity. As shown in Exhibit 1.2, several questions have multiple components and are answered through a combination of analytic approaches.

Exhibit 1.2. Validity and Reliability Study Research Question by Study Component

Research Question	**Content Validity:** *How well does the QRIS define quality?*	**Measurement Properties:** *How well does the QRIS perform as a measure of quality?*	**Concurrent Validity:** *How well does the QRIS differentiate between observed quality of programs?*	**Sensitivity:** *How do alternative rating approaches affect the distribution and validity of ratings?*
1. How effective are the California Common Tiers' structure and components/elements at defining and measuring quality in early learning settings?	●	●	●	
2. Do point values of each element and the final rating provide meaningful distinctions between programs and program types?		●	●	
3. Do element levels relate to each other in consistent ways (e.g., Classroom Assessment Scoring System [CLASS] or ERS score and their relationship to other elements)?		●		
4. How is the hybrid rating strategy and rating outputs representative of meaningful levels of quality?			●	●
5. How do QRIS ratings that use locally determined tiers differ from QRIS ratings calculated using recommendations in California's RTT-ELC QRIS Implementation Guide?				●

6. How effective is the rating protocol at determining valid ratings versus an annual 100 percent assessment protocol?

We draw on several data sources and methods to address the questions, including a content review of the Hybrid Rating Matrix, interviews with QRIS administrators, analysis of extant ratings data from the Consortia, and independent classroom observations to assess program quality separate from the rating system. Following each of these approaches is described; additional detail on the validation study methods can be found in Appendix B.

Interviews With QRIS Administrators

To provide a snapshot of the status of the implementation of the RTT-ELC QRIS to contextualize the study findings, AIR and RAND staff interviewed the administrators of each QRIS program in the Regional Leadership Consortia in May and June 2014. These interviews were designed to learn more about the work the Consortia had done on their quality improvement systems as of early summer 2014. As a starting point, the interviews used the information collected through AIR and RAND's interviews with Consortia conducted in spring 2013 for the *Local Quality Improvement Efforts and Outcomes Descriptive Study* (*Descriptive Study*) for the CDE, as well as from the brief interviews about relevant data that AIR staff conducted with the administrators in spring of 2014. Depending on the structure of the Consortia, AIR and RAND interviewed staff of local First 5 offices, county offices of education, and key partners. Using qualitative data analysis techniques, the study team analyzed the interview transcripts to gain an understanding of the work of each Consortium and to identify differences and common themes across Consortia.

Content Review

To situate the rating system in the broader context and to assess the extent to which the ratings that are the product of the matrix are effectively *defining* quality, we conducted a content review of the Hybrid Rating Matrix. This review included an examination of the research base for each of the elements included in the Hybrid Rating Matrix in order to confirm the relevance of each element and identify any additional elements that might warrant being added to the matrix for a more comprehensive picture of quality. We also reviewed common indicators included in other states' QRISs to examine how California compares and consider alternative approaches. The review also considered California's hybrid rating method and explored alternatives to this strategy for calculating ratings.

California's Common Data Elements

To assess the reliability and validity of the ratings, the study team collected extant data on the program characteristics and QRIS ratings of programs participating in the QRIS through each Consortium. These data, submitted to the state using the QRIS reporting requirements, are referred to as the Common Data Elements and include data on program type, enrollment, funding sources, languages spoken in the program, element scores, the sum of the element scores, the QRIS rating, and the program average CLASS scores used to calculate the CLASS element

scores. Data were available for 1,272 programs, though only 472 had full ratings, and the remaining 800 did not have full ratings.

Independent Classroom Observations

The study team collected classroom observation data from study sites in order to compare QRIS ratings and element scores against an independent measure of quality to assess concurrent validity. To measure classroom quality, we conducted classroom observations using three tools: CLASS, the Program Quality Assessment (PQA), and ERS. At the request of several of the Consortia and the CDE, we accepted some extant data from Consortia in lieu of conducting direct observations of classrooms if the data had been collected within nine months of the study's data collection period. Complete data on the CLASS or PQA were obtained for 175 sites, which make up the concurrent validity sample. ERS data were also collected in a subset of sites for sensitivity analyses.

Full Versus Provisional Ratings

- Programs with full ratings are defined as those with QRIS data that are complete and nonprovisional.
- Complete data is determined by having a QRIS rating and scores on each applicable rating element. (The number of applicable rating elements is determined by the Hybrid Rating Matrix and varies by program type, ranging from 4 to 7 elements.)
- Nonprovisional data further excludes programs awaiting classroom observations and thus without a finalized score on the associated elements.
- For the study, each Consortium identified the sites within their local QRIS that had provisional ratings as of January 2014.
- The study team excluded programs identified as having provisional ratings as well as those without complete data on the QRIS ratings and applicable element scores because inclusion of non-finalized QRIS ratings would bias the study results.

Challenges and Limitations

Several limitations to the study are important to highlight at the outset. First, just over a third of programs across California that are participating in the RTT-ELC QRIS had a full, nonprovisional rating and thus were eligible for inclusion in the study. The study was launched while the RTT-ELC QRIS was still in the early stages of implementation, and many programs participating in the QRIS had not yet received a full rating at the start of the study because they did not have finalized scores on all of the rating elements. Furthermore, these programs have limited variability in rating levels and differ from programs without full ratings in several ways.

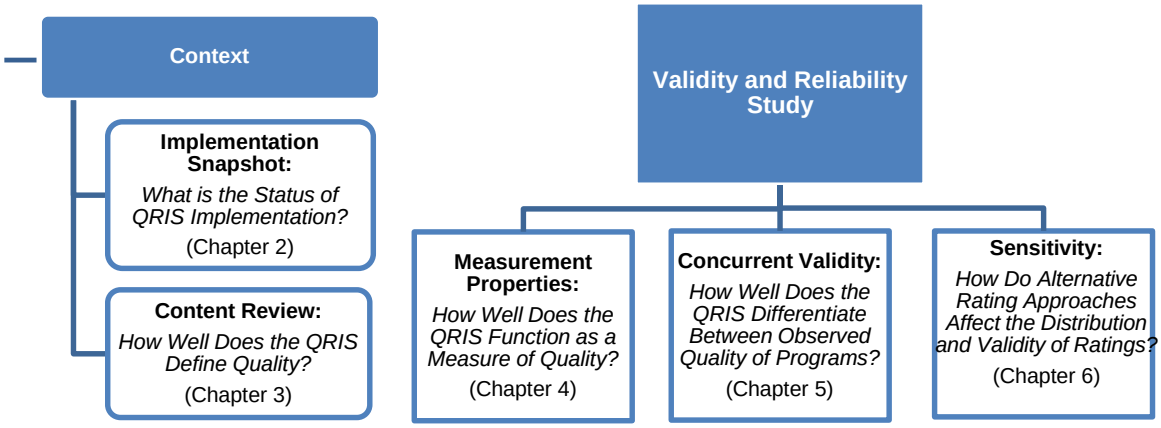
Second, the sample of programs that participated in data collection for the concurrent validity analyses had an insufficient number of family child care homes to permit statistical analysis, and the sample of centers in the concurrent validity sample had limited variability in QRIS ratings and was somewhat smaller than the anticipated sample size, in part because fewer programs were eligible for the study than anticipated and also because of delays in the start of recruitment for the study due to extended negotiations with the Consortia. The concurrent validity analyses cannot be considered conclusive because the small sample size and lack of variability in ratings among centers limits our ability to detect differences between each rating level. In addition, an implication of the limited samples is that the validation study results may not be generalizable to all programs participating in the RTT-ELC QRIS.

Third, there are some limitations to the validation research conducted because the RTT-ELC QRIS is relatively new and not fully implemented. Although examining the system and how it is performing at this early stage has value and can help the state consider possible revisions to the QRIS, results presented in this report should be interpreted within the context of the system's stage of development and current participants, and conclusions should be considered preliminary. Furthermore, additional results related to predictive validity will be presented in the final report; this aspect of validation is also important to consider when evaluating the system as a whole.

Organization of this Report

This half-term report, focusing on the Validity and Reliability Study of the Evaluation of California's RTT-ELC QRIS is organized into seven chapters, including this introductory chapter. The report is organized around the study components outlined in Exhibit 1.2 rather than by the specific research questions. However, we return to the research questions, which are typically addressed by analyses described in more than one chapter, in the final chapter of the report. Exhibit 1.3 provides a graphical overview of the structure of the report.

Exhibit 1.3. Structure of the Validity and Reliability Study and Report



Chapter 2 offers a snapshot of the implementation of the QRIS and the Hybrid Rating Matrix. It is early in the implementation of the system, thus the issues and challenges faced by Consortia are important context for interpreting the results and considering next steps for refinement of the system. This chapter draws on the interviews with QRIS administrators conducted in spring of 2014.

Chapter 3 provides a content review of the Hybrid Rating Matrix, including an examination of the research base and common practices across other QRISs for the quality elements as well as the rating approach.

Chapter 4 describes the distribution of the ratings obtained from the Common Data Elements and examines the measurement properties of the rating.

Chapter 5 focuses on concurrent validity results that compare classroom observation scores for programs with different ratings to assess the degree to which the rating is capturing quality differences as determined by an independent measure of quality. This chapter draws on primary classroom observation data collected for the study as well as the Common Data Elements provided by the Consortia.

Chapter 6 evaluates the sensitivity of the rating by comparing the distribution and concurrent validity of the ratings calculated under different rating approaches. Drawing on classroom observation data as well as the Common Data Elements, we present results for several different strategies including the local modifications to the Hybrid Rating Matrix that individual Consortia have adopted.

Chapter 7 summarizes the key findings presented in Chapters 2 through 6 organized by the research questions outlined in Exhibit 1.2. We also describe study limitations and present some preliminary observations and considerations for next steps.

Chapter 2. Snapshot of RTT-ELC QRIS Implementation

Although the focus of this report is on the validation of the QRIS in California, this chapter provides contextual information about California's QRIS implementation that is important for understanding the validation study findings.

Data and Sample

- Analyses in this chapter use data from interviews with administrators from each of the 17 Consortia.
- Interviews were conducted in spring 2014 and focused on implementation issues, challenges, and strategies.

Analysis Approach

- Data were analyzed using qualitative data analysis techniques to identify common themes across interview respondents and to describe the frequency of different experiences.

We begin by describing the current status of the state's QRIS. Still in its infancy as a system, California's QRIS is under development and continuous refinement. Some Consortia, especially those that did not have long-standing, integrated systems in place, have faced a number of challenges to reaching full implementation. For example, as of 2014, some Consortia had had years of experience in planning for and conducting valid, reliable, and independent classroom ECERS and CLASS observations, whereas in other Consortia with minimal experience or insufficient extant resources for these observations, this work has proven to be a considerable undertaking.

It is also critical to understand that California's QRIS is not one uniform system. For example, although all Consortia use a common, five-tiered Hybrid Rating Matrix, each Consortium is allowed to make local modifications to elements within Tiers 2 and 5. For example, some Consortia may have already added elements to Tiers 2 or 5, and others were considering making future adjustments to elements to make them even more appropriate indicators of quality than the current Tier 2 or 5 language describes. Because these adaptations can both impact validation analyses and shed light on elements that particular Consortia deem important, or alternatively, problematic, we also describe the local options that Consortia had either already implemented or were under consideration.

Status of Implementation

First we explore the current status of implementation of the RTT-ELC QRIS across the 17 Consortia, including progress toward accomplishing the activities outlined in their plans, reaching provider participation targets, and plans for releasing ratings information to the public.

Activities

Although the majority of the Consortia had not yet implemented all of their planned activities, most were on target with the deliverables outlined in their initial plans.

As of the interviews conducted between May and June 2014, four of the 17 Consortia had implemented all of their planned activities. The remaining 13 had not yet done so, though eight

of these 13 noted that they were following their initial plan and were on target with their deliverables.

The administrators of three Consortia emphasized that they had made adjustments and modifications to the initial plan, based on what they had discovered thus far during implementation. In a few instances, many of the changes in direction had to do with training and professional development. For example, the administrator of one Consortium noted that the implementation involved a continual process of internal evaluation and improvement. Although the Consortium had originally offered to pay tuition and books for one early care and education (ECE) class, the response from providers was so overwhelming that they ended up offering two classes. Then the Consortium found that the general education requirement posed a major barrier for moving from Tier 3 or 4. So now the Consortium is offering an English class. Thus, as the administrator summed up the status of the implementation, "I think we've tried everything we thought we were going to or wanted to do, it's still being tweaked."

The administrator of another Consortium explained that early on their plan had included "trying to have college courses to support ongoing, unit-bearing acquisition by our teachers and workforce. And what we know [now] is navigating those systems to be able to offer those courses is very challenging and time-consuming."

Administrators from three Consortia shared that the classroom observations and ratings were the biggest barrier to full implementation. For example, one QRIS administrator shared that they and many other Consortia had struggled with their assessments. Because they were just beginning the assessment piece as of the interview, they would have their first batch of full, complete ratings by the end of June 2014. The administrator of another Consortium explained that their observations and ratings were almost complete; however, they still needed to fulfill the requirements of having external, reliable assessors in their community who could conduct the ERS observations.

Provider Participation

Several Consortia had reached their targets for provider participation; however, most Consortia were adhering to phasing in enrollment, as projected in their initial plans and others reported some unexpected challenges in reaching participation targets.

As of the time of the interviews, at least four Consortia had reached full participation (not including enrollment as part of the augmentation cohort, if applicable). As one QRIS administrator explained, all of the centers and FCCHs they had intended to enroll—"and then some"—had joined their QRIS.

For more than half of the Consortia, participation was not yet complete as of the time of the interviews but was on target according to their plans. A few such Consortia had planned to phase in enrollment over the course of the grant period. For example, a QRIS administrator of one such Consortium that was doing rolling enrollment shared that although the Consortium did not have full enrollment as of the interview, administrators had never planned to have full enrollment by that point. In another Consortium, an administrator shared that they were on target—and if they added a prescribed number of people in each year of the grant, the Consortium expected to reach their target enrollment by the time the grant had ended.

Administrators of three Consortia shared that they had had some recent challenges with enrollment. Two of these Consortia had struggled with recruiting FCCHs in particular. As one administrator explained, they had expected to enroll 55 FCCHs but only had 16 or 17 FCCHs as of the time of the interview. But staff members were actively recruiting over the summer and fall of 2014. They planned to reach out to other counties for ideas on recruitment and hoped to reach their goal by October or November 2014. Similarly, in another Consortium, a subgroup of FCCHs were "not quite ready to join yet…and needed to have more discussions with [administrators] about it. But other than that, we're on target." A third administrator noted that they had underestimated the amount of time recruitment would take, adding, "We think that we'll make our number by next year, but nothing's going exactly as we've anticipated."

Planned Rollout for Publicized Ratings

Most of the Consortia were either holding off on developing a clear plan to roll out the ratings or had only had preliminary discussions about publicizing ratings.

As of June 2014, only one Consortium had made the ratings available to the public. In this Consortium, the public has access to a searchable, online database hosted on the website of the community's resource and referral agency, both to find programs involved in the QRIS and to review their ratings.

Administrators from five other Consortia shared that although they had not yet made their ratings publically available, they had been actively thinking about or working on a plan for the roll-out. All of these Consortia planned to publish ratings at the end of the grant period (i.e., around December 2015) and had been developing a marketing and communications plan in order to achieve that goal. For example, the administrator from one Consortium explained, "And so we will have our scores rolled out at the same time as our communication plan. We think it's very important that folks understand high quality and the nature of the scores before we release them. But we're committed to our communication plan, in concert with releasing scores."

Twelve of the 17 Consortia were either holding off on the plan for publicizing ratings or had had only preliminary discussions about publicizing ratings and were instead focusing their time and energy on other grant requirements. For example, one administrator explained that her Consortium was working on enhancing the level of quality of the programs by providing as much support as possible. Another administrator shared that her Consortium was focusing on getting its operations standardized and stabilized to be able to meet immediate needs, which included processing ratings and issuing grants. Because this Consortium was focusing on those immediate needs, they planned to devote more thought to and effort on publicizing ratings in calendar year 2015.

Seven QRIS administrators expressed concerns about publicizing ratings based on a tool that was not yet validated. For example, one QRIS administrator explained that they understood that publicizing the ratings is a requirement of the grant, but that doing so at that point seemed "premature," because the tool was still being validated and they needed to first make sure that it is measuring what it is intended to measure. Similarly, an administrator from another Consortium noted that the "jury is still out in terms of equating a high rating with high quality." This administrator added that they need to explain the ratings to the public before sharing them. For example, some privately funded programs might not get high ratings if they do not use the measurement tools required of subsidized programs.

Finally, administrators from two Consortia shared their reservations about the idea of consumer choice among families, particularly among parents who use subsidized child care. An administrator from one of these Consortia explained that although he agreed with the theoretical goal of the QRIS enabling parents to become better consumers and driving quality, in reality, the goal was not always attainable. This is particularly the case for parents with children in subsidized programs such as Head Start, who may be on a waiting list and would not necessarily have a choice in child care programs. For example, these parents might only have one Tier 2 site available to them and, therefore, would not have the option to enroll their children in a program with a higher rating. The administrator added that parents have limited choices: "You either put them there or you don't work."

Classroom Observations

An important part of implementing the QRIS and a significant job for the Consortia is conducting classroom observations, using both the CLASS and the ERS. QRIS administrators have voiced their concerns about the many challenges associated with this work. We highlight a few of the issues that were raised during our spring interviews. Specially, we describe Consortia's approach conducting observations, the costs of conducting observations, and the challenges related to the observations.

Observation Protocol: Frequency and Number of Classrooms Observed

The majority of the Consortia were following the Consortia Implementation Guide requirements for the frequency of classroom assessments, though several went beyond the minimum requirements and were conducting annual assessments.

At a minimum, the Implementation Guide requires Consortia to assess centers and FCCHs every other year with the age-appropriate version of the CLASS and the age- and setting-appropriate version of the ERS. According to the interviews with QRIS administrators in May and June of 2014, all of the Consortia were adhering to these minimum guidelines for observations, and nearly three quarters of the Consortia felt this frequency was sufficient. As a QRIS administrator from one Consortium explained, "We're not planning on doing additional assessments until it's required for their rerating, so it will be two years from the original rating date [before we assess again]." Another QRIS administrator shared that "in an ideal world, [programs] would get an independent CLASS one year, and then the next year, an independent ERS. [But] in some cases, they would get both in one year, just based on timelines."

Administrators from three Consortia reported conducting observations more frequently than what is required by the Implementation Guide, however. For example, at least three Consortia reported conducting both observations on an annual basis. One QRIS administrator explained that they planned to assess with ERS and CLASS on an annual basis, unless program changes were so marked that they triggered a second assessment within the same year. When asked why they were assessing annually instead of every other year, the interviewee shared, "We're just going to do it; we only have 18 months left. We want to be able to demonstrate improvement from this current year for our current programs. So I think after working with them a second year, we'll see even better scoring, for some sites. Some sites are already scoring so well, so I don't know how they're going to improve much. They're doing really well."

All Consortia were following the Implementation Guide's minimum guidelines for sampling classrooms at the site level, though again, several conducted more assessments than required.

The Implementation Guide also requires Consortia to observe a sample (i.e., approximately one third) of classrooms at the site level in centers that have multiple classrooms. Most Consortia are reportedly adhering to these guidelines, although at least one Consortium had modified the sampling slightly, based on the age groupings of the classrooms. The QRIS administrator in this Consortium explained that a sample of classrooms would be observed, with some exceptions. For example if a site had two classrooms—one infant classroom and one toddler classroom—the site would have two ITERS done, one in the infant classroom and one in the toddler classroom. However, at least three Consortia are assessing every classroom in multi-classroom centers that are participating in the QRIS—not just a sample. In one of these Consortia, for example, ERS observations are conducted one year and CLASS the alternating year—in every classroom at a site, versus a sample.

The QRIS administrator in one county conducting observations in all classrooms noted that, in terms of their quality improvement effort and plan, they "felt that it made a lot more sense…as long as [they could] do it." This administrator added, "We provide coaching to not only the directors of the site but to each and every teacher, and each and every teaching team. And each and every classroom teaching team creates a quality improvement plan. And it just lends itself to a better approach to that quality improvement plan, to have those scores at the classroom level rather than at the site level."

Cost for Observations

The cost of observations varied considerably among Consortia, from $180 to more than $1,000 per observation.

Consistent with findings from the *Descriptive Study*, QRIS administrators' reports of how much their Consortia pay assessors for CLASS and ERS assessments reveal considerable variability from Consortium to Consortium. Although four counties in one region had decided upon a similar rate to pay their CLASS and ERS assessors, most Consortia had developed their own rates. The cost of the CLASS observations ranged from $180 to more than $1,000 per observation. For example, of the 14 Consortia who shared the cost of their CLASS assessments, nine paid less than $500 for an observation, though several noted that this price did not cover everything. The Consortia with the lowest rate for the CLASS observations, for example, explained that the cost "certainly doesn't include the cost of the backend," such as "the reports." Another Consortium that paid less than $500 per observation noted that the rate did not include the "more global infrastructure costs of scheduling, QA (quality assurance) for the assessments, or training of assessors." Another QRIS administrator that quoted less than $500 for a CLASS observation, however, noted that the rate was inclusive of scheduling, conducting, and scoring the assessment. The rest of the Consortia that shared the cost of their assessments reported that they paid $500 or more per observation: Three Consortia paid $500, one paid $700, and another paid $800. An administrator from one Consortium noted that they have staff members who are trained and reliable on CLASS, so the cost for conducting CLASS assessments is absorbed in personnel costs.

On average, the ERS observations were more expensive than the CLASS observations. Among those 15 Consortia that reported rates for ERS observations, the rates ranged from $250 to $1,300. Although seven of the 15 Consortia paid between $250 and $400, the other eight Consortia reported rates paid $500 or more per ERS observation. Three Consortia paid $500, and five paid between $700 and more than $1,000. Consortia that paid higher rates reported doing so in order to retain assessors, while other Consortia noted that they simply did not have the resources to pay more.

Challenges Experienced With Observations

The most common challenge faced by Consortia in conducting observations was finding, training, and retaining qualified classroom assessors.

During the interviews, QRIS administrators were also asked to share overall challenges that they had experienced related to classroom observations. Such challenges included the cost of the observations, and logistics, such as scheduling. The most commonly cited challenge related to finding or keeping a sufficient number of assessors. Several counties did not have a cadre of trained and reliable assessors to draw from when needed and had to train a pool of people to reliability. A few Consortia struggled because once they did secure reliable assessors, they sometimes "lost them" to other counties that paid higher rates for each observation. As an administrator in one such Consortia explained, "You can't be mad about it because it comes down to business. But it just gets really hard when we actually need to have someone do an observation and they're already booked up because they can make [more money in another Consortium]." Two other Consortia that struggled with getting enough qualified assessors worked with the same contractor and group of assessors; however, because the group was doing both CLASS and ERS observations in two different counties, there were sometimes delays in reporting back to the providers.

Another challenge cited by three different Consortia was alignment between the observations done as part of RTT-ELC and those done for other quality initiative programs in the county or region. For example, one Consortium had partnered with a significant percentage of its Head Start programs for RTT. The beginning of this Consortium's partnership with Head Start coincided with Head Start's recompetition, which resulted in "a lot of movement and transition in terms of sites and number of classrooms, which…made it really challenging to get started in terms of conducting those first ERS and CLASS observations." In another Consortium, the challenges related to alignment with Child Signature Program (CSP) requirements. This Consortium ran into difficulties when they randomly selected classrooms for RTT-ELC, particularly in larger sites. If those randomly selected classrooms were not CSP classrooms, they had not been observed, and therefore the site had to have additional observations done. This administrator explained, "It would have been nice if there were some rules that would allow us to utilize existing CSP assessments for classrooms as the default for rating a site, as opposed to having to do additional classrooms on top of the CSP classrooms." A third Consortium had to balance the requirements of its local assessment schedule, which required assessments on every classroom in a site, with those of RTT-ELC and the RTT-ELC evaluation, and those of CSP and the CSP evaluation.

Finally, a couple of QRIS administrators shared struggles related to buy-in and understanding of the purpose of observations done by external staff. For example, an administrator in one Consortium noted that it was sometimes difficult for programs that were used to being observed by their own staff to get rated by independent assessors. She explained that some programs "were so used to doing their own self-assessments, that when they had an [external] assessor come in, they were actually very [upset] about what scores they received because they thought they were doing very well." In another Consortium, an administrator shared that a few of its child care programs did not have a clear sense of the benefits or expectations of participating in the QRIS program, so that "when the coaches started to go out, there were some sites and classrooms that were very difficult…to start this work and build a relationship." Because some of these sites did not understand the benefits of participating in QRIS, they saw observations as a burden rather than a learning tool.

Although some Consortia experienced some challenges with CLASS observations, ERS observations posed greater challenges to many Consortia.

QRIS administrators also described challenges specific to particular observation tools. For example, a couple of administrators mentioned the fact that the calibration process for CLASS was expected to change, so in addition to the yearly recertification, assessors would have a "six-month recalibration requirement," for which assessors would conduct observations with another CLASS assessor—"ideally from First 5 California" to compare scoring and prevent "drift." This six-month calibration would, of course, have cost implications.

ERS observations posed greater challenges than CLASS, according to QRIS administrators. Although finding reliable observers was sometimes a challenge for both CLASS and ERS, more administrators made specific mention of the challenges inherent in securing a cadre of ERS observers. As one administrator explained, it was much more difficult to get assessors for ERS, because

> *The standards for the state ERS assessors are much higher than for CLASS. The standard with the CLASS is that they have to pass the online reliability test that Teachstone offers and keep that current, whereas ERS assessors have been trained by a coach and have been trained by an author or a nationally certified anchor or one of those people. And then they have to go through constant reliability checks, which involve multiple reliability assessments. So it's a much more arduous process.*

Similarly, an administrator from another county noted that they did have people who were "familiar with ERS" but "the challenge was familiarity with ERS and being calibrated [for reliability] at 85 or 90 percent within a specific period of time," which is the standard agreed upon by the Consortia.

Administrators also discussed the costs inherent in getting staff and/or consultants trained to reliability. As one administrator explained, "No one in the county met the requirements for ERS. Of the five ERS assessors available to the county, four are paid staff." This Consortium had to train its own program staff and "invest very heavily in bringing those folks up to reliability." An administrator from another Consortium shared a similar experience, noting that it was "very, very difficult to get [their staff] to reliability" and that it involved "many, many days of training."

Some Consortia struggled with finding sufficient numbers of reliable assessors for particular age groups, settings, or language of instruction. For example, one administrator shared that their county was allowed to send two or three people to training sessions for various Consortia for each of the ERS tools (i.e., ECERS, FDCRS, ITERS) but only ended up with about three people who became reliable: two people were reliable on the ECERS, one person was reliable on FDCRS, and no one was reliable on the ITERS. Similarly, another Consortium's administrator noted that they had trouble finding age-group specific, calibrated assessors who could observe in the language of instruction. A third administrator noted that staff had struggled with finding reliable ITERS assessors; in fact, at the time of the interview, the Consortium did not have any reliable ITERS assessors.

Strategies to Make Observations More Affordable, Sustainable, and Manageable

Consortia implemented various strategies to make observations more affordable, sustainable, and manageable, including recruiting a large pool of classroom assessors and partnering with other Consortia or agencies in the area.

Despite the aforementioned challenges, QRIS administrators reported various strategies they had used to make the CLASS and ERS observations more affordable, sustainable, and manageable. The most commonly cited strategy was to develop and maintain a large pool of classroom assessors. As the representative from one Consortium advised, "You need to make sure you have a secure cadre of assessors and that the turnover will not be too high so you don't have to keep training people." To those ends, one administrator shared that for CLASS, their Consortium paid for the training for "a lot more people" than they anticipated needing. They offered two different sessions of CLASS training and were "willing to pay recertification fees as needed."

The second most commonly cited strategy to make observations more affordable, sustainable, and manageable was collaboration and partnership among Consortia. Consortia with considerable experience often mentored and supported other Consortia in various ways. As an administrator of one such Consortium explained, "Our assessors are working with other counties. So we basically have the technical expertise because we've been paying for this service for a very long time. So other counties are benefitting from our investments in that respect." Some counties reduce burden and costs by sharing assessors and following the same observation protocol and the same report guidelines. Other Consortia have negotiated similar rates for assessors in the region to reduce the likelihood that assessors will choose to work for other counties. One Consortium contracted with the Environment Rating Scales Institute and hosted an interrater reliability training in the spring of 2014, to which they invited other regional counties to send individuals to be trained. Another Consortium helped other counties that they have been mentoring to bring their assessors up to reliability; the hope is that counties being mentored can follow suit, resulting in a larger pool of assessors who could work throughout the region.

Another strategy mentioned by a few Consortia administrators was to establish unique partnerships with other agencies or institutions, either in their own county or in neighboring counties. For example, one Consortium administrator noted that they had been having difficulty finding reliable ERS assessors. This Consortium was mentoring an adjacent county and had considered reaching out to a local college with an ECE department in that county to take on ERS observations for the next year and a half as a special project. Staff from another Consortium were

working with a representative from their County Office of Education to create a cohort of reliable assessors.

Some Consortia administrators had found that it was more affordable and sustainable to have staff do observations as opposed to hiring consultants or contracted employees. In one such Consortium, conducting classroom observations was a primary role for one of their employees. And another Consortium's administrator explained that when they started the RTT-ELC work, they used contractors for the assessments. They discovered that it was challenging to manage quality assurance with contractors and that it would cost more to continue with that model. So they changed their approach and used full-time employees instead of contractors. They also have an in-house team of researchers who have been trained to support the large number of observations that they have in the coming year; this team will augment the team that is exclusively dedicated to conducting assessments.

On the other hand, a few Consortia reported that they found that outsourcing observations to consultants or contractors was a more effective strategy. An administrator in one such county explained that they felt that using full-time employees was not an effective strategy because of the burnout and drift that might result from doing observations year-round on a daily basis. This Consortium has returned to working with consultants.

Another strategy to make observations more affordable and sustainable was to institute alternative ways of scheduling and conducting the observations. For example, an administrator in one Consortium explained that although they subcontract for both ERS and CLASS observations, their own staff conduct the teacher interviews for ERS, because "that is very time-consuming and the teachers are usually with the children when the assessors do the site visits, so they have to go back at a different time—and that was really unmanageable when we first started. So now we have a streamlined process of having a paper interview where they respond to the questions on paper, and submit that two weeks before their review." This same Consortium also uses alternative scheduling strategies. So, for example, rather than scheduling an exact date with a program, they give the site a two-week window. If the teacher that an observer plans to visit is away from the classroom, the observer can then go to another classroom or site that was scheduled within that two-week block of time. In another Consortium, assessors are trained on both CLASS and ERS and administer both in one day. As the administrator explained, "Instead of having two data collectors in a day, you can do one. It may be a little bit of a longer visit but I think there's some cost savings there." At the time of the interviews, a third Consortium was also looking into ways to have both observation tools administered on the same day, though it might require some modification of the tools—such as completing fewer CLASS cycles or focusing on domains where improvements could be made—raising questions about the comparability of the scores across sites and across Consortia.

A few Consortia discussed specific ways to retain their assessors and therefore reduce the costs of training new people. For example, an administrator in one Consortium made specific mention of paying certain rates, noting that it was "one of the strategies that is vitally necessary—to pay enough to retain the same consistent rater." The administrator from another Consortium noted that when they provide a training opportunity on ERS or CLASS, they require those participants, once they reach reliability, to provide a minimum number of assessments in exchange for the free training. In a third Consortium, the administrator noted that when recruiting people for

training on observational tools, they make a point to find people who will likely stay in the community and maintain "this kind of training and certification" and "carry on this work."

Finally, one Consortium noted that a Web-based database had helped them reduce the costs of observations. Assessors can not only use this database to locate the program's address but also are able to enter the data from the observations directly into the Web-based tool on their iPads.

Local Variation in the Implementation of the Hybrid Rating Matrix

Local variation in implementation of the Hybrid Rating Matrix includes local adaptations to the rating levels, and also local differences in the way the element scores are recorded and documented. For local adaptations to the Hybrid Rating Matrix, Consortia had the option to make modifications to Tiers 2 and 5 of the Hybrid Rating Matrix. During the interviews in spring 2014, QRIS administrators were asked to share any changes they had made or planned to make to either tier. These modifications relate to the comparability of the ratings across Consortia and thus are presented here as context for the validation analysis presented in later chapters. Additionally, Consortia collected the data for the QRIS ratings locally, and some local variation in implementation is due to differences in the way Consortia interpreted Hybrid Rating Matrix requirements for recording and documenting scores.

Modifications to Tier 2

The vast majority of the Consortia were maintaining the common criteria for Tier 2.

As of the interviews, 15 of the 17 Consortia reported they were using the point system and maintaining the common criteria for Tier 2. A few of these Consortia noted that they were not making changes because modifying Tier 2 would add burden or cost. As an administrator from one Consortium explained, "Everything you add to the matrix is either a required element to check or to train upon." An administrator from another Consortium explained, "Adding more indicators would be going away from our philosophy of trying to keep things as simple as possible." An administrator from a third Consortium noted that modifying Tier 2 would incur additional costs for their database.

Administrators of two Consortia explained that they kept Tier 2 as is because they wanted to ensure that the system was achievable for both center-based care and FCCHs. As one administrator explained, "We also evaluated it with respect to our family child care providers because our goal has really not been to make this process look any different for family child care. And so we just felt like the way Tier 2 was written, it really did set achievable improvement for family child care." The administrator from one of the Consortia noted that their staff did not think it made sense to add criteria to Tier 2—"and then in Tier 3, it goes back to the matrix." This administrator added, "We were actually okay with what was in Tier 2." Finally, one of the Consortia's administrators said that one of the reasons that they did not change Tier 2 was the evaluation itself, adding that modifying Tier 2 "would complicate some of the pending analysis that the evaluators would be doing."[5]

[5] This latter point is a valid concern, and we have modified analyses to adjust for alterations that some Consortia have made (see Chapter 5).

As of the time of the interviews, only two of the 17 Consortia indicated they had made changes to Tier 2. One of these two Consortia had changed Tier 2 from a point structure to a blocked structure (as is done for Common Tier 1).[6] The other Consortium modified the requirements for two points on the Minimum Qualifications for the Lead Teacher or FCCH element because they wanted the lead teacher to be familiar with two documents, *California Preschool Learning Foundations (Foundations)* and *California Preschool Curriculum Frameworks (Frameworks)*. This change could potentially affect other rating levels in addition to Tier 2, if the adaptation occurs before points are summed.

Administrators from six Consortia explained that although they had not yet changed Tier 2, they were considering making modifications in the future. One such Consortium was going to make Tier 2 a blocked structure instead of a point structure, beginning in fiscal year 2014–15. An administrator from this Consortium explained that their Consortium did not have any Tier 1 or 2 sites—only Tiers 3, 4, and 5, and "so by nature of participating in our project, you're never going to be lower than a '3.'"

Administrators from two different Consortia mentioned adding requirements around inclusion; one of these administrators also discussed requirements around family engagement and working with English language learners (ELLs). Another administrator explained that they would likely revisit Tier 2's criteria for Developmental and Health Screenings, which states that a Health Screening Form is used at entry, then annually or ensures that vision and hearing screenings are conducted annually. The administrator's rationale for modifying this criterion was that it is cost prohibitive and the matrix does not offer clear instructions about what providers are supposed to do with the Health Screening Form, other than collect it.

Modifications to Tier 5

The majority of the Consortia either had made modifications to Tier 5 or were considering making adjustments to it in the future.

QRIS administrators also shared whether they had made modifications to Tier 5. Although many of the adaptations were implemented as an additional block requirement to reach Tier 5, in some cases criteria were added to obtain 5 points on an element, and in those cases the local adaptation could potentially affect other tiers as well. As of the interviews, 10 of the 17 Consortia were maintaining the criteria to receive 5 points on each element delineated in the Hybrid Rating Matrix. As with Tier 2, the most common reasons for not modifying Tier 5 were to not complicate the process and not add criteria that would create more burden. One of the Consortia's administrators shared that they had considered changing the ERS requirement so that the criteria at Tier 5 were more stringent (i.e., higher than the current requirement of an average overall score of 5.5 or above on all subscales), but the more [they] talked to the evaluators and to different counties, they realized that a 5.5 on the ERS was already "a very high ceiling," and they chose to keep it as is.

As of the time the interviews were conducted, two Consortia had made changes to both Tier 2 and Tier 5. One of these two Consortia had modified the Developmental and Health Screenings

[6] The administrator took the job position after this decision was made and therefore did not know the rationale for changing Tier 2 to block from points.

element because administrators wanted the lead teacher to be familiar with *Foundations* and *Frameworks;* for the Ratios and Group Size element, they wanted all teachers to have training in both the *Foundations* and *Frameworks.*

Among those five Consortia that had only made changes to Tier 5, the types of and rationale for modifications varied. For example, an administrator from one Consortium noted that they had added requirements around working with ELLs and children with special needs because they wanted to support full inclusion in the classrooms. This administrator shared that many of their providers had not yet received training on best practices for working with ELLs and children with special needs and could benefit from such opportunities. Another Consortium had added a provision the ERS subscale that relates to working and partnering with families; if the subscale score was less than 6.0, a quality improvement plan would have to be put into place. The provider would also have to offer links to community-based resources that support families with young children; these resources have to be visible or available in writing from the provider, and the provider must share information on family strengthening protective factors related to social and emotional competence of children.

Three Consortia had made modifications to Tier 5 to reflect or align with other quality initiatives or priorities in the county. For example, one Consortium had added requirements for associate teachers that would support alignment with CSP. The administrator of another Consortium that had added an additional element requiring accreditation noted, "From the beginning, when we did all of our outreach and engagement at the community, the community basically insisted on having accreditation at the top level for both family child care and centers." In a third Consortium in which staff modified Tier 5 so that the lead teacher must participate in CARES Plus, the administrator noted, "We really wanted to make sure that programs and teachers understood how all of these services really wrap in together."

Five of the six Consortia that noted that they might adjust Tier 2 in the future also noted that they might modify Tier 5. As with Tier 2, two of the Consortia's administrators mentioned adding requirements around inclusion. Two other Consortia disagreed with the requirement that providers must use DRDPtech, specifically, in order to earn 5 points for the Child Observation element. Given administrative burden, cost, and the lack of research connecting ERS and positive child outcomes, one Consortium noted that they would consider revisiting the requirements for ERS in the future. One Consortium also planned to discuss the option for an Administrative Credential under the Director Qualifications element,[7] noting that an administrative credential does not necessarily guarantee quality.

Differences in Documentation of Element Scores

There is some variation in Consortia approaches to calculating and recording rating data.

As part of the locally driven approach to California's QRIS, Consortia developed their own local systems to record and store the data collected for element scores and ratings. While some groups of Consortia collaborated to develop shared data systems, other Consortia developed their own

[7] To achieve 5 points for the Director Qualifications element: Master's degree with 30 units core ECE/CD including specialized courses plus eight units management or administration OR administrative credential AND 21 hours professional development annually.

data system or made adaptations to already-existing data systems in their county. As a result, there is variation in the amount and type of information included in the QRIS rating data files, and in the way Consortia interpreted the Hybrid Rating Matrix scoring requirements. For example, the Hybrid Rating Matrix includes minimum criteria to earn 1 point in certain elements, but other elements have no criteria to earn 1 point. For the elements with no criteria to earn 1 point, programs that did not meet criteria for 2 or more points were assigned a default of 1 point by most Consortia, but a few Consortia instead assigned a default of 0 points in such cases. This difference in local interpretation of the Hybrid Rating Matrix could in some cases cause a program to receive a different QRIS rating depending on which Consortium calculated the element scores and rating. Additionally, some Consortia stored raw scores or indicators used to determine element scores, but most Consortia did not do so in the early phase of implementation.

Summary

In summary, although the majority of the Consortia had not yet implemented all of their planned activities or reached their total anticipated number of QRIS participants as of early summer 2014, most were on target with the timeline outlined in their initial plans.

Classroom assessments were one integral piece of these planned activities. The majority of the Consortia were following the RTT-ELC's requirements for the frequency and sampling of classroom assessments, though several went beyond the minimum requirements and were conducting annual observations or expanding the sample size of observed classrooms at a site.

Although Consortia shared common experiences around classroom assessments, there were some notable differences as well. For example, the cost of ERS and CLASS observations varied considerably among Consortia. And although a few Consortia had years of experience with observations, which meant that the process was streamlined, this was not the case for many of the Consortia. Several Consortia voiced common challenges around finding, training, and retaining qualified classroom assessors, particularly for the ERS. To address these challenges and make observations more affordable, sustainable, and manageable, Consortia had implemented a number of strategies, including recruiting a large pool of classroom assessors and partnering with other Consortia or agencies in the area.

Another key planned QRIS activity was the publication of ratings. As of July 2014, only one Consortium had made the ratings available to the public through a searchable, online database, both to find programs involved in the QRIS and to review their ratings. In contrast, most of the Consortia were focusing their time and energy on other grant requirements and were therefore holding off on a plan for rolling out the ratings—or had only had preliminary discussions about publicizing ratings. Almost half of the Consortia expressed concerns about publicizing ratings based on a tool that was not yet validated. These concerns were also shared in the *Local Quality Improvement Efforts and Outcomes Descriptive Study: Final Report* (AIR and RAND 2013). During the site visits to the focal systems in spring 2013, a number of QRIS administrators, R&R agency representatives, providers, parents, and others voiced concerns about the public release of ratings information, as required by the RTT-ELC grants.

In terms of validation, it is important to reiterate that California's QRIS is not one uniform system. For example, although all of the Consortia use a common, five-tiered Hybrid Rating

Matrix, each Consortium is allowed to make local modifications to elements within Tiers 2 and 5, and additional local variation occurred in the way consortia interpreted the Hybrid Rating Matrix. As of June 2014, only two of the 17 Consortia had made changes to Tier 2 or to element requirements for 2 points, although several were considering making modifications in the future. One of these two Consortia had changed Tier 2 from a point structure to a blocked structure (as is done for Common Tier 1). The tendency to keep Tier 2 as a point structure aligns with findings shared in the *Descriptive Study* report, in which the majority of counties AIR and RAND interviewed were in support of the combination scoring system and few counties said that they would have preferred a block system. During the 2014 interviews, many respondents said that the hybrid system was strengths based and would be more inclusive of private providers and FCCHs. In contrast, several Consortia had made changes to Tier 5 or to element requirements for 5 points, and others were considering making future adjustments to elements they thought were not the most appropriate indicators of quality. These adaptations can both affect validation analyses and shed light on elements that particular Consortia deem most important.

Chapter 3. Content Analysis

California is implementing a unique strategy to rate and improve the quality of early childhood programs. The state's QRIS Hybrid Rating Matrix includes five tiers with common elements of program quality that are applied across the state's county-based early childhood systems; at the same time, two of the tiers allow counties to include additional elements if desired.

In this chapter of the report, we analyze the content of California's Hybrid Rating Matrix in light of the most current research about early childhood program quality. The chapter is presented in four major sections. First, we review the Hybrid Rating Matrix elements as compared with those in other state QRISs and delve into the evidence base for the program quality elements in the Hybrid Rating Matrix as well as those for some indicators of quality included in other QRISs. Second, we review the evidence base for the program quality assessment instruments that are featured in the Hybrid Rating Matrix and for assessment tools used in other systems. Third, we provide a brief review of research on the rating structure itself and highlight information about QRISs that use points, blocks, or a combination of the two approaches. The chapter concludes with a summary of our findings regarding the content of California's QRIS Hybrid Rating Matrix and a discussion of their potential implications for system design.

It should be noted that the validation study consists of two components—the content review (including a review of the evidence base for each component), and the validation analysis (which is based on actual data from programs). This chapter addresses the first part of the validation study; Chapters 4, 5, and 6 present the validation analysis.

The Research Base for the Elements in the Hybrid Rating Matrix and Other Indicators of Quality

California's QRIS Hybrid Rating Matrix has three core domains: Child Development and School Readiness; Teachers and Teaching; and Program Environment: Administration and Leadership. These domains are in turn assessed by seven program quality elements: Child Observation, Developmental and Health Screening; Minimum Qualifications for Lead Teacher or FCCH; Effective Teacher-Child Interactions: CLASS Assessments; Ratios and Group Size; Program Environment Rating Scale(s); and Director Qualifications. In the following discussion, we provide an overview of the Hybrid Rating Matrix elements and place them in the context of other state QRISs. We then define and summarize the research base for each program quality element as well as for other potential elements that are not currently included in the Hybrid Rating Matrix.

The Hybrid Rating Matrix (Exhibit 3.1), as the name suggests, is the tool used for determining the rating level of an early learning and development program. The RTT-ELC Quality Continuum Framework also includes an accompanying document, the Continuous Quality Improvement Pathway, which in turn includes Core Tools and Resources, with a number of resources that the Consortia are required to include in their Quality Improvement Plan. Although some of the tools and resources cited in the Core Tools and Resources document are also included in the Matrix, others, such as the Center on Social and Emotional Foundations for Early Learning (CSEFEL) Teaching Pyramid Overview, *Preschool English Learner Guide*, USDA Child and Adult Care Food Program Guidelines, and Strengthening Families Five Protective Factors are not included in the rating structure.

Exhibit 3.1. California RTT-ELC Quality Continuum Framework—Hybrid Rating Matrix With Elements and Points for Consortia Common Tiers 1, 3, and 4

ELEMENT	BLOCK (Common Tier 1) Licensed In-Good Standing	2 POINTS	3 POINTS	4 POINTS	5 POINTS
CORE I: CHILD DEVELOPMENT AND SCHOOL READINESS					
1. Child Observation	☐ Not required	☐ Program uses evidence-based child assessment/observation tool annually that covers all five domains of development	☐ Program uses valid and reliable child assessment/ observation tool aligned with CA *Foundations & Frameworks* twice a year	☐ DRDP 2010 (minimum twice a year) and results used to inform curriculum planning	☐ Program uses DRDP 2010 twice a year and uploads into DRDP Tech and results used to inform curriculum planning
2. Developmental and Health Screenings	☐ Meets Title 22 Regulations	☐ Health Screening Form (Community Care *Licensing form LIC 701 "Physician's Report - Child Care Centers" or equivalent)* used at entry, then: 1. Annually OR 2. Ensures vision and hearing screenings are conducted annually	☐ Program works with families to ensure screening of all children using a valid and reliable developmental screening tool at entry and as indicated by results thereafter AND ☐ Meets Criteria from point level 2	☐ Program works with families to ensure screening of all children using the ASQ at entry and as indicated by results thereafter AND ☐ Meets Criteria from point level 2	☐ Program works with families to ensure screening of all children using the ASQ & ASQ-SE, if indicated, at entry, then as indicated by results thereafter AND ☐ Program staff uses children's screening results to make referrals and implement intervention strategies and adaptations as appropriate AND ☐ Meets Criteria from point level 2
CORE II: TEACHERS AND TEACHING					
3. Minimum Qualifications for Lead Teacher/ Family Child Care Home (FCCH)	☐ Meets Title 22 Regulations [Center: 12 units of Early Childhood Education (ECE)/Child Development (CD) FCCH: 15 hours of training on preventive health practices]	☐ Center: 24 units of ECE/CD[8] OR Associate Teacher Permit ☐ FCCH: 12 units of ECE/CD OR Associate Teacher Permit	☐ 24 units of ECE/CD + 16 units of General Education OR Teacher Permit AND ☐ 21 hours professional development (PD) annually	☐ Associate's degree (AA/AS) in ECE/CD (or closely related field) OR AA/AS in any field plus 24 units of ECE/CD OR Site Supervisor Permit AND ☐ 21 hours PD annually	☐ Bachelor's degree in ECE/CD (or closely related field) OR BA/BS in any field plus/with 24 units of ECE/CD (or Master's degree in ECE/CD) OR Program Director Permit AND ☐ 21 hours PD annually
4. Effective Teacher-Child Interactions: CLASS Assessments (*Use tool for appropriate age group as available)	☐ Not Required	☐ Familiarity with CLASS for appropriate age group as available by one representative from the site	☐ Independent CLASS assessment by reliable observer to inform the program's professional development/improvement plan	☐ Independent CLASS assessment by reliable observer with minimum CLASS scores: Pre-K ▪ Emotional Support – 5 ▪ Instructional Support –3 ▪ Classroom Organization – 5 Toddler ▪ Emotional & Behavioral Support – 5 ▪ Engaged Support for Learning – 3.5	☐ Independent assessment with CLASS with minimum CLASS scores: Pre-K ▪ Emotional Support – 5.5 ▪ Instructional Support – 3.5 ▪ Classroom Organization – 5.5 Toddler ▪ Emotional & Behavioral Support – 5.5 ▪ Engaged Support for Learning – 4

[8] For all ECE/CD units, the core 8 are desired but not required.

ELEMENT	BLOCK (Common Tier 1) Licensed In-Good Standing	2 POINTS	3 POINTS	4 POINTS	5 POINTS
CORE III: PROGRAM AND ENVIRONMENT - Administration and Leadership					
5. Ratios and Group Size (Centers Only beyond licensing regulations)	☐ Center: Title 22 Regulations Infant Ratio of 1:4 Toddler Option Ratio of 1:6 Preschool Ratio of 1:12 ☐ FCCH: Title 22 Regulations *(excluded from point values in ratio and group size)*	☐ Center - Ratio:Group Size Infant/Toddler – 4:16 Toddler – 3:18 Preschool – 3:36	☐ Center - Ratio:Group Size Infant/Toddler– 3:12 Toddler – 2:12 Preschool– 2:24	☐ Center - Ratio:Group Size Infant/Toddler – 3:12 or 2:8 Toddler – 2:10 Preschool – 3:24 or 2:20	☐ Center - Ratio:Group Size Infant/Toddler – 3:9 or better Toddler – 3:12 or better Preschool – 1:8 ratio and group size of no more than 20
6. Program Environment Rating Scale(s) (Use tool for appropriate setting: ECERS-R, ITERS-R, FCCERS-R)	☐ Not Required	☐ Familiarity with ERS and every classroom uses ERS as a part of a Quality Improvement Plan	☐ Independent ERS assessment. All subscales completed and averaged to meet overall score level of 4.0	☐ Independent ERS assessment. All subscales completed and averaged to meet overall score level of 5.0	☐ Independent ERS assessment. All subscales completed and averaged to meet overall score level of 5.5
7. Director Qualifications (Centers Only)	☐ 12 units core ECE/CD+ 3 units management/ administration	☐ 24 units core ECE/CD + 16 units General Education + 3 units management/ administration OR Master Teacher Permit	☐ Associate's degree with 24 units core ECE/CD + 6 units management/ administration + 2 units supervision OR Site Supervisor Permit AND ☐ 21 hours PD annually	☐ Bachelor's degree with 24 units core ECE/CD + 8 units management/ administration OR Program Director Permit AND ☐ 21 hours PD annually	☐ Master's degree with 30 units core ECE/CD including specialized courses + 8 units management/administration, OR Administrative Credential AND ☐ 21 hours PD annually
TOTAL POINT RANGES					
Program Type	Common-Tier 1	Local-Tier 2[9]	Common-Tier 3	Common-Tier 4	Local-Tier 5[10]
Centers 7 Elements for 35 points	Blocked (No Point Value) – Must Meet All Elements	Point Range 8 to 19	Point Range 20 to 25	Point Range 26 to 31	Point Range 32 and above
Infant-only Centers 6 elements for 30 points	Blocked (No Point Value) – Must Meet All Elements	Point Range 7 to 15	Point Range 16 to 21	Point Range 22 to 26	Point Range 27 and above
FCCHs 5 Elements for 25 points	Blocked (No Point Value) – Must Meet All Elements	Point Range 6 to 13	Point Range 14 to 17	Point Range 18 to 21	Point Range 22 and above
Infant-only FCCHs 4 Elements for 20 points	Blocked (No Point Value) – Must Meet All Elements	Point Range 5 to 10	Point Range 11 to 13	Point Range 14 to 17	Point Range 18 and above

[9]Local-Tier 2: Local decision if Blocked or Points and if there are additional elements

[10] Local-Tier 5: Local decision if there are additional elements included

Note: Point values are not indicative of Tiers 1-5 but reflect a range of point values. December 17, 2013

Overview of the Hybrid Rating Matrix Elements

The California RTT-ELC Quality Continuum Framework—Hybrid Rating Matrix, in Exhibit 3.1 above, shows how the seven program quality elements are listed among the three core domains. In addition, the Framework indicates that the first tier is assessed using a block approach, where an early childhood program must meet all of the criteria in order to receive a Tier 1 rating. For Tier 2, local Consortia may determine whether the level will be based on achieving a certain number of points or whether the program or provider must meet all criteria, as required by a block approach. Finally, for Tiers 3 through 5, programs must meet a minimum point range.

Our review of the literature indicates that the *Compendium of Quality Rating Systems and Evaluations* (Tout and others 2010a) and the more recently developed online Compendium offer the most comprehensive review to date of systems across the country. Tout and colleagues (2010a) included systems in 22 states and the District of Columbia in their study, as well as three regional systems—for a total of 26 systems. Since then, many more states have adopted QRISs. The Compendium website includes information for a total of 38 systems, including California's RTT-ELC system in 16 counties, and three single-county-based regional systems. In addition, more states are in the process of developing systems.

Exhibit 3.2 provides a summary of the most common program quality elements, referred to as "quality indicators" in the online Compendium included in the 38 systems and also shows which of the indicators is included in California's RTT-ELC Hybrid Rating Matrix.

Overall, California's Hybrid Rating Matrix explicitly includes three of the five most common quality indicators found in QRISs: staff qualifications; environment; and program administration, management, and leadership.

Although nationally the number of QRISs including family partnership as an element has grown from 89 percent in 2010 to 93 percent in 2014, the Hybrid Rating Matrix does not have a separate indicator for family partnership (QRIS Online Compendium 2014). Family involvement is mentioned in the indicator for developmental and health screenings, and family involvement appears to be included in the Program Environment Rating Scales by virtue of its being one of the subscales in the ERS. Family engagement is also included in the RTT-ELC Continuous Quality Improvement Pathways, a document adopted by the Consortia as a companion to the Hybrid Rating Matrix. This document includes tools and resources listed in the federal application that the Consortia are required to include in their Quality Improvement Plan. In accepting the RTT-ELC funds, the Consortia agreed to adopt the Quality Continuum Framework and its tools and resources. However, their utilization does not count toward points in the Hybrid Rating Matrix.

The Hybrid Rating Matrix does not include a separate element for curriculum, although the Child Observation element refers to using child assessment tool findings to inform curriculum planning. The percentage of states including curriculum in their QRISs has risen from 52 percent in 2010 to 78 percent in 2014 (QRIS Online Compendium 2014). However, the implementation of this element varies greatly. Some states provide a list of recommended curricula; a growing number require alignment of the curricula with state early learning foundations. However, many

of these states have neither an identified curriculum nor have a review process in place to ensure that the curriculum is aligned with educational standards.

Exhibit 3.2. Most Common Quality Indicators in 38 QRIS Systems, 2010 and 2014, as Compared to California's RTT-ELC Hybrid Rating Matrix in 2014

Quality Indicators Included	Percentage of Systems as of 2010	Percentage of Systems as of 2014	Used in California's RTT-ELC Hybrid Rating Matrix?
Staff Qualifications	96%	100%	Yes
Family Partnership and Engagement	89%	93%	Not as separate element, but included in Program Environment Rating Scale and in Pathways
Environment	88%	93%	Yes
Program Administration, Management and Leadership	85%	85%	Yes
Curriculum	52%	78%	Not as separate element, but mentioned in Child Observation
Health and Safety	15%	63%	Developmental screening and health examination required for the child, but no provisions other than in Program Environment Rating Scale for promoting nutrition, exercise, etc. Some aspects of Health and Safety are included in the Pathways.
Ratio and Group size	48%	60%	Yes
Child Assessment	44%	55%	Yes
Accreditation	78%	53%	No
Provisions for Special Needs[11]	34%	50%	No
Continuous Quality Improvement[12]	—	50%	Not in rating matrix, but included in Pathways.
Interaction	—	48%	Yes
Cultural and Linguistic Diversity[13]	30%	33%	Not in rating matrix, but included in Pathways.
Community Involvement[14]	26%	40%	Not in rating matrix, but included in Pathways.

Source: QRIS Online Compendium 2014

[11] However, provisions for special needs are assumed for all programs using the Desired Results Developmental Profile (DRDP) and Desired Results (DR)-associated program standards.
[12] Provisions for Continuous Quality Improvement are included in DRDP and DR-associated program standards.
[13] DRDP and Title 5 standards contain provisions for Cultural and Linguistic Diversity.
[14] DRDP and Title 5 standards contain provisions for Community Involvement.

Health and Safety is defined to include several different components, including individual developmental and health screenings for children participating in early learning and care settings, but it also includes broader program health practices, such as nutrition and exercise promotion. California's requirement for developmental and health screenings, therefore, meet one of the definitions of this element. At least 15 of the 38 QRISs across the states include provisions for developmental screening, of which 11 specifically cite the Ages and Stages protocol (QRIS Online Compendium 2014). However, the Hybrid Rating Matrix does not address the broader definition of Health and Safety, which includes program health practices. The RTT-ELC Continuous Quality Improvement Pathways does have a separate section entitled Health, Nutrition, and Physical Activity. This document also lists tools and resources, such as the *California Preschool Learning Foundations* (*Foundations)* and *California Preschool Curriculum Frameworks (Frameworks)*—Health and Physical Development and the USDA Child and Adult Care Program Guidelines. However, use of these tools is not included in the Hybrid Rating Matrix.

The inclusion of Child Assessment is a key feature of the newer QRIS systems, now present in more than half of QRISs across the states, including California's RTT-ELC QRIS. The increased focus on child outcomes has been driven in part by accountability systems placing increasing attention on the readiness of incoming kindergartners to meet more rigorous K–12 standards (Zellman and Perlman 2008). In addition, federal requirements that RTT-ELC grant recipients conduct QRIS validation studies have led a number of states to focus attention on child assessments.

Overall, the number of elements or indicators included in QRIS systems has increased since 2010, with the addition of two new elements, Interaction and Continuous Quality Improvement, and expanded inclusion of existing elements, with a particular jump for Health and Safety. One indicator that has declined as a feature of QRISs is accreditation, from 78 percent of systems in 2010 to 53 percent in 2014[15] (QRIS Online Compendium 2014).

Evidence Base for the Elements and Other Indicators of Quality

In the following discussion, we briefly define and summarize the research base for each program quality element in the Hybrid Rating Matrix. Then we turn to other indicators of high-quality programs that are not currently included in the Hybrid Rating Matrix but are components of other QRISs or are commonly discussed in reviews of early childhood program quality. Where applicable, we also briefly refer to the approaches used to measure the various elements, with a more comprehensive discussion in the next section, which covers program quality assessment tools.

Core I: Child Development and School Readiness

Child Observations. Observing children has long been considered a benchmark of high-quality early childhood instruction (National Education Goals Panel 1998). Although not often featured

[15] While accreditation is most often viewed as an alternative pathway to a high rating, the Compendium here views it as a system element.

in early QRIS systems, it is present in most of the systems developed over the last five years (AIR and RAND 2013).

Child observations allow teachers to monitor children's progress in order to guide their interactions and instruction. In an early childhood context, authentic assessments are often used to guide teachers' instructional decisions with individual students and promote children's learning. Some research shows that the use of assessment data can enhance children's learning. One study found that K–3 students enrolled in classrooms that use a curriculum-embedded assessment instrument showed greater gains in reading compared to students who were not in such classrooms (Meisels and others 2001). Two studies, including one of infants and toddlers in Early Head Start and another of preschool-age children, provide evidence that early evaluations of literacy skills are feasible and predictive of later achievement (Greenwood and others 2011; Missall and others 2007). These studies point to the possibility of using data to inform teaching practices and curriculum and to improve child outcomes.

Once acknowledging the benefits of child observation, the question becomes what approach or tool works best. One issue concerns whether the particular tools used or specified for child assessment can be used not only to inform and improve instruction, but also to measure child performance for accountability purposes. In the Hybrid Rating Matrix, receiving 2 points in this element calls for the use of an evidence-based child assessment tool that covers all five domains of development. Receiving 3 points requires the use of a tool aligned with California's *Foundations* and *Frameworks.* Receiving 4 points requires using the Desired Results Developmental Profile (DRDP), a tool specifically developed for use in state-contracted child care and preschool programs in California. Anecdotal evidence from teachers (during pilot testing) suggests that teachers appreciate having a research-based tool that maps children's typical development across developmental domains and that is also designed to be inclusive of children with special needs (i.e., DRAccess). However, the tool was initially designed to be used to track children's progress and inform instruction and curriculum planning, and not as an outcome measure, and although the tool has been adapted, some concerns may remain regarding the validity of using tools for a purpose other than the one for which it was developed (AERA, APA, and NCME 1999; Scott-Little, Kagan, and Clifford 2003). As noted in California's RTT-ELC application, the DRDP-School Readiness tool was developed to determine readiness upon kindergarten entry, inform curriculum planning, and as an outcome measure.

Developmental and Health Screenings. Developmental screenings are used to understand and evaluate children's development when they enter an early childhood program. They provide an opportunity to share information about a child between caregivers and family members. Importantly, they also identify children who may have medical (hearing, dental, or vision) needs who may be eligible for special education-related support services; in addition, the screenings may identify children at risk of developmental delay who require consistent follow-up screenings. Research shows that when children are screened using a standardized tool, the identification of developmental delay and referrals increase (Guevara and others 2013). Based on the well-researched premise that early intervention for identified problems prevents later challenges that require more intensive and costly intervention, the American Academy of Pediatrics offers guidelines and research-based tools that can be used to conduct these screenings. The research base demonstrating the positive impact of early intervention on children's progress is expansive. Findings from the National Early Intervention Longitudinal

Study revealed that infants and toddlers who participated in Part C services had greater than expected improvements in their motor, social, and cognitive development. The data also showed that 72 percent to 76 percent of the children also had greater than expected growth in terms of their social relationships, use of knowledge and skills, and their ability to take care of their personal needs (Goode, Diefendorf, and Colgan 2011).

Again, although the benefits of health and developmental screening are widely acknowledged for purposes of providing a pathway for young children to needed early interventions, the question becomes what approach or tool to use for the screening. The Hybrid Rating Matrix requires the use of a Health Screening Form at entry and annually or annual vision and hearing screenings. In addition, programs must use the Ages and Stages Questionnaire (ASQ) or Ages and Stages Questions–Social-Emotional (ASQ-SE) to receive the highest number of points for this element.

Core II. Teachers and Teaching

The Teachers and Teaching section of the Hybrid Rating Matrix includes two dimensions: teacher qualifications and teacher-child interactions. Following, we provide a brief review of the literature on each dimension.

Minimum Qualifications for the Lead Teacher. Decades of observational studies and experimental or quasi-experimental evaluations of specific ECE program models (for example, Perry Preschool Program, Chicago Child-Parent Centers, and specific State Preschool programs) have supported the conclusion that formal preservice early childhood education improves the quality of care delivered in ECE settings and promotes stronger child developmental outcomes (Barnett 2003; Karoly and Zellman 2012). Some research also indicates that children in classrooms led by teachers with formal ECE units or certification in early childhood education have lower rates of grade retention and special education placements in the early elementary years and, ultimately, better outcomes reflected in increased high school graduation, reduced incarceration, and stronger employment histories (Schweinhart, Barnes, and Weikart 1993).

However, positive associations between formal teacher education and children's outcomes or classroom practices have not been replicated in every study (U.S. Department of Education 2010). Most notably, secondary analyses of studies that include a large sample of students and teachers in state-funded prekindergarten programs and Head Start programs did not find a positive impact of bachelor's degree or a bachelor's degree with early childhood coursework (Early and others 2007, cited in U.S. Department of Education 2010). Another study suggests that a bachelor's degree predicts quality in community-based child care centers but not in programs with greater resources, such as prekindergarten programs (Vu, Jeon, and Howes 2008, cited in U.S. Department of Education 2010).

The mixed results on the impact of a bachelor's degree on children's learning has prompted researchers and advocates alike to rethink early childhood teacher preparation programs. Specifically, there is a new emphasis on the importance of focused early childhood coursework linked with high-quality field placements: two features of formal education that research shows often do translate into more effective practice. For example, a study conducted in New Jersey found that teachers with early childhood certification provided higher quality environments and stronger literacy practices than their counterparts with elementary certification (Seplocha and

Strasser 2008). A recent review of 44 studies on coaching by Isner and others (2011) found consistent evidence of positive effects of coaching—in both home and center settings, delivered alone or in combination with other professional development—on observed quality, practices with children, and child language and literacy outcomes. Thus, California's approach to give programs credit for more formal coursework and its specificity with regards to coursework in early childhood education is grounded in some research, but it is important to note that the quality of teachers' preparation, and in particular access to coaching and mentoring, warrants attention. Furthermore, despite the promise of coaching, available research is as yet unable to identify the specific coaching elements (for example, dosage, frequency, topics) that are critical to ensuring its effectiveness (AIR and RAND 2013).

One additional consideration within this program element is California's focus on the qualifications of the "lead teacher." In many early childhood classrooms, the concept of a lead teacher does not reflect the reality of the teachers working together on a daily basis. Some programs have multiple adults who share equal responsibilities in the classroom. Other programs, especially those that provide 12 or more hours of child care a day, may have different adults in a leadership role at different times during the day, one who is responsible for the first several hours and another who takes responsibility later in the day. Research from Colorado's QRIS measure of child-adult ratios noted frequent staffing changes within a given classroom during the day (Le and others 2006). In light of this research, it is worth investigating how California determines whose qualifications to evaluate and how present the lead teacher is for the classrooms that are being rated. However, the RTT-ELC implementation guide does include specific guidelines about defining lead teacher in classrooms that have multiple teachers; the guidelines are based on the National Association for the Education of Young Children (NAEYC) definitions used for accreditation visits.

Teacher-Child Interactions: CLASS Assessments. Teacher-child interactions directly influence children's day-to-day experience in early childhood programs. The quality of interactions include the tone that teachers use when speaking with their students, the language they use to extend children's thinking abilities, and the way they facilitate social interactions. Studies of teachers and students in state-funded prekindergarten programs consistently find that high-quality, responsive, and engaging teacher-child interactions are the most significant predictors of children's developmental outcomes (Howes and others 2008; Mashburn and others 2008). In a study that used the same data set of prekindergarten classrooms to replicate QRIS structure, the researchers found that interactions as measured by the CLASS were related to children's academic language skills and social-emotional skills (Sabol and others 2013).

As indicated by the title of this element, the Hybrid Rating Matrix specifies measurement of teacher-child interactions with a formal program-quality assessment tool, the CLASS. Receiving 2 points calls for familiarity with the CLASS, 3 points for an independent CLASS assessment, and 4 and 5 points for particular minimum CLASS scores.

For a more comprehensive discussion of the research on the validity and reliability of the CLASS for measuring teacher-child interaction, see the section on the Research Base for the Program Quality Assessment Tools in the Hybrid Rating Matrix later in this chapter and in Appendix B.

Core III. Program and Environment—Administration and Leadership

The Program and Environment section of the Hybrid Rating Matrix includes three dimensions: ratios and group size, program environment ratings scales, and director qualifications. A brief review of the literature on each dimension follows.

Ratios and Group Size. Low adult-child ratios and group sizes are designed to increase a teacher's ability to provide responsive care, facilitate more positive peer interactions, and tailor instruction, which may be constrained by large group sizes and high child-adult ratios (Vandell and Wolfe 2000). The NAEYC accredits programs that have a preschool class size of no more than 20 students and a ratio of 1 adult to 10 children and smaller groups and ratios for younger age groups. One literature review indicates that low teacher-child ratios and group size are positively related to teachers' practices, the classroom's social and emotional environment, and children's developmental outcomes (National Research Council 2001). Not all research has come to this conclusion, however. Studies of state-funded prekindergarten did not detect a relationship between ratios and group size and children's outcomes (Mashburn and others 2008; Sabol and others 2013).

Another consideration is the way that ratios are measured within the QRIS context. Research from Colorado's QRIS noted that when ratios are measured at one time point within a two-hour window, the ratios are underestimated as compared to when they are measured at multiple points within an eight-hour window (Le and others 2006). Therefore, it is important to think critically about how this program element is captured in order for it to be valid.

California's Hybrid Rating Matrix addresses ratios and group size with a tiered approach, with 1 point for following state licensing requirements, 4 points essentially conforming to NAEYC accreditation criteria, and 5 points for setting more protective ratios more similar to Early Head Start or Head Start.

Program Environment Rating Scales. Program environment rating scales measure the global quality of an early care and education program, which includes the material aspects of a classroom, daily schedule, health and safety procedures, and some aspects of teacher interactions with students. They aim to capture the many features of a classroom that contribute to a positive environment for children. Few tools measure program environment; California's Hybrid Rating Matrix specifies the use of the ERS, beginning with familiarity with the scale and its use a part of a Quality Improvement Plan for 2 points and independent ERS assessments for 3 to 5 points. Some research shows a positive yet modest association between ERS scores and children's outcomes (Burchinal, Kainz, and Cai 2011), but results are not consistent. A full discussion of the validity and administration of the ERS measures is provided in the following section.

An additional consideration with program environment rating scales is how they are used and here the results are noteworthy. According to the U.S. Department of Education's (2010) systematic review of the effective features of professional development, the use of an observational measure of quality can help to provide specific and articulated goals for quality improvement that are beneficial for teachers (Bryant and others 2009).

Director Qualifications. Directors of early childhood programs often have the dual responsibilities of managing a small organization and serving as instructional leaders for their teaching staff. Studies of administrators in public schools show that principals account for one fourth of a school's total impact on student achievement (Leithwood and others 2004), and these results may also be relevant to early childhood programs. Some studies have found a positive association between director education and staff retention and program quality (Whitebook and Sakai 2004). Beyond the research that links director qualifications with improved program, classroom, or teacher outcomes, professional consensus is that program leaders are a critical pathway toward program quality because of the many decisions they make about elements of quality, such as teacher qualifications, the environment, and curricula.

California's Hybrid Rating Matrix requirements for Director Qualifications increase in points, beginning with 12 units core ECE/CD plus three units of management/administration for 1 point and increasing gradually to a master's degree with 30 units core ECE/CD and eight units of management/administration or an administrative credential for 5 points.

Summary of Elements in the Hybrid Rating Matrix

Our review of the literature suggests that many of the elements of the Hybrid Rating Matrix have been shown to influence general measures of program quality and, in some instances, children's development. The strength of the evidence, however, is not consistent. The evidence is strongest for effective teacher-child interactions. The research base for the other elements—teacher qualifications, director qualifications, and class size/ratios—is not as strong but may promote important pathways toward better children's outcomes. In support of this premise, the NICHD ECCRN (2002) study found that structural factors moderate the relationship between process quality and child outcomes: that is, they increase the likelihood of better interactions (Zaslow and others 2010). In addition, a study by Mashburn and Pianta (2010) found a moderating relationship between structural factors and process quality. Importantly, as will be discussed in Chapter 4 on the Distribution and Reliability of QRIS ratings, the research shows that the way these data elements are measured within the QRIS context can impact their validity in evaluating early childhood program quality.

Other Elements or Indicators of Quality

Curriculum and Instruction

As noted above in Exhibit 3.2, although many QRISs include a separate element entitled Curriculum, the Hybrid Rating Matrix refers to curriculum only in the element entitled Child Observation. The Pathways include California's *Frameworks* and *Foundations*, documents intended to inform instruction, and the DRDP to inform curriculum planning. However, the Hybrid Rating Matrix neither includes a list of recommended curricula nor does it require alignment of curricula with the *Frameworks* and *Foundations*. Also, no review process is specified to determine whether a program has an intentional instructional component.

When looking at the impact of early childhood curriculum on children's development, research shows larger effect sizes when programs have an intentional instructional component (Burchinal and others 2010). Put another way, when implemented with fidelity, curricula can help improve

child outcomes. However, simply having a curriculum and occasionally consulting it, without carefully implementing it as intended, does not have the same effect. A wide variety of curricula are available to promote strong instructional practices, yet not all curricula have an evidence base, so their impact on children's outcomes is uncertain.

Two important features of any given curriculum should be considered when evaluating its impact on children. First, the learning goals the curriculum promotes and the structure of learning activities are important. For example, the Building Blocks curriculum that provides hands-on learning in which carefully sequenced content is introduced and reinforced in a variety of experiences throughout the school day has been shown to improve children's mathematical knowledge, particularly for children at risk of school failure (Clements and Sarama 2007).

Second, an evidence-based curriculum needs to be implemented with fidelity. Research supports curriculum implementation interventions that provide coaching or professional development that target instructional practices intended to support particular domains of children's development, such as literacy, mathematics, and social and emotional development. Specifically, a systematic literature review of professional development strategies reports that curriculum implementation supports have a positive impact on teaching practices and the child outcomes aligned with the curricula (U.S. Department of Education 2010). Thus, evidence-based curricula that are implemented well and supported with coaching or other professional development are a promising element of program quality that can positively impact children's development.

Although research supports the benefits of curriculum implemented with fidelity, little evidence indicates the selection of one curriculum over all others. Hence, the California Early Learning Quality Improvement System (CAEL QIS) Advisory Committee (2010) recommended aligning curriculum with the *Foundations* and *Frame*works as opposed to requiring that programs use one of the curricula on an approved list. The *Foundations* contain curricula and other quality criteria and are aligned with kindergarten and Common Core State Standards. To implement this recommendation, the committee recommended for Tier 1 that the program must have an education plan with a philosophy statement, and for Tier 2 that the program explore integrating the *Foundations* and *Frameworks* in its program, and have an education plan with a developmentally, culturally, and linguistically appropriate (DCLA) curriculum. For Tiers 3 and 4, the committee recommended that programs have an education plan with all domains linked to child assessments and a professional development plan including training on the *Foundations* and *Frameworks*. For the highest tier, the committee recommended that programs must include all domains of learning in an integrated fashion in lesson plans linked to a DCLA curriculum.

The RTT-ELC Hybrid Rating Matrix does not include the CAEL QIS emphasis on curriculum, and the reference in the matrix to the *Foundations* and *Frameworks* is limited to the child assessment and observation tool. There is no progression of requirements related to curriculum in the Hybrid Rating Matrix or in any way measured in the ratings. The RTT-ELC Continuous Quality Improvement Pathways document refers to the *Foundations* and *Frameworks* as a tool and resource related to promoting school readiness; social-emotional development; and health, nutrition, and physical activity. No specific reference or guidance is provided on how their utilization related to selection or implementation of curricula will be measured.

Family Engagement

Although most QRISs include family partnership, engagement, or involvement as a separate element in their systems, the Hybrid Rating Matrix does not. However, the CAEL QIS Committee recommended using the ERS measure for family involvement[16] and the Title 22 licensing requirements related to family engagement as proxies for this element of the rating scale (CAEL QIS Advisory Committee 2010); the Consortia appear to have followed this recommendation in their development of the Hybrid Rating Matrix. In addition, the RTT-ELC Continuous Quality Improvement Pathways document sets a goal of families receiving "family-centered, intentional supports framed by the Strengthening Families Protective Factors to promote family reliance and optimal development of their children."

Family engagement, also referred to as family-sensitive caregiving, addresses the ways that early childhood programs support families and invite their participation in program activities. This program element includes parent-teacher conferences, family dinners, and volunteer opportunities and also may extend further to include practices that accommodate family needs (e.g., flexible schedules and fees) and support families' active participation in their children's learning (e.g., shared literacy resources). This expansive definition of family engagement, which considers the ways that programs are sensitive to the needs of families, is informed by research demonstrating that families have a greater influence than early childhood programs on child development (Bromer and others 2011).

The Chicago Child-Parent Centers found that family engagement is not only an essential component of a high-quality early learning program but is also a key factor associated with more positive student outcomes and greater family involvement in the elementary school years (Miedel and Reynolds 1999). Lopez (2010) identifies three essential components to partner with families effectively: strengthening the family-child bond and acknowledging the primacy of the family in child development; addressing diversity and understanding cultural and socioeconomic variables; and building trust with families by sharing knowledge about child rearing and other topics.

Bromer and colleagues (2011) identify three dimensions of early childhood programs that can affect the degree of family-sensitive caregiving: ECE providers' attitudes, knowledge, and practices. They note that studies demonstrate positive associations between parent participation in school activities and outcomes for prekindergarteners and kindergartners (e.g., Mantzicopoulos 2003; McWayne and others 2004). Bromer and colleagues also describe qualitative studies that identify many ways that FCCHs offer scheduling, financial, and emotional support that helps family functioning. They note, however, that little research has investigated early childhood providers' attitudes or knowledge in relation to children's developmental outcomes.

Although the literature supports the importance of family engagement in early childhood programs, little consensus exists on how best to measure it (AIR and RAND 2013). Faced with this dilemma, some researchers recommend early piloting of any measure of family engagement

[16] Family involvement is in ECERS-R subscale "Parents & Staff," item 38; ITERS-R subscale "Parents & Staff," item 33, and FCCERS-R subscale Parent & Provider, item 35.

and assessing the implications of including it in a QRIS system (Karoly and Zellman 2012). One new measure—the *Family and Provider/Teacher Relationship Quality* measure (Kim and others, 2014)— has recently been developed to assess site staff's knowledge, practices, and attitudes around family engagement.

Children With Disabilities

Although a growing number of QRISs across the nation include provisions for children with special needs in their rating systems, the Hybrid Rating Matrix does not include this element as a separate indicator. However, in its recommendations for the development of a QRIS, the CAEL QIS Committee (2010) saw the requirements for aligning the program with the *Foundations* and *Frameworks* as a proxy for this element because that document, developed with input from many state and national experts, contains criteria on inclusion that are also aligned with kindergarten standards.

Within the early childhood system, two types of programs serve children with disabilities: (1) general early childhood programs and (2) early intervention or early childhood special education programs. It is important to consider the unique ways that both types of programs support the special and wide-ranging needs of children who have (or are at risk of) developmental disabilities. Spiker, Hebbeler, and Barton (2011) identify five elements of programs serving children with special needs that are important when evaluating early childhood program quality: interactions, program features, staff characteristics and program structure, administrative characteristics, and parent partnerships., It is critically important to consider how well general early childhood practices fit the individual needs of children with (or at risk of) disabilities so that they may fully participate in a program's learning opportunities. Early childhood programs serving children with disabilities may have better global quality in general than those that do not, in part explained by teacher education levels and teacher-child ratios (Peth-Pierce 1998). However, little research identifies the specific practices in inclusive programs that lead to better child outcomes.

For a discussion of the instruments used to measure the program quality as related to children with disabilities, see the later section on High-Quality Programs.

Dual Language Learners (DLLs)

In conjunction with the element of Cultural and Linguistic Diversity, some systems have an element focused on supporting young children who are learning both their home language and English. Although this element is not directly included in the Hybrid Rating Matrix, the matrix does require use of the DRDP for higher tiers, and the DRDP assesses the skills and abilities of young dual language learners using the child's home language. The DRDP includes four measures of English language development. Although this element is not more discretely identified in the Hybrid Rating Matrix, a companion document, the RTT-ELC Continuous Quality Improvement Pathways, includes the *Foundations* and *Frameworks*, which address English language development and support for dual language learners. The Pathways also mention the Preschool English Learner Guide as a tool to support the goal of all children receiving individualized instruction and support for optimal learning and development informed by child observation and assessment data.

Studies consistently have found that intentional bilingual instruction for DLL preschool-age children is positively related to their Spanish language and literacy skills without negative impacts on their English language and literacy skills (Durán, Roseth, and Hoffman 2010; Farver, Lonigan, and Eppe 2009). However, in the absence of a structured bilingual program, evidence of the benefits of using Spanish in the classroom is not consistent (Bumgarner and Brooks-Gunn 2011; Burchinal and others 2012).

In a study of two-way immersion (dual language instruction) versus monolingual English instruction for DLL children, Barnett and colleagues (2007) found no effects on English language skills but a positive association with Spanish language skills.

Other research indicates that programs that simultaneously teach content in both English and Spanish support literacy development in English for young DLLs while also supporting home language skills (Castro, Garcia, and Markos 2013). A study conducted in the public school context shows that students who are instructed in two-way bilingual programs (in which children receive instruction in their native language as long as possible) outperform their peers who are in English-only programs (Garcia, Kleifgen, and Falchi 2008). However, a comprehensive literature review identified some value of DLLs' participation in well-regulated early childhood programs for later school success yet was unable to detect specific characteristics of a program (e.g., language of instruction) that contributed to children's language and literacy development (Buysse and others 2013). Further research in the education and care of young children who are emerging bilinguals is needed to specify the instructional and programmatic practices that best support their growth and development. See the later section on High-Quality Programs for a brief discussion of the instruments sometimes used to measure this element.

Cultural Competency

Again, although QRIS planners and implementers considered including Cultural Diversity and Competency as an element, it is not a separate part of the Hybrid Rating Matrix. The CAEL QIS Advisory Committee recommended alignment of the program with the *Foundations* and *Frameworks* and having an education plan with a developmentally, culturally, and linguistically appropriate (DCLA) curriculum as the best approach to this element. In the Consortia's trimming of elements for the Rating Matrix, the *Foundations* and *Frameworks* were moved to the Pathways. The more compact Hybrid Rating Matrix developed by the Consortia does not include that language.

Cultural competency addresses the capacity of early childhood programs and teachers to be responsive to a culturally diverse group of children. Shivers, Sanders, and Westbrook (2011) developed a conceptual framework for analyzing cultural competency that takes into consideration several factors: teacher preparation, organizational quality, process quality as reflected in the environment, curriculum, and everyday practices. The authors note that to date, little evidence shows a relationship between cultural competency and teaching practices or child development outcomes. It is important to note that the lack of evidence does not minimize the importance of cultural competency; rather the measures and research methods for exploring the relationship have not been developed.

See the later section on High-Quality Programs for mention of instruments sometimes used to measure this element.

Health Practices

About two thirds of QRISs include health practices as a quality element in their rating systems. Although the Hybrid Rating Matrix includes Developmental and Health Screenings, these focus on individual children entering or participating in the programs, not the program's overall practices or activities to promote child health. The *Pathways* companion document does specifically include a goal that "children receive support for optimal physical development, including health, nutrition, and physical activity." Tools and resources for addressing this goal include the *Foundations* and *Frameworks*—Health and Physical Development, Infant/Toddler Program Guidelines, California Infant/Toddler *Foundations* and *Frameworks*—Perceptual/Motor, and USDA Child and Adult Care Food Program Guidelines.

Children's health and well-being is essential for their participation and learning in early childhood programs. Illness and injury prevent children from fully engaging in the early learning experiences that may promote social and cognitive development (Hegland and others 2011). Beyond keeping children safe and healthy, early childhood programs can also be structured to promote children's physical well-being. In an analysis of early childhood quality measures, La Paro and others (2012) suggest that the way that an early childhood program promotes health is an important aspect of quality services that warrants further research. For example, the physical characteristics of the early care and education setting, and the degree of access to safe outdoor play, affects the amount of exercise children are able to experience. Although some evidence shows the importance of health practices, Hegland and colleagues (2011) note that effectiveness studies are needed that show a positive relationship between an early childhood program's health practices and children's outcomes.

Summary of Other Indicators of Quality

The literature demonstrates that researchers and practitioners are investigating many elements that may be related to program quality but are not yet included in California's Hybrid Rating Matrix. Based on this review, curriculum implementation and instruction appear to have the strongest evidence base, and it is notable that these elements are not yet included in either the Child Development and School Readiness or the Teachers and Teaching core components. In particular, the QRIS administrators may want to reconsider the CAEL QIS Advisory Committee recommendations that specify alignment with the *Foundations* and *Frameworks* and a progressive set of requirements, such as lesson plans and documentation of curriculum used. Of course, a review process for ensuring this alignment would also need consideration.

Adding specific reference to the *Foundations* and *Frameworks* in the Hybrid Rating Matrix might also help to address some of the other domains not fully represented in the Hybrid Rating Matrix, such as Cultural and Linguistic Diversity, DLL, Cultural Competency, Special Needs, and Health Practices. These other program elements discussed in this section have some theoretical foundation, but either the research does not exist yet to show whether they do or do not link to improvements in teaching practices or children's outcomes, or as will be discussed in the next section, there is little consensus on how to measure the elements. The absence of

evidence does not mean the elements are not important. Adding alignment of curricula with the *Foundations* and *Frameworks* to the Hybrid Rating Matrix would offer a modest approach to addressing these elements.

The Research Base for the Program Quality Assessment Tools in the Hybrid Rating Matrix and Other Assessment Tools

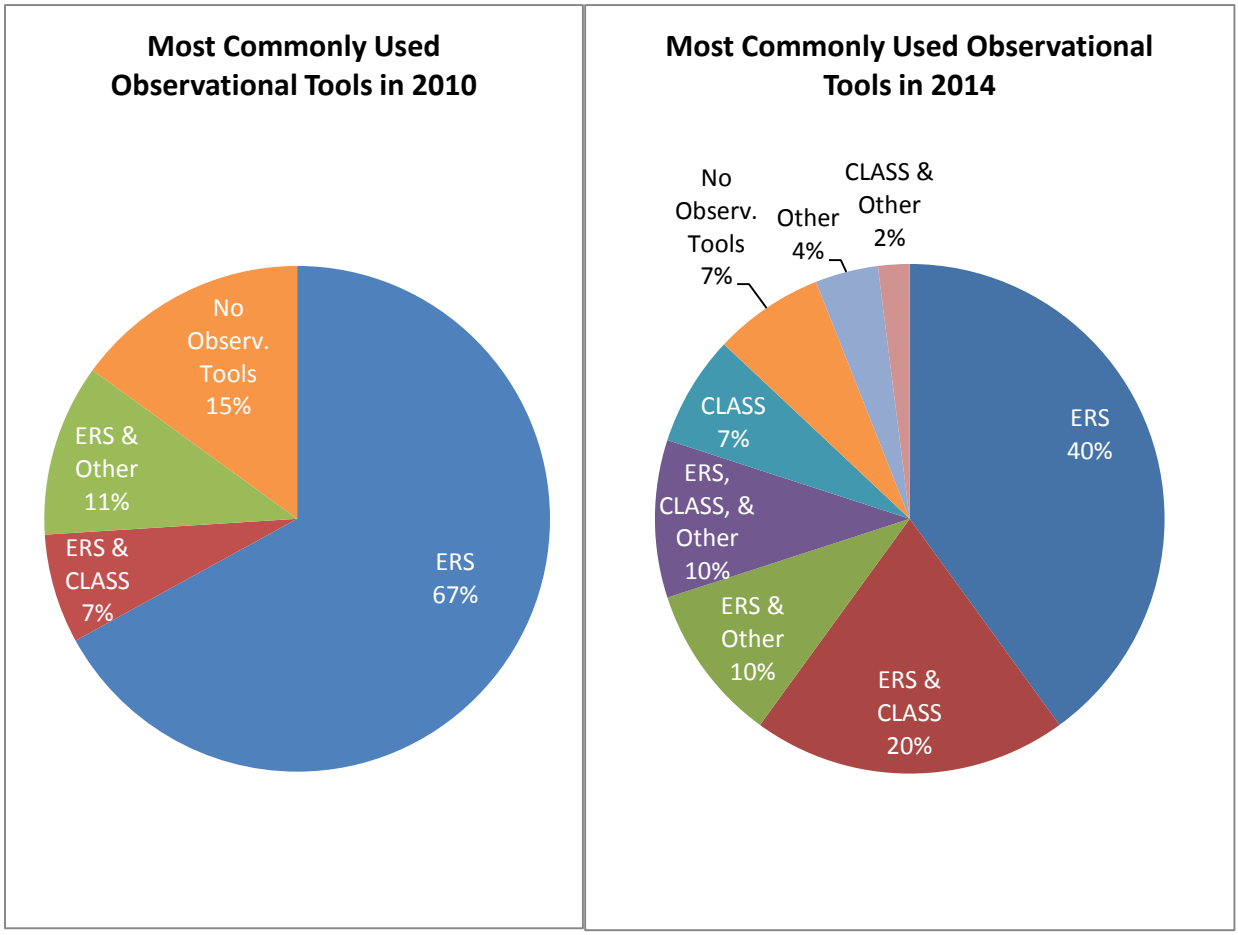
To consider further the content validity of the Hybrid Rating Matrix, we now turn to a discussion of the measurement tools that are used to rate program quality. We briefly discuss the validity, reliability, and research on each tool that is specified within the Hybrid Rating Matrix. We then provide a table that summarizes other tools that measure the constructs within the Hybrid Rating Matrix and tools that measure the other quality constructs discussed in the previous section.

Program Quality Assessment Measures within the Hybrid Rating Matrix

In the following discussion, we describe the research base for the two quality measures included in the Hybrid Rating Matrix: CLASS and ERS.

Exhibit 3.3 shows the most common program quality assessment measures used in QRIS systems and how the utilization of instruments has changed between 2010 and 2014. Overall, the percentage of systems relying on the ERS alone has declined from 67 percent in 2010 to 40 percent in 2014 (QRIS Online Compendium 2014). At the same time, the percentage of systems using both ERS and CLASS (or a third instrument) has increased from 7 percent in 2010 to 30 percent in 2014.

Exhibit 3.3. What Were the Most Commonly Used Tools to Observe Program Quality in 2010 and 2014?



Source: QRIS Online Compendium 2014

The following discussion draws heavily on Halle, Whittaker, and Anderson's (2010) report *Quality in Early Childhood Care and Education Settings: A Compendium of Measures*. We have supplemented the information from the Compendium with more current research on these measures.

Interactions Instrument: CLASS

The CLASS is an observational instrument that was developed to assess classroom quality in preschool through third grade classrooms. The dimensions were derived from a review of constructs assessed in classroom observation instruments used in child care and elementary school research, literature on effective teaching practices, focus groups, and extensive piloting. According to the authors of the tool, it can be used as a research tool, a professional development tool, or as a program development and evaluation tool (Halle, Whittaker, and Anderson 2010).

Interrater Reliability and Internal Consistency. Average interrater reliability on the CLASS is reported as 87 percent, and internal consistency among the CLASS dimensions is high:

Emotional Support (alpha = 0.89); Classroom Organization (alpha = 0.77); and Instructional Support (alpha = 0.83) (Halle and others 2010).

Predictive Validity. The CLASS also demonstrates strong predictive validity. As noted previously in the discussion of teacher-child interaction, studies of teachers and students in state-funded prekindergarten programs consistently find that high-quality, responsive, and engaging teacher-child interactions, as measured by the CLASS, are the most significant predictors of children's developmental outcomes (Howes and others 2008; Mashburn and others 2008). Similarly, in a study that used the same data set of prekindergarten classrooms to replicate QRIS's structure, the researchers found that interactions as measured by the CLASS were related to children's academic and language skills and their social-emotional skills (Sabol and others 2013). The most consistent and robust classroom quality domain for predicting achievement was the Instructional Support of the classroom as assessed by the CLASS. In addition, the Classroom Organization domain has been linked to children's self-control, engagement, and literacy gains (Ponitz and others 2009). For more detail on the validity and reliability of the CLASS, see the full review in Halle and others (2010) compendium.

New Research on Thresholds. Although research has shown a positive relationship between the CLASS and classroom and child outcomes, the strength of these relationships is modest or inconsistent. To explain this, researchers have conducted thresholds analysis to determine if a classroom needs to have a particular score on the CLASS to see a more robust association. In one such study, Burchinal and colleagues (2010) empirically investigated specific thresholds on the CLASS related to children's academic, language, and social skills. They found evidence of threshold effects on both instructional quality and emotional climate subscales. Specifically, when teacher-child interactions were rated in the 5–7 range on the CLASS's Emotional Support scale (the upper quartile of the sample), prosocial outcomes improved and behavioral problems decreased. In terms of academic outcomes, they found that children gained academic skills when the classroom reached a 3.25 on the CLASS Instructional Quality Dimension (the upper 15 percent of the distribution) and that higher quality instruction produces more academic gains. According to the authors, "it is likely that below that point, there is too little explicit instruction or guided child-centered teaching for academic learning to occur" (p. 174). In a more recent analysis of the CLASS, Burchinal and colleagues (2014) found different results with a sample of 1,200 children living in rural areas. Specifically, behavior problems diminished in classrooms that had higher scores in the instruction and classroom organization subscales (at least a score of 5), but no thresholds were detected related to academic outcomes.

Based on this review of the literature on CLASS, there appears to be support both for inclusion of the tool in the Hybrid Rating Matrix and for specifying particular CLASS scores in the higher tiers of the Hybrid Rating Matrix

Program Environment Instrument: ERS

The Early Childhood Environment Rating Scale–Revised (ECERS-R) is intended to measure global quality in center-based early childhood programs. The ECERS-R can be used to measure how well a program meets children's needs, that is, whether children receive the protection, learning opportunities, and opportunities for the positive relationships they need for successful development (Cryer, Harms, and Riley 2003). The ECERS-R is intended for use by researchers,

practitioners, program monitors, and early childhood professionals providing technical assistance to programs. Note again that this discussion draws heavily on Halle, Whittaker, and Anderson (2010). For a more comprehensive review of the ERS, see page 150 of that report.

Interrater Reliability and Internal Consistency. According to the authors as cited in *Quality in Early Childhood Care and Education Settings: A Compendium of Measures,* the ECERS-R is reliable at the levels of indicator, item, and total score (Halle, Whitaker, and Anderson 2010). Across all 470 indicators, the agreement is 86 percent, and all items have agreement of at least 70 percent among indicators. However, the authors note that conflicts of interest by observers can affect the reliability and accuracy of the scores. Total scale internal consistency is 0.92, and subscale internal consistencies range from 0.71 to 0.88 (Harms, Clifford, and Cryer 1998, 2). Again, the authors urge care in interpreting the subscale scores (Halle, Whitaker, and Anderson 2010).

Concurrent Validity. Again, according to the tool authors as cited in *Quality in Early Childhood Care and Education Settings: A Compendium of Measures* (Halle, Whitaker, and Anderson 2010), the total score on the ECERS-R has been found to correlate with two dimensions of the CLASS (Pianta, La Paro, and Hamre 2008): Emotional Climate, $r = 0.52$ and Instructional Support, $r = 0.40$. The total score of the ECERS-R has also been shown to be correlated with the Early Language and Literacy Classroom Organization (ELLCO) (Smith and others 2002) total classroom observation score ($r = 0.41$) and the Literacy Environment Checklist ($r = 0.44$). Finally, the total score of the ECERS-R has also been found to positively correlate with the Caregiver Interaction Scale (CIS) (Arnett 1989) ($r = 0.69$).

Predictive Validity. Research also suggests a positive relationship between the social interaction subscale of the ECERS-R and children's early number concept development (Halle, Whitaker, and Anderson 2010). For example, Clifford, Reszka, and Rossbach (2009) found associations between thee ECERS-R and the Woodcock-Johnson-R mathematics achievement applied problems subset. Clifford and colleagues (2009) also found higher scores on the ECERS-R to be associated with children's development of receptive language, print awareness, and book knowledge and with children's social-emotional development. In addition, the previously mentioned researchers found several subscales of the ECERS-R to be associated with children's scores on measures of independence, concentration, cooperation, and conformity skills in preschool (Clifford, Rezka, and Rossbach 2009).

In more recent analyses of the ECERS-R with national data sets, however, the results have been mixed. Across five studies that used the ECERS or ECERS-R, associations were positive but very modest (0.06 for language, 0.03 for cognitive, and 0.02 for social and emotional development (Burchinal, Kainz, and Cai 2011). In another investigation, Gordon and colleagues (2013) found small effect sizes for regressions predicting child outcomes and moderate effect sizes for regressions predicting teacher-reported quality based on data from the Early Childhood Longitudinal Study Birth Cohort. Researchers using data from state-funded prekindergarten programs found a modest significant relationship between ECERS-R scores and children's language skills (Sabol and others 2013).

New Research on ECERS-R Scoring. A number of critiques of the ECERS-R have to do with the scoring system. Some have challenged the weighting of the ECERS items, particularly as the

tool has been applied to diverse populations (Lambert and others 2008). Other researchers have looked at the ordering of the items to ensure they are sequenced appropriately. Specifically, Gordon and others (2013) found significant disordering; in other words "when observers follow the scoring instructions and assign a low score on an item due to the indicators of one of the dimensions, the higher quality of the center on other dimensions is missed. When observers violate the scoring instructions and assign a higher score on an item due to indicators of other dimensions, the lower quality of the center on one dimension is missed. Both situations may occur in practice." Additional investigations of the ECERS-R indicate that the scoring strategy in which raters stop scoring once a classroom does not meet one criterion is also problematic. One study found that about one quarter of the rated sites would receive higher ratings and be eligible for greater funding when all indicators were taken into account versus when the standard to stop-scoring was used (Hofer 2010).

New Research on Thresholds. Researchers have recently investigated thresholds for ECERS-R and the Infant Toddler Environment Rating Scale-Revised (ITERS-R) based on the notion that there may be a nonlinear relationship between ERS scores and children's outcomes such that classroom quality only impacts development in the higher ranges of quality care. In one study that empirically tested thresholds in the ECERS-R, researchers found evidence of two thresholds: one at a score of 3.4 and the other at a score of 5.4 on the 7-point scale (Le, Schaack, and Setodji 2013). At these points, the researchers found a positive relationship between the ECERS-R scores and a range of cognitive outcomes. Linear regressions failed to detect these relationships. In similar research with the ITERS-R, programs scored within the range of 3.8 and 4.6 showed a positive influence on children's cognitive outcomes: outside of this range, there was no impact (Setodji, Le, and Schaack 2013)

The ECERS-3. The authors of the ECERS-R are currently in the process of revising the instrument, as they have done periodically since the tool was originally developed (Cryer 2014). The new tool responds to the aforementioned research and lessons learned from users. The ECERS-3 (2014) has a greater emphasis on teacher-child interactions throughout the measure and adds numerous items that support young children's preacademic skills, including emerging mathematics, language, and literacy abilities. The new tool also addresses scoring issues, such as the way some items are scaled, eliminating self-reporting, and only including items that an assessor observes. Now that the tool has been released, the authors intend to test its reliability using item-response theory.

Summary

Some research on the CLASS and the ECERS-R shows that they predict modest improvements in children's outcomes, but these relationships are not found consistently across studies, although the evidence base for the CLASS appears stronger than that for the ECERS-R. Importantly, the literature points to some considerations concerning how the tools are applied. New research has found that scores on the tools are not linearly associated with developmental outcomes. As indicated in Exhibit 3.4, when programs reach specific threshold scores, they can see impacts that exceed the modest findings commonly noted in the literature, suggesting that policymakers might want to consider the idea of linking criteria to scoring thresholds. The evidence also suggests that policymakers should consider the implications of having raters score all items on the ECERS-R (as opposed to stopping scoring when a program fails to meet the requirements of

the first item). More research is needed, but this strategy might provide more information to programs about the areas where they are in need of improvement and also give programs credit for their strengths, which is critically important in a high stakes context.

Exhibit 3.4. Impact of Thresholds of CLASS and ERS Scores on Child Development

Instrument	Thresholds
CLASS	Teacher-child interactions that reach a 5–7 on the Emotional Support scale are associated with better social outcomes and fewer behavioral problems.
	Teacher child interactions that reach 3.25 on Instructional Support scale are associated with better mathematics, reading, and language scores.
ECERS-R	Thresholds at 3.4 and 5.4 are associated with better cognitive outcomes. No thresholds were found in terms of social development.
ITERS-R	Programs that score in the range of 3.8 and 4.6 see benefits in children's cognitive outcomes.

Summary Information About Other Quality Measures

Exhibit 3.5 provides information about measures for the quality elements currently captured in the Hybrid Rating Matrix, which might be considered as supplements or even replacements to the current tools. Exhibit 3.6 presents instruments to assess the quality elements, such as curriculum, children with special needs, and dual language instruction, that are not yet included in the Hybrid Rating Matrix but that might be considered. It should be noted that this table does not include all possible measures of quality. Halle and colleagues (2010) provide a detailed review of measures, and much of the information included in the table is drawn from their comprehensive resource. For this chapter, we have selected tools that are intended for external observations and that can be used for the purpose of improving early childhood programs and holding them accountable. For the section of this chapter on possible alternatives or supplemental measures for program elements already included in the Hybrid Rating Matrix, and for the section on Curriculum and Instruction, we focus on tools that already have a considerable track record and have undergone reliability and validity testing. For the section on possible tools for measuring less researched elements, it was necessary to include some instruments that have not yet been fully tested (Halle and others 2010). Although the review may serve to highlight possible alternatives to the tools currently used in the Hybrid Rating Matrix and other tools that could be used to supplement the system, the merits of any additional instrument must be balanced with concerns about imposing additional time or cost burdens on programs and QRIS administrators. Should California's decision makers elect to pursue alternative measures, therefore, more comprehensive consideration of the possible tools would be needed.

Exhibit 3.5. Possible Alternative or Supplemental Measures for Elements in the Hybrid Rating Matrix

Teacher-Child Interactions		
Name	**Description**	**Validity**
Teacher Behavior Rating Scale (TBRS; Landry and others 2001)	An observational tool designed to assess the quantity and quality of general teaching behaviors, language use, and literacy instruction in early childhood classrooms. The prekindergarten TBRS can be used to observe teachers and caregivers of children 3 to 5 years of age and in a variety of early care and education settings. The completion of the TBRS requires two to three hours of observation time while the teachers of interest are with their children.	The validity of the TBRS has been demonstrated with significant correlations (range: 0.25 to 0.40) between TBRS items and teacher self-reports of knowledge. Evidence of convergent validity is seen in multiple instances in which teachers with higher scores on the TBRS also have students who score higher on measures of early literacy.
Program and Environment		
Name	***Description***	***Validity***
The Preschool Program Quality Assessment (PQA; HighScope Educational Research Foundation 2003)	The measure identifies the structural characteristics and dynamic relationships that effectively promote the development of young children, encourage the involvement of families and communities, and create supportive working environments for staff. It is recommended that raters spend at least one full day reviewing a program before completing PQA ratings, allocating a half-day to observing in the classroom (first three sections) and a half-day to conducting interviews (last four sections).	PQA scores are significantly related to children's developmental outcomes, both while children are in preschool and kindergarten, and is associated with established measures of child development (e.g., DIAL-R, High/Scope COR) and teacher ratings. A confirmatory factor analysis identified five factors accounting for 58 percent of the variance and their content aligned with the five corresponding PQA sections.
Program for Infant/Toddler Care Program Assessment Rating Scale (PITC PARS; WestEd Center for Child & Family Studies 2007)	An instrument that measures the extent to which caregiving practices, the care environment, program policies, and administrative structures promote responsive, relationship-based care for infants and toddlers. It uses observations, interviews, and reviews of the programs written materials. The authors recommend conducting observations in the care environment for a minimum of three hours, followed by an interview with a program administrator, and review of written program materials	Correlations between the PITC PARS and the ERS range from 0.81 to 0.88. Correlations with the Arnett Scale of Caregiving Behavior are moderately high. A confirmatory factor analysis identified three factors that are consistent with the structure of the scale.

Program and Environment		
Name	***Description***	***Validity***
Program Administration Scale (PAS; Talan and Bloom 2004)	Reliable and easy-to-administer tools for measuring the overall quality of administrative practices of early care and education programs (PAS) and family child care homes (BAS). It is designed to be a useful guide to improve programs. Formal assessments typically take two hours for an interview with the program administrator and two to four hours for document review.	Concurrent validity was determined through a correlational analysis with two other instruments that measure early childhood organizational effectiveness. Lower and Cassidy (2007) found a statistically significant moderate correlation ($r(54) = 0.291$, $p = .031$) between the PAS and global classroom quality measured by the ECERS-R. A positive correlation ($r(25) = 0.331$, $p = .098$) was also found between the PAS and the Organizational Climate scale of the ECERS-R.

Exhibit 3.6. Possible Tools for Curriculum and Instruction (Element Not Yet in Matrix)

Curriculum and Instruction		
Name	***Description***	***Validity***
The Early Childhood Environment Rating Scale-Extension (ECERS-E; Sylva and others 1998).	The ECERS-E is an extension of the ECERS-R instrument that focuses on aspects of an early childhood curriculum: literacy, science, and mathematics.	The tool has been validated in England, where it was developed. The predictive validity of the ECERS-E in relation to cognitive progress was found to be better than the ECERS-R. The ECERS-E average total was significantly associated in a positive direction with prereading scores, early number concepts, and nonverbal reasoning (Sylva and others 1999).
Early Language and Literacy Classroom Observation (ELLCO) Toolkit Research Edition (Smith and others 2002)	The ELLCO Toolkit is composed of three interdependent research tools. These parts are the Literacy Environment Checklist, completed first as a means to become familiar with the organization and contents of the classroom; the Classroom Observation and Teacher Interview, used second to gather objective ratings of the quality of the language and literacy environment experiences in a classroom; and the Literacy Activities Rating Scale, completed last to provide summary information on the nature and duration of literacy-related activities observed" (Smith and others 2002, p. 1). It takes one to one and a half hours to complete.	The validity of the ELLCO has been demonstrated through its correlations with the Learning Environment subscale of Assessment Profile for Early Childhood Programs (r=.44) and the total score on the Early Childhood Environment Rating Scale [ECERS-R] (r=.41). Other studies showed that scores on the Classroom Observation accounted for 80 percent of the between-classroom variance in vocabulary and 67 percent of the between-classroom variance in early literacy (Dickinson and others 2000 in Smith and others 2002).

Curriculum and Instruction		
Name	***Description***	***Validity***
Early Literacy Observation Tool (E-LOT; Grehan and Smith 2004)	An observation instrument designed to measure research-based instructional practices, student activities, and environmental settings in early childhood classrooms where teachers are engaged in teaching the foundations of reading and other literacy processes. The process takes at least 90 minutes to complete the observation of literacy practices and summary.	Descriptive results examining the relationship between the E-LOT and student achievement suggests a positive correlation between the scores on the observation measure and student achievement. In addition, these descriptive results suggest that the E-LOT converges with the Classroom Observation component of the ELLCO.

Tools are also under development for measuring quality elements that are less researched. Exhibit 3.7 provides information about a few of these instruments. It should be noted that this is just a sample of possible new tools, one in each of the domains of interest: family engagement, children with disabilities, dual language learners, and cultural competence.

Exhibit 3.7. Potential Tools for Measuring Less Researched Elements

Family Partnerships and Engagement		
Name	**Description**	**Validity**
Family and Provider/Teacher Relationship Quality (FPTRQ) (Kim and others 2014)	The FPTRQ assesses the relationships between parents and teachers or providers using three separate measures: a provider/teacher measure, a parent survey, and a director survey, each of which takes 10 minutes to complete. There is also a survey for family service providers and separate survey for parents about their relationship with the family service provider; each takes 10–15 minutes. Constructs measured include knowledge, attitudes, and practices related to family engagement.	The FPTRQ shows some reliability results, but validity studies have not yet been completed.

Children with Disabilities		
Name	**Description**	**Validity**
The Inclusive Classroom Profile (Soukakou 2012).	A structured observation rating scale designed to assess the quality of provisions and daily practices that support the developmental needs of children with disabilities in early childhood settings. The tool is modeled after the format of the Early Childhood Environment Rating Scales and can be administered simultaneously during a 2 or 3 hour-long observation.	The first study leading to validation of the ICP was completed recently. According to the authors, "the measure has acceptable inter-rater agreement, is internally consistent, and shows a good factor structure. Correlations with another measure of global classroom quality (ECERS-R) provided initial evidence for construct validity."
Dual Language Learners		
Name	***Description***	***Validity***
ELLCO Addendum for English Language Learners (ELLCO: Addendum for ELL; Castro 2005)	The measure has been developed as an addendum to the ELLCO to obtain information about the specific classroom practices related to promoting language and literacy development among children who are English language learners. It can also be used as a stand-alone instrument because it is scored separately from the ELLCO. It requires one to one and a half hours to be administered.	This instrument is under development and information about the validity of the measure is not available.
Cultural Competency		
Name	***Description***	***Validity***
Quality Benchmark for Cultural Competence Project (QBCCP; NAEYC 2009)	The NAEYC developed a framework for evaluating the cultural competence of early childhood programs for program discussion and implementation	No information about the psychometric properties of the QBCCP was found.

Summary

Considerable research supports the validity and reliability of the CLASS and support for setting thresholds within tiers that programs and providers have to attain. Like the CLASS, the setting of thresholds for the ERS is a strength of the Hybrid Rating Matrix. Although the ERS has proved difficult to sustain (see Chapter 2) and has some questions regarding its scoring system, deleting it would require replacing it with other tools to measure family engagement, health and safety, and the like that have no better reliability and validity. In addition, inserting these tools might not save time or prove less burdensome. As discussed in Chapter 6 taking into account the difficulty finding adequate number of trained assessors, it may be advisable to consider further limiting the number of tiers for which ERS assessments are required.

Having a methodology to assess curriculum and instruction would be helpful, but it is not clear that adding an additional measurement tool is the best way to approach this goal. Rather having a central monitoring process to determine whether curricula are aligned with the *Foundations* and *Frameworks* might be a more efficient method. Having an instrument to measure family engagement would also be useful, but as we noted earlier in this chapter, the subscale of the ECERS-R only addresses limited aspects of the concept, and there is no other widely used comprehensive measure of family engagement. Although it has not yet been validated, the FPTRQ looks promising and it might be an option for the future. An instrument to measure cultural and linguistic diversity would be helpful, but not enough testing has been conducted on the available instruments to support their addition at this time. The Inclusive Classroom Profile and the ELLCO: Addendum for ELL warrant more consideration as more information on their validity and reliability becomes available.

Rating Structure Analysis

We now turn to a discussion of the strategies that states use to calculate a program's final rating. We share information about the features and implications of calculating ratings using each of the three strategies: a building block approach, a point approach, and a hybrid approach that combines building blocks and points.

The building blocks approach has been the most common rating structure nationwide. However in 2014, as indicated in Exhibit 3.8, the number of QRISs with hybrid structures increased to match the number of systems using blocks, each with 14 systems (QRIS Online Compendium 2014). Ten additional systems relied on a point structure for their ratings.

Exhibit 3.8 State and Local QRIS Rating Structures in 2014

Rating Structure	Number of State and Local Systems
Building Blocks	14
Points	10
Hybrid	14

Source: QRIS Online Compendium (2014)

California has a hybrid approach, which prioritizes flexibility and variation. As noted previously, the structure uses a block system for the entry level of programs licensed in good standing. Points are used to derive ratings for Tiers 2, 3, 4, and 5 so that programs throughout the state can reach these levels in different ways. Then, layered on top of that, counties have the flexibility to add elements to determine how programs can reach Tiers 2 and 5. Finally, participating county Consortia have the option to use a block rating structure for Tier 2, so some counties may use a block approach, and others may use points. Thus, a Tier 2 site can convey common quality in one county that uses a block system but variable quality in a county that uses points.

The building block approach requires programs to meet all of the criteria in one level before they can move up and attain the next quality rating level. The programs at one level meet the same standards; for instance, all Tier 1 programs employ teachers with at least the minimum credentials required by licensing. This approach has several benefits. First, the consistency helps providers and the public understand the system. Second, it identifies the common minimum criteria, such as health and safety, that QRIS designers think parents believe to be really important and which they may already assume that programs meet. Third, the simplicity adds transparency to the system.

The building block approach, however, may have disadvantages as well. Programs that excel in particular areas are not recognized for their accomplishments. Moving from one tier to the next in certain areas, such as staff qualifications, depends on financial resources that some programs may not be able to acquire, especially if the QRIS does not offer generous financial incentives. It assumes all of the quality elements are equally important. The building block approach may contribute to low ratings for certain categories of programs, such as home-based settings, which families value for particular characteristics, and thereby discourage participation by these providers in the QRIS. From a research perspective, the building block strategy does not provide sufficient data elements for analysts to consider whether a measure provides enough variability to distinguish levels of quality (Lahti and others 2013).

In a point system, every standard or program quality element is assigned a number of points. The total number of points a program receives determines its final quality rating level. The point approach encourages participation by a more diverse group of programs because it does not require uniform compliance with certain elements of the system. QRIS designers can give more weight to the quality elements they believe to be more important by assigning more points in those areas. A point approach also gives researchers the ability to analyze the validity of the system. Researchers can examine the distribution of indicator scores and determine whether a measure provides enough variability to distinguish levels of quality. The research on Minnesota and Virginia's QRIS found skewed distributions in some areas, which can point to weaknesses in the measures themselves or in the strategy used for measurement (Lahti and others 2013). In other words, the point strategy highlighted limitations of the QRIS that QRIS administrators could then address to ensure the QRIS was meeting its objectives.

However, using a point approach has some drawbacks as well. The flexibility can make it difficult to communicate standards at a given level. Furthermore, a point strategy can allow programs to reach high ratings based on their strengths in certain areas while neglecting some aspects of quality that can affect child and family experiences.

The third strategy to derive a rating is to use a combination of building blocks and points: the hybrid approach. States that use a hybrid often use building blocks for the first level and then programs earn points to reach the higher rating levels. QRIS designers that select a hybrid strategy try to maximize the benefits of the building block and point approaches. Many states use a hybrid approach to ensure all rated programs meet some common benchmarks at the lower levels while the higher levels accommodate program variation in meeting higher standards. The complexity may deter some participants or inhibit transparency and hinder accountability.

New research sheds light on the implications of using a building block, point, or hybrid structure. Tout and colleagues (2014) used the Early Childhood Longitudinal Study-Birth cohort (ECLS-B), a national data set that included early childhood program data, to replicate a QRIS and then examined how the rating structure influences the distribution of programs across ratings levels, the linkages of ratings with measures of observed quality, and the scores of individual quality components within each structure. The study revealed that: Whereas fewer than one fifth of programs achieved a Level 3 or 4 in the block structure, more than 70 percent of programs achieved a Level 3 or 4 in the points and hybrid structures. Rating levels produced by each of the three structures were significantly correlated with observed quality as measured by the ECERS-R. However, the points structure was the only structure to produce quality levels in which observed quality was significantly different between each of the levels. The points structure also captured the greatest range of ECERS-R scores with a 1.61 point spread between Level 1 and 4 compared to 0.13 and 1.14 point spreads for the block and hybrid structures respectively, enabling greater score variability in the points structure. Scores across rating levels in the rating structures showed different patterns for specific quality components. For example, some domains, such as Health and Safety and Assessment and Accreditation, scored high regardless of level and structure. However, others, such as Family Partnership, Teacher Qualifications, and Director Qualifications, demonstrate significant differences across structures (Tout and others 2014).

Other research uses the NCEDL-SWEEP data on children who attended state-funded prekindergarten programs to replicate nine states' QRIS. This study found that the rating structure affects whether the rating level is related to differences in child outcomes; only three of the nine models had significant associations between the rating level and at least one measure of children's school-readiness skills (Sabol and others 2013). These findings show that the rating structure—points, building block, or hybrid as well as the scoring decisions applied to each rating criteria—can significantly influence the distribution of programs that reach each level in the QRIS. It also has implications for the content that each rating level conveys.

Summary and Possible Policy Considerations

California's QRIS structure underscores the value of data-driven decision making to inform program improvement. The program quality elements are measured using tools that can guide administrators' and teachers' practices. The content review revealed several important considerations for California's decision makers as they move forward with the implementation of the QRIS.

The first part of the chapter examines the elements included in the Hybrid Rating Matrix. The literature indicates that all elements have a research basis, with the strongest evidence for teacher-child interactions and the program environment. The literature on other quality elements

suggests that California may wish to consider adding an element that relates to curriculum implementation and that family engagement may deserve a place as a separate element, although more consideration of how to measure this dimension comprehensively is needed. Other aspects of quality, such as support for DLLs, cultural competency, and support for children with special needs, are worthy of more specific attention in the broader QRIS framework and perhaps in the Hybrid Rating Matrix itself. A broader definition of health and safety that encompasses program practices to promote exercise and nutrition also deserves consideration.

The second part of the chapter looked at the measures currently used to rate teaching and the program environment. The literature shows that the CLASS and ECERS-R tools can be administered reliably. Further, the tools are modestly predictive of children's outcomes, with the CLASS showing stronger evidence of predictive validity (Sabol and others 2013). There are, however, some important considerations concerning the way that the tools are used to derive program ratings. First, for both tools it is critical to set research-informed thresholds that programs need to attain in order to reach the highest quality levels. Second, researchers have identified issues related to the scoring of the ECERS-R and the sequencing of the underlying factors. Based on this research, policymakers may wish to consider the implications of having raters score all items in the instrument, instead of stopping scoring once a program fails to meet one of the criteria in a dimension; of course, doing so would require consultation with the authors of the instrument because it may not be possible without violating the psychometric properties of the tool. Of course, it is important to note that this chapter is based solely on a literature review. In the end, decisions on whether to continue use of both the CLASS and the ERS in the Hybrid Rating Matrix must also take into account the validation analysis specific to the implementation of the RTT-ELC in California and the experience of state and local QRIS administrators.

The review of other tools does not provide strong support for the addition of any tools to the Hybrid Rating Matrix. Some of the tools that may eventually be good candidates to measure cultural and linguistic diversity, for example, are not available or have not been tested sufficiently for reliability and validity. Furthermore, the possible addition of any further measures cannot be based on the reliability and validity of the tool alone; other factors, such as the cost of training people to administer the assessment, the time needed to administer the assessments, and the burden placed on programs being assessed, are all important. The one slight modification that California might consider is the identification of the ECERS-R subscale as an independent measure of family engagement, but it may be wise to await the validation of the FPTRQ, which may measure this program quality element more comprehensively.

Finally, the review of the research showed that California's decision makers may wish to re-examine its hybrid rating structure. The research demonstrates that using points or blocks can significantly alter the distribution of programs across the rating structure. The flexibility that counties have with respect to using points or blocks at Tier 2 should be reviewed in light of this research to ensure that the Hybrid Rating Matrix is designed to meet its goals.

Chapter 4. Measurement Properties of QRIS Ratings

This chapter explores the distribution of QRIS ratings and element scores and the measurement properties of the QRIS ratings, among programs with full ratings.[17] The purpose of these analyses is to determine how well the ratings function as a measure of program quality. In this chapter, we address the following research questions:

- RQ 1. How effective are the California Common Tiers' structure and components/elements at defining and measuring quality in early learning settings?
- RQ 2. Do point values of each element and the final rating provide meaningful distinctions between programs and program types?
- RQ 3. Do element levels relate to each other in consistent ways (e.g., CLASS or ERS score and their relationship to other elements)?

In this chapter and throughout the report, we present ratings that were simulated using the state's rating criteria, without local adaptations that include blocking criteria at Tiers 2 or 5 in some Consortia, in order to examine more comparable ratings in study analyses.[18] Henceforth, the simulated ratings using unadapted criteria are referred to as California QRIS ratings, while the ratings calculated with local adaptations are referred to as Consortia QRIS ratings. Additional details on the methods used in this chapter can be found in the sidebar to the right and in Appendix B.

Analysis Approaches

- Rating Distributions: Description of the number of programs at each QRIS rating level and element score level
- Predictors of Ratings: Ordinal logistic regression analysis indicating if program characteristics predict ratings
- Internal Consistency: Cronbach's alpha statistics assessing the extent to which the QRIS rating measures a single latent construct of program quality
- Relationships Between Element Scores and Ratings: Descriptive and correlational analysis describing how the element scores relate to each other and the overall rating

Data and Sample

- Analyses in this chapter use QRIS rating data, including ratings and element scores, from 472 programs across the state with full QRIS ratings as of January 2014.
- Programs with full QRIS ratings in 2013 tend to be high in quality and differ from other programs participating in the QRIS in 2013 without full ratings (see Appendix B). Therefore, results of the analyses in this chapter may not apply to the broader range of programs participating in the QRIS.

[17] See definition of full ratings in Chapter 1.

[18] Note that we simulated the QRIS ratings using the Consortia's element scores. In most cases, Consortia used the same criteria for element scores, but two of the Consortia added unique local criteria to the California QRIS criteria for element scores and could not provide raw data to determine element scores without the local criteria. In those two counties, the simulated ratings are not perfectly comparable to other counties.

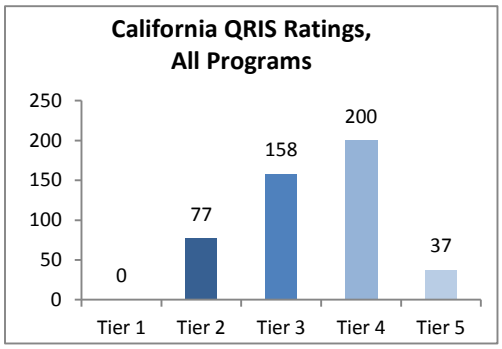
Distribution of Ratings and Element Scores

Among programs with full ratings, the distribution of QRIS ratings is limited and does not span all five possible QRIS rating levels.

Exhibit 4.1 shows the distribution of California QRIS ratings among all 472 programs with full QRIS ratings as of January 2014. Among this sample of programs, no programs were rated at Tier 1 using California QRIS criteria.[19] This indicates that fully rated programs participating in the rating system in 2013 were able to accrue at least enough points to meet the state's Tier 2 criteria. The lack of programs rated at Tier 1 among those with full ratings reflects that QRIS participation is voluntary, and programs may be more likely to participate or finalize their full QRIS ratings if they are eligible for at least Tier 2.[20]

Relatively few programs (8 percent) were rated at Tier 5 using California QRIS criteria, which requires that programs achieve approximately 90 percent of the possible points awarded in the state's hybrid rating matrix (the exact percentage ranges from 88 percent of possible points for family child care to 91 percent of possible points for centers). This suggests that the Tier 5 rating is a high bar for many programs, especially at this stage in the development of the QRIS.

Exhibit 4.1. Distribution of California QRIS Ratings, All Fully Rated Programs



The distribution of ratings differs markedly for centers and family child care homes (FCCH).

Exhibit 4.2 shows the distribution of California QRIS ratings separately for the 365 centers and 107 FCCHs with full ratings as of January 2014. The distributions are quite different by program type. The most common rating for centers is Tier 4, and 86 percent of centers were rated at Tiers 3 or 4. In contrast, the most common rating for a FCCH is Tier 2, and 85 percent of FCCHs were rated at Tiers 2 or 3. Exhibit 4.3 shows the percentage of programs rated at each level among centers, FCCHs, and all programs combined. The differences in the rating distributions between centers and FCCHs are not surprising given that centers tend to have greater access to resources that accrue points on elements, such as professional development supports and child screening

[19] Similarly, less than 1 percent of programs received Tier 1 ratings using the Consortia ratings with local adaptations, i.e. requiring programs to score at least 2 points in each element to receive a rating of Tier 2 or higher.
[20] Among programs with provisional ratings (which were not included in the study analyses because the ratings use data that are not finalized, and thus their ratings are not comparable to the full QRIS ratings), 19 percent were provisionally rated at Tier 1. Programs may seek to earn enough points for Tier 2 before finalizing a full rating. Appendix B provides additional information about provisional and full ratings.

tools. This finding is also consistent with literature suggesting that centers tend to offer higher quality care than FCCH using similar types of quality criteria to the California QRIS ratings (Fuller and others 2004; Li-Grining and Coley 2006; Rigby, Ryan, and Brooks-Gunn 2007).

Exhibit 4.2. Distribution of California QRIS Ratings for Fully Rated Centers and FCCHs

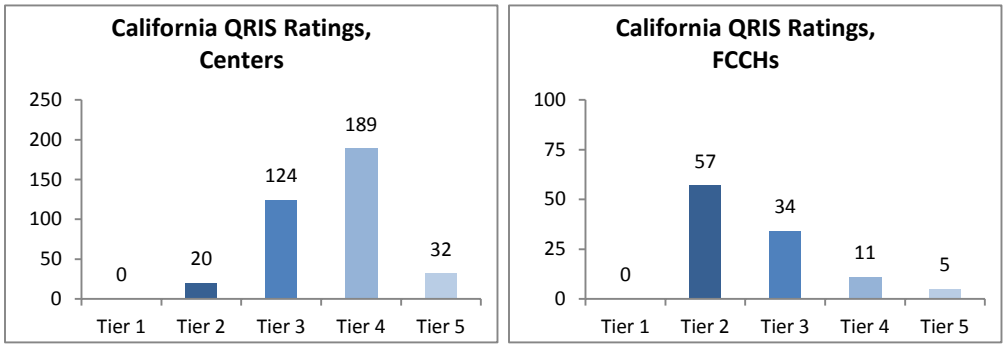


Exhibit 4.3 Percentage of Programs at California QRIS Rating Levels, by Program Type

	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total *N*
Centers	0.0	5.5	34.0	51.8	8.8	365
FCCHs	0.0	53.3	31.8	10.3	4.7	107
All programs combined	0.0	16.3	33.5	42.4	7.8	472

The difference in the distribution of ratings is affected in part by different California QRIS rating criteria for centers and FCCHs. Ratings for centers are determined by the number of points earned in seven quality domains with a total of 35 points, or in six quality domains with a total of 30 points for centers serving only infants (of which there are just two out of 365 in our January 2014 sample). Ratings for FCCHs are determined by the number of points earned in five quality domains with a total of 25 points, or in four domains with a total of 20 points for FCCHs serving only infants (of which there are none in the sample). Accordingly, the minimum number of points required for each rating level also varies by program type.

Still, differences in rating distributions persist between FCCHs and centers when ratings are calculated using the same FCCH rating criteria for both program types, as shown in Exhibit 4.4. In fact, 37 centers that had Tier 4 ratings using the criteria for centers were rated at Tier 5 using criteria for FCCHs, suggesting that the distribution of quality as measured by the California QRIS truly does differ between centers and FCCHs. However, the differences in ratings between centers and homes may be due in part to characteristics of the programs included in these analyses.

The differences in QRIS ratings that are observed between centers and FCCHs in the study sample are specific to programs with full ratings in January 2014—early participants in the state's RTT-ELC QRIS. In the early phases of RTT-ELC implementation, California prioritized enrollment of programs receiving public funding in the QRIS, in response to RTT-ELC guidelines on the inclusion of programs serving high-needs children. As a result, a high

percentage of centers participating in California's QRIS in its early implementation were receiving standards-based public funding, such as State Preschool, Child Signature Program (CSP), or Head Start funding, which require programs to meet specific quality standards. Almost all centers with full ratings in the study sample received standards-based public funding (95.8 percent), compared with fewer than half of FCCHs (42.7 percent). The requirements for standards-based public funding are consistent with some requirements for high scores on California QRIS rating elements, so programs with these funding sources are unlikely to have low QRIS ratings. Exhibit 4.5 shows that low ratings were less common among FCCHs with standards-based public funding than FCCHs without such funding, as is also the case with centers.

Exhibit 4.4. Distribution of California QRIS Ratings Using Family Child Care Rating Criteria for Both Program Types, Centers and FCCHs

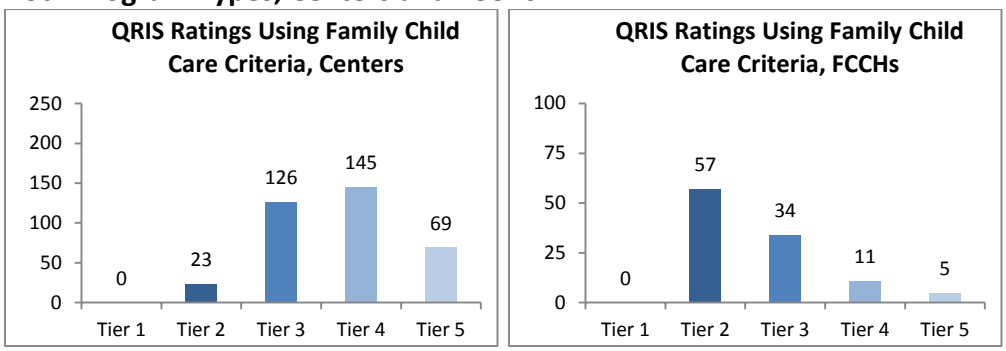
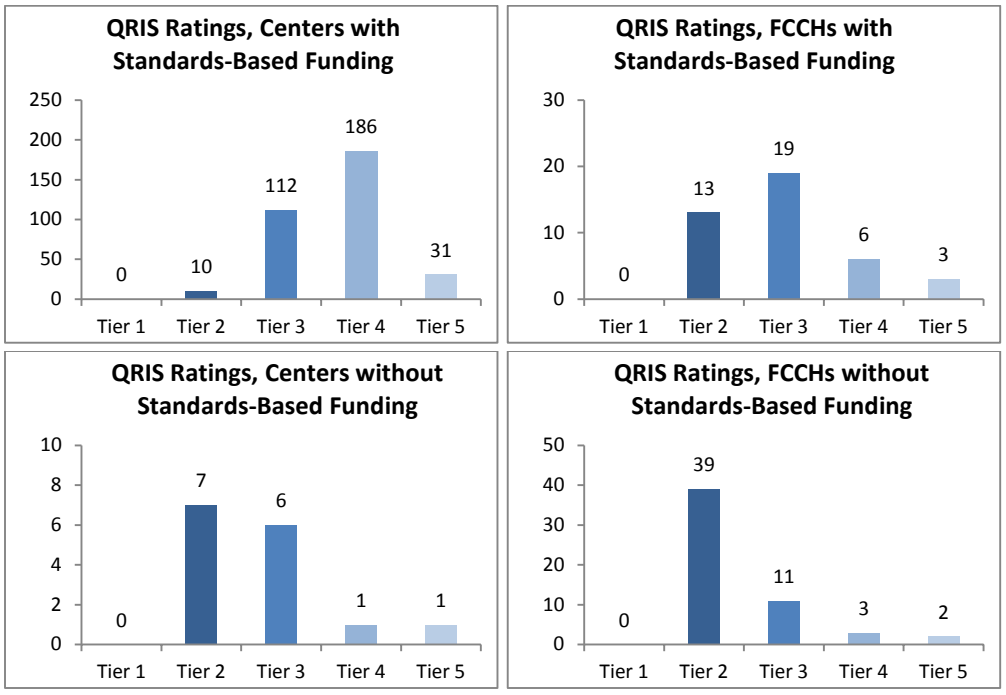


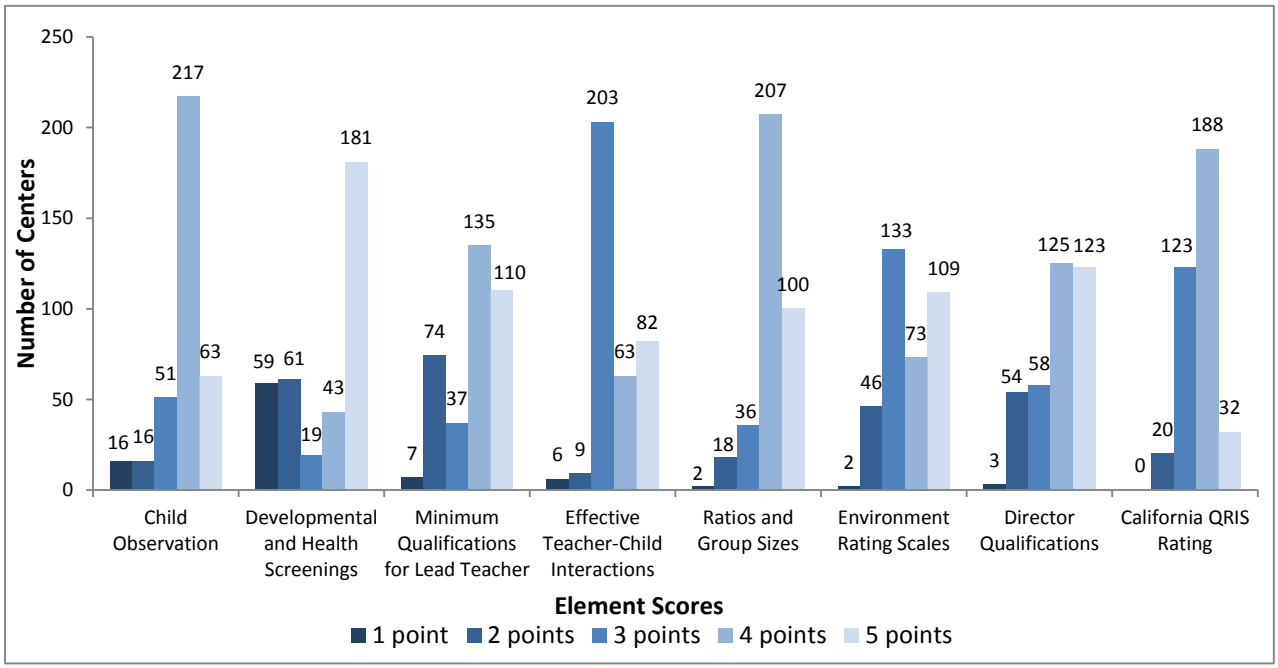
Exhibit 4.5. Distribution of California QRIS Ratings Among Centers and FCCHs, With and Without Standards-Based Public Funding



The charts above show the distribution of ratings among a small number of centers without standards-based public funding (15 centers). It is not known if the patterns of ratings will be similar when a larger and more diverse sample of centers and FCCHs have received full QRIS ratings.

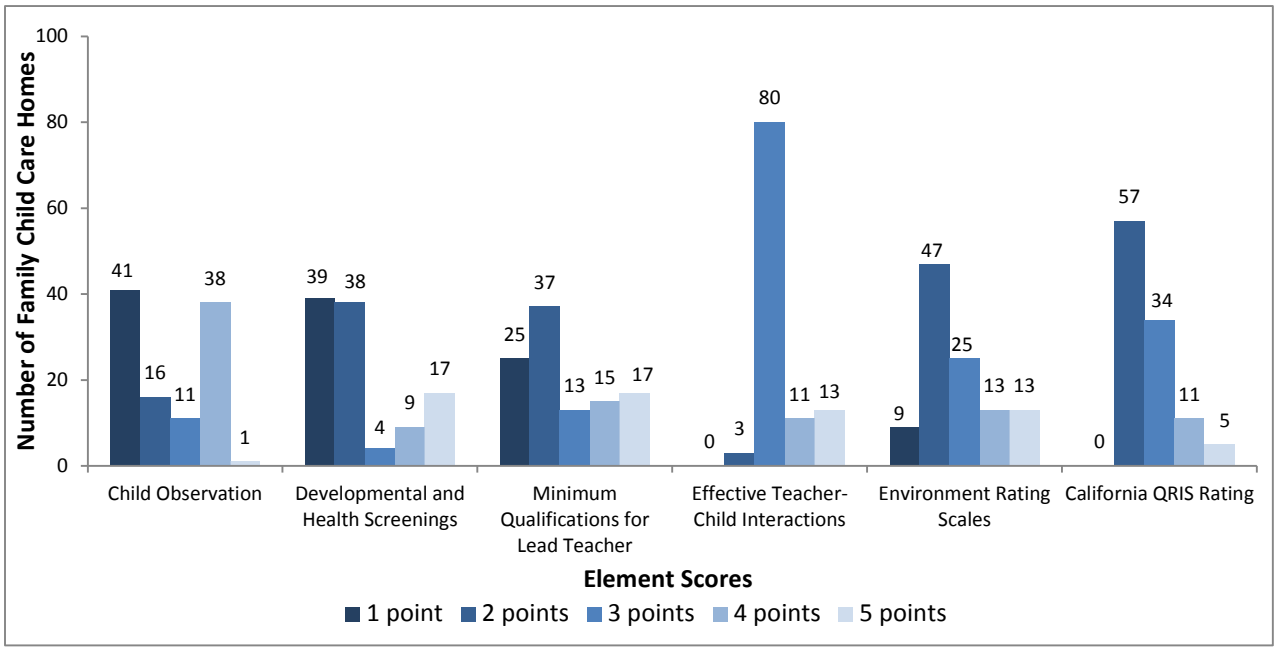
The distribution of element scores differs considerably between centers and FCCHs with full ratings. Exhibit 4.6 shows the distribution of element scores as well as the QRIS ratings for 363 centers with full QRIS ratings (excluding two centers serving only infants because rating criteria are different), and Exhibit 4.7 shows the distribution of element scores and QRIS ratings for 107 FCCHs with full QRIS ratings (none serve only infants). On four of the seven elements that centers are scored on, a large percentage of centers (half or more) receive the same score. Among FCCHs, this occurs in only one element, and the other element scores tend to be more dispersed.

Exhibit 4.6. Distribution of Element Scores, Centers



N = 363 centers, excluding two centers serving only infants

Exhibit 4.7 Distribution of Element Scores, FCCHs



N = 107 FCCHs

In centers, structural quality element scores skew high and thus may not differentiate centers well, while process quality elements have more variation[21]. In contrast, there is more variability in structural quality element scores among FCCHs, but less variability in one of the two process quality elements.

Among centers, element scores skewed high (more than 60 percent of centers received an element score of 4 or 5) for all of the element scores related to structural quality, including child observation, development and health screenings, minimum qualifications for lead teacher, ratios and group sizes, and director qualifications, as shown in Exhibit 4.8. As described previously, the high element scores for centers—particularly those related to structural quality—may be due to the requirements of standards-based public funding received by almost all centers with full ratings in the study sample. The ratios and group size and child observation elements skewed especially high, with more than 75 percent of centers receiving an element score of 4 or 5. This suggests that the rating criteria for these structural quality elements may not differentiate centers well, at least among the study sample of 363 fully rated centers that are not infant-only.

There is somewhat more variability among ratings in the process quality domains, including effective teacher-child interactions (based on the CLASS) and program environment rating scales (based on the ERS). Programs are eligible for element scores of 3 or higher if they receive an independent observation using the respective measures, CLASS and ERS. Almost all centers (96 percent) received CLASS element scores indicating that they had a CLASS observation and

[21] Structural quality refers to easily measurable program characteristics that contribute to high quality, such as staff qualifications, curricula and assessment tools used by the program, adult-child ratios, and group sizes. Process quality refers to interactions between adults and children in classrooms, and includes constructs such as teacher sensitivity and instructional quality, measured by classroom observation instruments such as the PQA and CLASS.

most (87 percent) received ERS element scores indicating that they had an ERS observation. However, there is variability in the element scores since programs must attain specific scores on the CLASS and ERS for element scores of 4 and 5.

Exhibit 4.8. Element Scores and California QRIS Ratings, Percentage of Centers

Element	Percentage of Centers With Element Score or Rating				
	1	2	3	4	5
Child Observation (CO)	4.41	4.41	14.05	59.78	17.36
Developmental and Health Screenings (DHS)	16.25	16.80	5.23	11.85	49.86
Minimum Qualifications for Lead Teacher or FCCH (MQ)	1.93	20.39	10.19	37.19	30.30
Effective Teacher-Child Interactions: CLASS (CLASS)	1.65	2.48	55.92	17.36	22.59
Ratios and Group Sizes (RGS)	0.55	4.96	9.92	57.02	27.55
Program Environment Rating Scales (ERS)	0.55	12.67	36.64	20.11	30.03
Director Qualifications (DQ)	0.83	14.88	15.98	34.44	33.88
California QRIS Rating	0.00	5.51	33.88	51.79	8.82

N = 363 centers (two centers serving infants only are not included)

Among FCCHs, as shown in Exhibit 4.9, element scores are more dispersed on the structural quality indicators (among those that that are applicable to California QRIS ratings for FCCHs: child observation, developmental and health screenings, and minimum qualifications) in comparison to centers. FCCHs in the study sample are less likely than centers to have a teacher or provider with advanced degrees, consistent with other studies of early childhood education programs (e.g., National Survey of Early Care and Education Project Team 2013) and also are less likely to meet high element score criteria for child observation and screenings. In particular, FCCHs are less likely to meet criteria for using specific instruments, including the DRDP for child observation and the ASQ for screenings.

Among the process quality elements, FCCHs have little variability on the effective teacher-child interactions element score (based on the CLASS), but they do vary in scores on the program environment rating scales element score (based on the ERS). Almost 75 percent of FCCHs scored a 3 on the teacher-child interactions element, indicating that most FCCHs participated in CLASS observations (which does not have a specific version for FCCHs) but were unable to score high enough to qualify for 4 or 5 points. In contrast, more than half of FCCHs scored below a 3 because they did not participate in an ERS observation (which does have a specific version for FCCHs). This may be related to challenges expressed by Consortia in obtaining certified ERS observers for the QRIS observations, although the programs included in the study sample are those with full ratings, meaning that programs scoring a 2 on the ERS element are not waiting for an observation to be scheduled and have finalized their full QRIS rating without receiving an ERS observation.

Exhibit 4.9. Element Scores and California QRIS Ratings, Percentage of FCCHs

Element	Percentage of FCCHs With Element Score or Rating				
	1	2	3	4	5
Child Observation (CO)	38.32	14.95	10.28	35.51	0.93
Developmental and Health Screenings (DHS)	36.45	35.51	3.74	8.41	15.89
Minimum Qualifications for Lead Teacher or FCCH (MQ)	23.36	34.58	12.15	14.02	15.89
Effective Teacher-Child Interactions: CLASS (CLASS)	0.00	2.80	74.77	10.28	12.15
Program Environment Rating Scales (ERS)	8.41	43.93	23.36	12.15	12.15
California QRIS Rating	0.00	53.27	31.78	10.28	4.67

N = 107 FCCHs

Characteristics of Programs That Predict QRIS Ratings

The distribution of California QRIS ratings is different for centers and FCCHs with full ratings. Due to the differences in rating criteria for each rating type, predictors of QRIS ratings are evaluated separately for centers and FCCHs.

As described previously, the distribution of ratings differs for centers and FCCHs, even when the same rating criteria are applied to both program types. Program type is clearly a strong predictor of QRIS ratings. However, analyses of the characteristics of programs that predict QRIS ratings are conducted separately for FCCHs and centers because such analyses should not be conducted using ratings calculated with different criteria.

Among centers, descriptive tabulations suggest variation in mean enrollment, languages used in classrooms, funding streams, and Consortium by California QRIS rating level. However, only certain types of public preschool funding are statistically significant predictors of California QRIS rating level among centers with full ratings, controlling for other program characteristics.

Some Consortia had particularly high percentages of centers with high California QRIS ratings. In San Diego and San Joaquin, all centers were rated at Tiers 4 or 5. Among other Consortia, the percentage of programs rated at 4 or 5 ranged from 22 percent to 80 percent. Exhibit 4.10 shows the percentage of programs at each rating level, within each Consortium.

Exhibit 4.10. Percentage of Centers Within Each Consortium at California QRIS Rating Levels

Consortia	*N*	Percentage of Centers at California QRIS Rating Level				
		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
Alameda	12	0.00	16.67	50.00	33.33	0.00
Contra Costa	5	0.00	0.00	20.00	80.00	0.00
Fresno	3	0.00	0.00	33.33	33.33	33.33
LA OCC	14	0.00	0.00	21.43	78.57	0.00
LAUP	76	0.00	9.21	68.42	22.37	0.00
Orange	8	0.00	0.00	37.50	37.50	25.00
Sacramento	15	0.00	20.00	33.33	46.67	0.00
San Diego	76	0.00	0.00	0.00	76.32	23.68
San Francisco	89	0.00	0.00	42.70	55.06	2.25
San Joaquin	13	0.00	0.00	0.00	69.23	30.77
Santa Clara	11	0.00	0.00	36.36	54.55	9.09
Ventura	41	0.00	19.51	24.39	46.34	9.76
Total	363	0.00	5.51	33.88	51.79	8.82

As shown in Exhibit 4.11, serving infants and toddlers appears to be less prevalent among centers with higher California QRIS rating levels. The average enrollment size differs by rating level but does not steadily increase or decrease across rating levels. Use of any language other than English, as well as Spanish specifically, appears to be higher in centers rated at Tier 2 than at other rating levels, while the prevalence is quite similar between Tiers 3, 4, and 5.

A high percentage of centers rated at Tier 5 receive CSP funding as well as Title 5 (State Preschool, General Child Care, or CalSAFE) funding, while fewer than half of centers rated at Tiers 2 and 3 receive these funds. The percentage of programs with Head Start or Early Head Start funding and child care subsidies varies by rating level, but without any apparent pattern of steady increase or decrease across rating levels.

Exhibit 4.11. Program Characteristics by California QRIS Rating Level, Centers

Program Characteristic	*N*	California QRIS Rating Level				
		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
General Characteristics						
Mean enrollment, all ages	360	--	38.2	60.9	50.5	47.4
Percentage of programs serving infants and toddlers	360	--	35.0%	26.2%	15.4%	6.7%
Percentage of programs using any language other than English	348	--	70.9%	55.1%	55.3%	59.4%
Percentage of programs using Spanish	348	--	64.7%	54.2%	53.0%	59.4%
Funding Streams						
Percentage with First 5 California CSP 1 or CSP 2 funding	354	--	41.2%	44.1%	64.2%	87.5%
Percentage with California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	354	--	47.1%	45.8%	75.4%	78.1%
Percentage with Federal Head Start or Early Head Start funding	354	--	11.8%	50.0%	38.5%	21.9%
Percentage with children receiving State/Federally Funded Child Care Subsidy Vouchers	354	--	23.5%	37.3%	28.3%	6.3%
Number of centers at rating level, full sample	363	0	20	123	188	32

Note: Exhibit excludes two centers serving only infants because the rating criteria are different for these centers.

CSP funding and Title 5 (State Preschool, General Child Care, or CalSAFE) funding are significant positive predictors of California QRIS rating level among centers, after controlling for Consortia, enrollment, serving infants and toddlers, and other funding streams (see Exhibit 4.12). This is not surprising because requirements for public funding streams are closely aligned with requirements for high scores on some QRIS elements. Consortium membership also predicted QRIS ratings, but none of the other program characteristics were significantly related to ratings for centers. However, only a limited number of predictor characteristics were included in the models due to data limitations, so these results should be interpreted with caution.

Some combination of public funding types significantly predicted each element score among centers, although the specific funding type varied significantly by element. Enrollment is a significant predictor of several element scores, including child observation, developmental and health screening, and ratios and group sizes, but the magnitude of the relationship is extremely small. Serving infants and toddlers is negatively associated with child observation scores, and using a language other than English is positively associated with developmental and health screenings—but these relationships are difficult to interpret and again should be interpreted with caution due to the limited data available on program characteristics.

Exhibit 4.12. Ordered Logistic Regressions of QRIS Ratings and Element Scores on Program Characteristics, Centers

Program Characteristic	Odds Ratios for Each Dependent Variable in Ordinal Logistic Regression Models							
	QRIS	CO	DHS	MQ	CLASS	RGS	ERS	DQ
General Characteristics								
Enrollment	1.00	1.02**	0.99*	1.00	1.00	0.98***	1.01	1.01
Serves infants and toddlers	0.79	0.43*	1.06	0.55	1.83	1.30	0.71	0.95
Uses language other than English	1.21	0.91	2.35*	0.66	0.91	0.98	1.07	0.89
Funding Streams								
First 5 California CSP 1 or CSP 2 funding	5.31**	33.34***	4.33×10^5	1.39	0.45	1.19	47.70***	0.85
California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	4.01***	2.71**	3.10***	1.27	1.86*	2.53**	2.27**	0.82
Federal Head Start or Early Head Start funding	1.53	3.41***	1.35	0.54*	0.90	1.88*	0.95	1.04
State/Federally Funded Child Care Subsidy Vouchers	0.38	0.83	0.29	0.08***	0.45	1.28	0.44	0.18*

* $p < .05$, ** $p < .01$, *** $p < .0001$

n = 346 centers for all models

Note: Each column represents a separate ordinal logistic regression model, which also included fixed effects for Consortia.

QRIS = California QRIS rating; CO = Child Observation element score; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher or FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; RGS = Ratios and Group Sizes; ERS = Program Environment Rating Scales; DQ = Director Qualifications

Among FCCHs, descriptive tabulations suggest variation in mean enrollment, languages used in classrooms, funding streams, and Consortium by California QRIS rating level. However, none of these characteristics are statistically significant predictors of California QRIS rating level among FCCHs with full ratings.

Although some Consortia had most or all FCCHs rated at Tier 2, other Consortia had no FCCHs rated at Tier 2. In Los Angeles Office of Child Care (LA OCC), Los Angeles Universal Preschool (LAUP), and Alameda, more than 70 percent of FCCHs were rated at Tier 2, while in San Diego, San Francisco, and Santa Clara, all FCCHs were rated higher than Tier 2. Exhibit 4.13 shows the percentage of programs at each rating level, within each Consortium.

Exhibit 4.14 shows that serving infants and toddlers is less prevalent among FCCHs with higher California QRIS rating levels, similar to centers. There is no apparent pattern of mean enrollment by rating level. Use of Spanish during caregiving is less common as ratings increase from Tiers 2

to 4, although all five FCCHs rated at Tier 5 use Spanish. There is no apparent pattern by rating level for specific types of public funding.

Exhibit 4.13. Percentage of FCCHs Within Each Consortium at California QRIS Rating Levels

Consortia	*N*	Percentage of FCCHs at California QRIS Rating Level				
		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
Alameda	5	0.00	100.00	0.00	0.00	0.00
Contra Costa	3	0.00	33.33	33.33	33.33	0.00
Fresno	2	0.00	0.00	0.00	50.00	50.00
LA OCC	38	0.00	71.05	28.95	0.00	0.00
LAUP	21	0.00	85.71	9.52	0.00	0.00
Orange	0	—	—	—	—	—
Sacramento	11	0.00	45.45	36.36	18.18	0.00
San Diego	13	0.00	0.00	46.15	23.08	30.77
San Francisco	12	0.00	0.00	66.67	33.33	0.00
San Joaquin	0	—	—	—	—	—
Santa Clara	2	0.00	0.00	100.00	0.00	0.00
Ventura	0	—	—	—	—	—
Total	107	0.00	53.27	31.78	10.28	4.67

Exhibit 4.14. Program Characteristics by California QRIS Rating Level, FCCHs

Program Characteristic	*N*	California QRIS Rating Level				
		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
General Characteristics						
Mean enrollment, all ages	107	—	8.2	10.5	9.5	9.0
Percentage of programs serving infants and toddlers	107	—	73.7%	44.1%	27.3%	20.0%
Percentage of programs using any language other than English	96	—	65.4%	53.3%	44.4%	100.0%
Percentage of programs using Spanish	96	—	63.5%	53.3%	44.4%	100.0%
Funding Streams						
Percentage with First 5 California CSP 1 or CSP 2 funding	96	—	0.0%	33.3%	44.4%	0.0%
Percentage with California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	96	—	21.2%	20.0%	11.1%	20.0%
Percentage with Federal Head Start or Early Head Start funding	96	—	3.9%	10.0%	22.2%	40.0%
Percentage with children receiving State/Federally Funded Child Care Subsidy Vouchers	96	—	71.2%	70.0%	77.8%	0.0%
Number of FCCHs at rating level, full sample	107	0	57	34	11	5

Apart from Consortium membership, there were few significant predictors of element scores among FCCHs. As shown in Exhibit 4.15, Mean enrollment was negatively associated with scores on Program Environment Rating Scales, and funding types predicted some element scores.

Exhibit 4.15. Ordered Logistic Regressions of QRIS Ratings and Element Scores on Program Characteristics, FCCHs

Program Characteristic	Odds Ratios for Each Dependent Variable in Ordinal Logistic Regression Models					
	QRIS Rating	CO	DHS	MQ	CLASS	ERS
General Characteristics						
Enrollment	1.03	1.09	1.01	1.07	0.94	0.88*
Serves infants and toddlers	0.56	0.25	1.43	0.32	1.53	1.14
Uses language other than English	0.67	1.38	3.02	0.49	0.68	0.90
Funding Streams						
First 5 California CSP 1 or CSP 2 funding	4.12	84.73	0.00	1.35	21.65	18.96
California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	4.41	8.34	655.59**	0.97	0.95	9.72*
Federal Head Start or Early Head Start funding	3.51	7.52**	0.91	0.70	1.27	12.71***
State/Federally Funded Child Care Subsidy Vouchers	0.84	0.48	1.60	1.24	0.43	0.51

* $p < .05$, ** $p < .01$, *** $p < .0001$

$n = 96$ for all models

Note: Each column represents a separate ordinal logistic regression model, which also included fixed effects for Consortia.

QRIS = California QRIS rating; CO = Child Observation element score; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher/ FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; ERS = Program Environment Rating Scales

Internal Consistency

Internal consistency is low, but this is expected given the multi-dimensional nature of QRIS ratings.

Internal consistency measures the extent to which a group of variables (in this case, element scores) produce consistent scores, thereby describing the extent to which they measure a single underlying construct (in this case, program quality). In the context of a QRIS, internal consistency analyses provide information about the extent to which the QRIS ratings are unidimensional in nature (i.e., measuring a single program quality construct). QRIS ratings are designed to measure multiple aspects of program quality, and to assign a single rating reflecting the program's overall level of quality based on these multiple aspects of quality. The use of a

single rating implies that the rating measures one underlying construct of quality, but this may or may not be the case. If the multiple aspects of program quality measured in the QRIS are not strongly related to each other, the internal consistency will be low, indicating that the overall quality level measured by the QRIS rating is not unidimensional. Low internal consistency does not suggest that the QRIS rating is flawed, but rather that the rating does not measure quality in the same way across programs. California QRIS ratings were designed to allow a broad range of definitions of program quality, so a low level of internal consistency is to be expected.

Among the seven domains collected on centers, internal consistency is low ($\alpha = 0.54$) according to criteria used for unidimensional scales, as expected. Generally, a Cronbach's α of 0.8 or higher is considered to indicate a high level of reliability of a unidimensional scale, and between 0.7 and 0.8 is considered to be acceptable for such a scale. Internal consistency of the California QRIS ratings would increase somewhat ($\alpha = 0.67$) if the Developmental and Health Screenings Element and the Ratios and Group Sizes element were both removed from the overall rating and reported separately. Internal consistency is also relatively low among the five domains collected on FCCHs ($\alpha = 0.63$) but would not be increased by removing any element scores. The low levels of internal consistency indicate that the QRIS ratings do not measure a unidimensional program quality construct, especially among centers. In other words, the overall QRIS ratings do not represent a single type of quality, but rather represent diverse types of program quality. These predictably low levels of internal consistency also serve as an important reminder about the likely relationships between QRIS ratings and the observed measures of program quality collected for the concurrent validity analyses. As noted previously, low internal consistency across the multiple rating domains in the QRIS rubric underscores the point that the overall QRIS rating includes different elements of quality that may not be closely related to each other and are not necessarily expected to be. As a result, the relationship between QRIS ratings and the observed quality measures collected for the concurrent validity analyses may not be that strong. We may expect far closer relationships between scores on those elements that measure aspects of quality that relate most closely to the observed measures of program quality collected for the concurrent validity analyses. For example, we can reasonably expect a strong relationship between the effective teacher-child interactions element and the CLASS scores; we might also expect a strong relationship between the minimum qualifications for lead teacher/ FCCH element and the PQA Staff Qualifications and Staff Development scores.

How Element Scores Relate to Each Other and the Overall Rating

The elements included in California QRIS ratings are not redundant; indeed, some pairs of elements have very low correlations.

Among centers, none of the element scores were redundant, indicating that the element scores measure different aspects of program quality. The Ratios and Group Size and Developmental and Health Screening elements had particularly low correlations with other elements (Spearman's ρ below .10 for most element pairs), as shown in Exhibit 4.16, while other pairs of elements had Spearman's ρ correlations ranging from .11 to .46. The percentage of centers with the same number of points earned on two elements ranged from 19 percent (Ratios and Group Size and CLASS) to 44 percent (ERS and CLASS). These percentages are low, but the QRIS is designed to measure diverse aspects of quality, and programs are expected to earn different scores on

rating elements. The low level of overlap in ratings and the low correlations are reflected in the relatively low internal consistency of the QRIS ratings.

Exhibit 4.16. Correlations (Spearman's rho) Among Element Scores, Centers

	CO	DHS	MQ	CLASS	RGS	ERS	DQ
Child Observation (CO)	1.000						
Developmental and Health Screenings (DHS)	0.348*	1.000					
Minimum Qualifications for Lead Teacher or FCCH (MQ)	0.233*	0.077	1.000				
Effective Teacher-Child Interactions: CLASS (CLASS)	0.106*	0.077	0.303*	1.000			
Ratios and Group Sizes (RGS)	–0.030	0.061	–0.128*	–0.081	1.000		
Program Environment Rating Scales (ERS)	0.195*	–0.012	0.305*	0.324*	–0.058	1.000	
Director Qualifications (DQ)	0.351*	0.051	0.464*	0.135*	–0.078	0.149*	1.000

n = 363 centers (excludes two centers serving only infants because the element score requirements are different). Correlations are calculated using Spearman's *ρ*, a nonparametric correlation coefficient that can be interpreted in a similar way to Pearson's *r*.

* $p < .05$

Among FCCHs, again no element scores were redundant. As shown in Exhibit 4.17, the correlations (Spearman's ρ) range from 0.03 (Minimum Qualifications and CLASS) to 0.41 (ERS and CLASS), and the lowest correlations occur between the Effective Teacher-Child Interactions: CLASS and other elements. The percentage of programs with the same number of points earned on two elements ranged from 9 percent (Developmental and Health Screenings and CLASS) to 38 percent (Developmental and Health Screenings and Child Observations). As described previously, these relatively weak relationships between elements are expected in a multidimensional measure of quality such as a QRIS.

Exhibit 4.17. Correlations (Spearman's rho) Among Element Scores, FCCHs

	CO	DHS	MQ	CLASS	ERS
Child Observation (CO)	1.000				
Developmental and Health Screenings (DHS)	0.367*	1.000			
Minimum Qualifications for Lead Teacher or FCCH (MQ)	0.258*	0.228*	1.000		
Effective Teacher-Child Interactions: CLASS (CLASS)	0.095	0.155	0.031	1.000	
Program Environment Rating Scales (ERS)	0.298*	0.371*	0.197*	0.406*	1.000

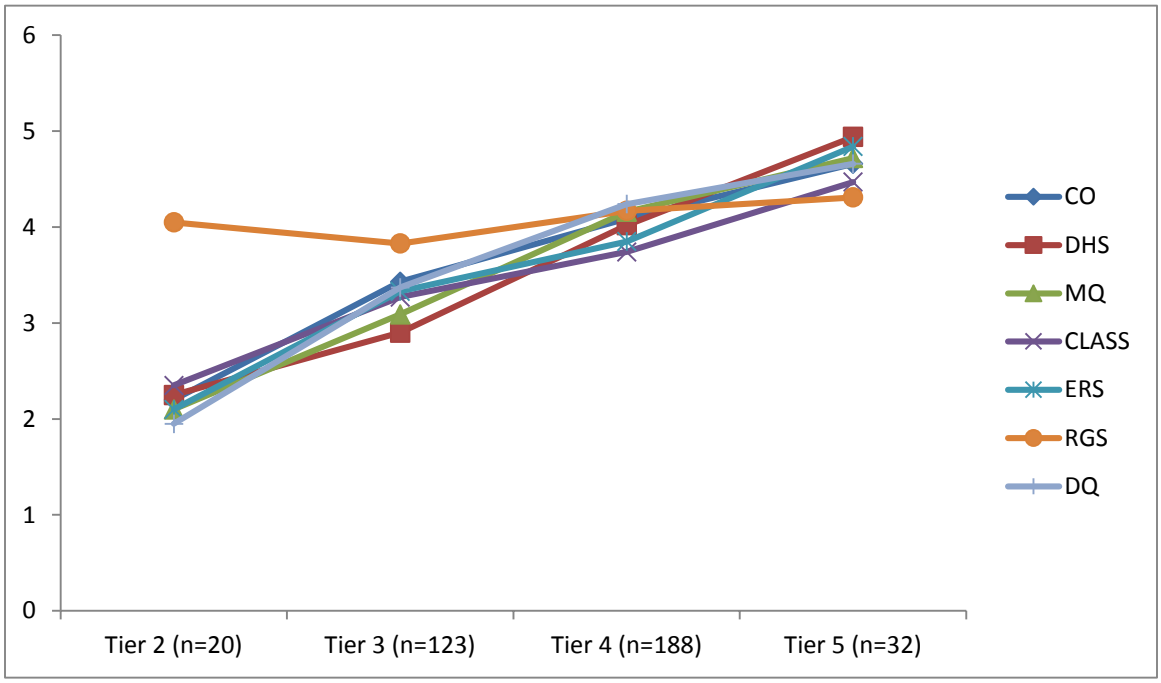
n = 107. Correlations are calculated using Spearman's *ρ*, a nonparametric correlation coefficient that can be interpreted in a similar way to Pearson's *r*.

* $p < .05$

Elements with limited variability tend to be weakly related to the overall QRIS rating and other element scores.

Among centers, the Ratios and Group Sizes element is weakly correlated with the overall QRIS rating (Spearman's ρ = 0.16), while other element scores are moderately correlated with the overall QRIS rating (Spearman's ρ ranging from .45 to .57, see Exhibit D.3). Exhibit 4.18 illustrates this weak relationship, with similar average scores on the Ratios and Group Size element at each rating level (also see Exhibit D.5). Although the other elements have a consistent positive relationship with ratings—as rating level increases, element scores increase— the slope for Ratios and Group Size is much flatter, with limited variation in the element score across rating levels. The Ratios and Group Sizes element also has the least amount of variability (see Exhibit 4.8), which may explain why it detracts from the internal consistency of the QRIS rating and has weak relationships with other elements.

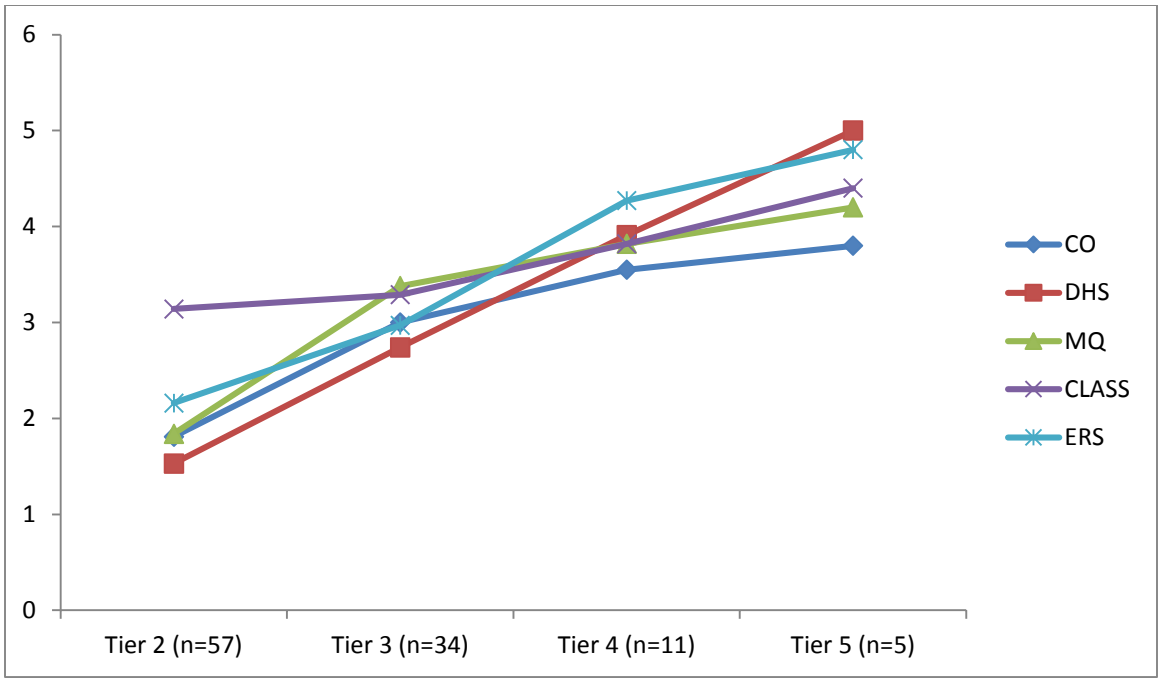
Exhibit 4.18. Average Element Scores by California QRIS Rating Level, Centers



Element score name abbreviations are as follows: CO = Child Observation; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher/ FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; RGS = Ratios and Group Sizes; ERS = Program Environment Rating Scales; DQ = Director Qualifications

Among FCCHs, the Effective Teacher-Child Interactions (CLASS) is the element most weakly correlated with the overall QRIS rating (Spearman's ρ = 0.35), a low to moderate correlation. Exhibit 4.9 shows that this element also has limited variability among FCCHs, which contributes to its relatively weak correlation with QRIS ratings. Exhibit 4.19 illustrates this weak relationship, with limited variability in the average scores on the Effective Teacher-Child Interactions (CLASS) element at each rating level (also see Exhibit D.6). Other element scores are more strongly correlated with the overall QRIS rating (Spearman's ρ ranging from .53 to .62, see Exhibit D.4).

Exhibit 4.18. Average Element Scores by California QRIS Rating Level, FCCHs



Element score name abbreviations are as follows: CO = Child Observation; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher/ FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; RGS = Ratios and Group Sizes; ERS = Program Environment Rating Scales; DQ = Director Qualifications

These results indicate that some elements are not contributing as much to the overall rating. With a more diverse group of programs in the system, we might see more variability among the element scores, which might, in turn improve the degree to which these elements are able to differentiate programs.

Summary: How Well Does the QRIS Perform As a Measure of Quality?

This chapter examined the performance of California's QRIS as a measure of quality. This included a look at the relationship between program characteristics and QRIS ratings, the distribution of ratings and element scores, how element scores relate to each other and the overall QRIS rating, and the internal consistency of the ratings.

As shown in Exhibit B.1, many of the programs with full QRIS ratings are California Title 5 State Contracted Programs (State Preschool, General Child Care, or CalSAFE) or CSP sites. Examination of the characteristics of programs that predict QRIS ratings reveals that only CSP funding and Title 5 (State Preschool, General Child Care, or CalSAFE) funding are statistically significant predictors of California QRIS rating level among centers. And among FCCHs, none of the tested program characteristics significantly predicted QRIS rating level, perhaps due to the limited sample size. The small number of significant predictors may be due to limited data available on program characteristics, and also to the limited distribution of QRIS ratings among programs with full ratings.

Descriptive analyses indicate that the distribution of ratings is limited and does not span all five possible QRIS rating levels. This may be due to the population of programs participating in the system, as it is a voluntary system: as a voluntary system, programs that might score lower have little motivation to become involved. We know that the sample of 472 programs with full ratings differs from other programs voluntarily participating in the system but which have only incomplete or provisional ratings thus far (see Appendix B). In addition, many of the programs with full QRIS ratings are State Contracted Title 5 programs (State Preschool, General Child Care, or CalSAFE) or CSP sites—programs with specific quality requirements for funding—so the QRIS ratings of these programs may be higher than those of other types of early childhood programs in California. In addition, many of the programs that already completed full ratings have a history of participating in other quality improvement efforts in California (prior to RTT-ELC funding and the development of the Hybrid Rating Matrix) and have already had the benefit of significant professional development and quality improvement resources. The limited distribution means that the full range of ratings cannot be fully evaluated.

The distribution of ratings is very different for centers and FCCHs, and the distribution of ratings is even more limited within each program type. In centers, structural quality element scores are skewed high and thus may not differentiate centers well, while process quality elements have more variation. In contrast, there is more variability in structural quality element scores among FCCHs, but less variability in one of the two process quality elements.

Elements with limited variability in scores (such as Ratios and Group Sizes for centers and Effective Teacher-Child Interactions [CLASS] for FCCHs) are weakly related to QRIS ratings, and to other element scores. Internal consistency analyses indicate that the QRIS ratings do not represent a single type of quality, but rather represent diverse types of program quality. As described previously, low internal consistency does not suggest that the rating is flawed, but rather that the aspects of quality measured for the QRIS are not always closely related to each other.

Chapter 5. Concurrent Validity

After exploring the psychometric properties of the ratings in Chapter 4, Chapter 5 turns to an assessment of the concurrent validity of the California QRIS ratings. Evaluating the concurrent validity of the ratings involves comparing the ratings assigned by Consortia to independent measures of quality to see how closely they are aligned. This chapter focuses on the following two research questions:

- RQ 1. How effective are the California Common Tiers' structure and components/elements at defining and measuring quality in early learning settings?
- RQ 2. Do point values of each element and the final rating provide meaningful distinctions between programs and program types?

Concurrent validity analyses are conducted separately for centers and FCCHs because the California QRIS ratings use different criteria for each program type and also because the distributions of ratings are very different for centers and FCCHs, so the results combining both would be difficult to interpret. In addition, the PQA has separate instrument forms for centers and FCCHs that are somewhat different.

Analysis Approaches

- Concurrent validity: Analysis of variance (ANOVA) models for centers and descriptive analyses for FCCHs, examining the average scores on independent measures of observed quality by QRIS rating level or by element score level. ANOVA models for centers indicate whether the average scores differ significantly by rating or element score level.

Data and Sample

- Analyses in this chapter use QRIS rating data (including rating levels and element scores) and classroom observation data, from programs with full QRIS rating levels that agreed to participate in the study. The classroom observation data includes CLASS scores in 139 centers and 20 family child care homes; PQA Form A scores in 140 centers and 27 FCCHs; and PQA Form B scores in 124 centers.
- Programs with full QRIS ratings in 2013 tend to be high in quality and differ from other programs participating in the QRIS in 2013 without full ratings; there are also some differences between sites that did and did not participate in classroom observations (see Appendix B). Therefore, results of the analyses may not apply to the broader range of programs participating in QRIS.

Concurrent Validity of California QRIS Ratings

California QRIS ratings are significantly and positively related to CLASS total scores in centers, although differences from one level to the next are not all significant.

California QRIS ratings have a significant and positive relationship with centers' average CLASS total scores (combining total scores for preschool and toddler classrooms). CLASS total scores increase steadily as the California QRIS rating increases, but differences are small in magnitude between Tiers 3 and 4, with a somewhat larger increase to Tier 5. Only Tiers 3 and 5 differ significantly from each other, as shown in Exhibit 5.1, but the majority of programs in the sample are rated at Tiers 3 and 4, which are not significantly different.

Exhibit 5.1. CLASS Total and Preschool Domain Scores by California QRIS Rating Level and ANOVA Results, Centers

California QRIS Rating Level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	—	0	—	—	—	0
Tier 2	—	3	—	—	—	2
Tier 3	4.81 (0.50)[e]	56	5.79 (0.46)	5.36 (0.64)	2.91 (0.86)[e]	55
Tier 4	4.94 (0.69)	68	5.97 (0.69)	5.56 (0.74)	3.01 (0.86)[e]	66
Tier 5	5.39 (0.34)[c]	12	6.23 (0.50)	5.88 (0.54)	3.74 (0.70)[c, d]	12
All levels	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	$F[3,135] = 3.40$*		$F[3,131] = 2.12$	$F[3,131] = 2.23$	$F[3,131] = 3.24$*	
Kruskall-Wallis results	$H = 12.81$**					

Cells show the mean and standard deviation for the CLASS scores at each California QRIS rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

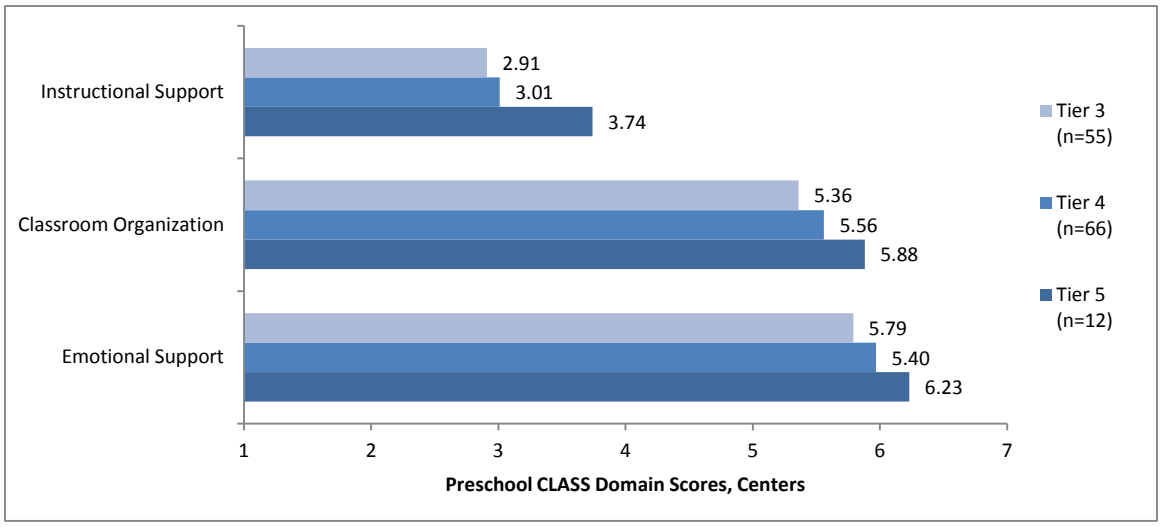
Average scores in all three CLASS domains increase steadily as California QRIS ratings increase, but the differences are mostly small in magnitude and only the relationship with Instructional Support scores is statistically significant.

Among preschool classrooms in centers, instructional support scores increased steadily by California QRIS rating level, shown in Exhibit 5.2. The differences were large and statistically significant between centers rated at Tier 5 and those rated at Tier 3 ($d = 0.97$, close to one standard deviation in magnitude), and also between centers rated at Tier 5 and those rated at Tier 4 ($d = 0.82$). This magnitude of difference is meaningful, particularly since the instructional support domain is the most difficult one to score well on and is most strongly predictive of child cognitive skills among the CLASS domains (Howes and others 2008; Mashburn and others 2008). However, mean differences between Tiers 3 and 4 are smaller and not significant, and the majority of programs in the sample are rated at these tiers.

There were no significant differences between rating levels on the emotional support or classroom organization domains, but the means on both of these domains did increase a small amount as the California QRIS rating levels increase.

Exhibit 5.2 illustrates the average preschool CLASS domain scores by California QRIS rating level in centers and the magnitude of differences between them.

Exhibit 5.2. Average Preschool CLASS Domain Scores by California QRIS Rating Level, Centers



Note: Exhibit 5.2 excludes the two centers in the preschool CLASS sample that were rated at Tier 2 because average CLASS score data are not reliable for rating levels with fewer than five observations.

In centers, the relationship between California QRIS ratings and PQA Form A total scores is positive but not statistically significant.

As shown in Exhibit 5.3, the center-average PQA Form A total scores increase consistently as California QRIS ratings increase, but the differences are small and are not statistically significant.

Exhibit 5.3. PQA Form A Total and Preschool Domain Scores by California QRIS Rating level and ANOVA Results, Centers

California QRIS Rating Level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	—	0	—	—	—	—	0
Tier 2	—	2	—	—	—	—	1
Tier 3	3.40 (0.48)	54	3.54 (0.52)	3.25 (0.59)	3.16 (0.59)[d, e]	4.09 (0.63)	53
Tier 4	3.55 (0.52)	72	3.67 (0.50)	3.30 (0.61)	3.58 (0.75)[c]	4.16 (0.52)	68
Tier 5	3.81 (0.59)	12	3.95 (0.43)	3.43 (0.65)	3.85 (0.71)[c]	3.93 (0.91)	12
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	*F*[3,136] = 2.38		*F*[3,130] = 2.28	*F*[3,130] = 0.60	*F*[3,130] = 5.54*	*F*[3,130] = 0.60	

Cells show the mean and standard deviation for the PQA scores at each California QRIS rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

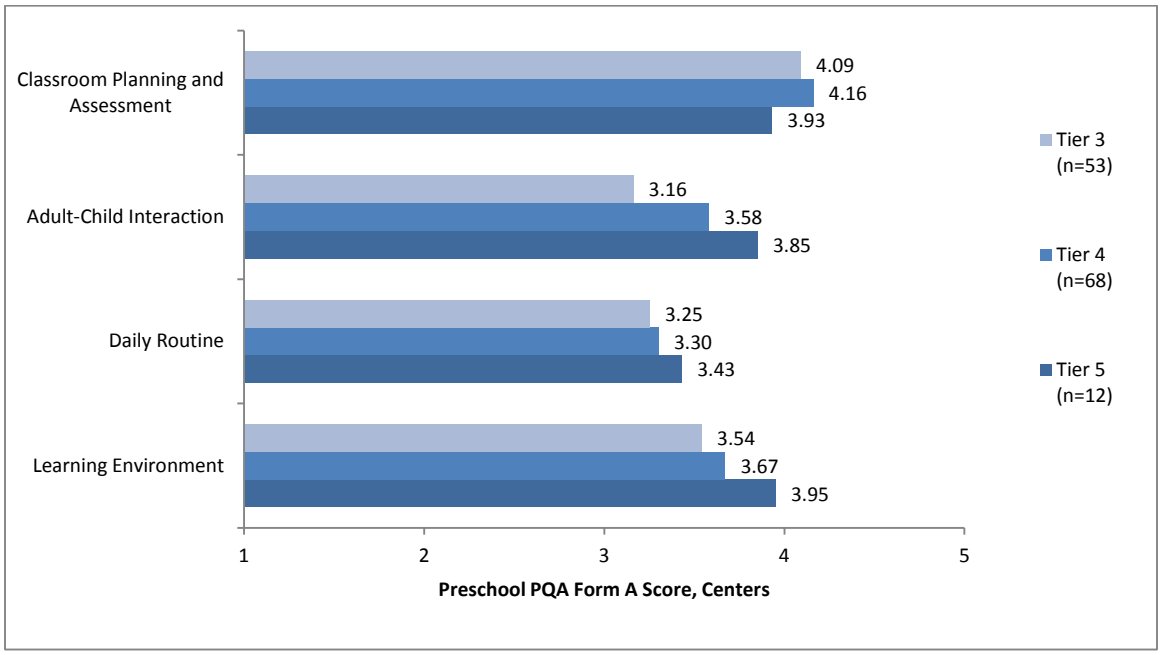
There is a positive relationship between California QRIS ratings and three of the four PQA preschool domain scores, but only the positive relationship with Adult-Child Interaction scores is statistically significant.

Among centers, adult-child interaction scores increased steadily for Tiers 3, 4, and 5. The differences were large and statistically significant between Tier 3 and Tier 4 ($d = 0.58$) and between Tier 3 and Tier 5 ($d = 0.96$). This domain measures many aspects of the quality of interactions between children and adults, including warmth and sensitivity, communication, child directedness, encouragement, and problem solving. The difference between Tiers 4 and 5 is smaller and not significant.

There were no significant differences in other preschool Form A subscale scores by rating level, but the direction of the relationship was consistently positive for the learning environment and daily routine domains. The learning environment domain assesses environment safety, the quality of equipment and materials, and organization of the classroom space. The daily routines domain assesses routines and scheduling of the day, grouping, child-directed activities, and transitions. The relationship was not consistently positive for the curriculum planning and assessment domain.

Exhibit 5.4 illustrates the average preschool PQA domain scores by California QRIS rating level in centers and the magnitude of differences between them.

Exhibit 5.4. Average Preschool PQA Form A Domain Scores by California QRIS Rating Level, Centers



Note: Exhibit 5.4 excludes the one center in the preschool PQA sample that was rated at Tier 2 because average PQA score data are not reliable for rating levels with fewer than five observations.

There are no significant relationships between California QRIS ratings and PQA Form B scores in centers, and the relationships are not consistently positive in direction.

Among centers, as shown in Exhibit 5.5, there are no significant relationships between California QRIS ratings and PQA Form B total and domain scores, and in all domains the differences between Tiers 3 and 4 are very small and slightly lower for programs rated at Tier 4 than for programs rated at Tier 3, contrary to expectations. The parent involvement and family services domain measures the level of parent involvement in program management as well as classroom activities and child learning, and also measures the center's level of screening, referrals, and supports for children with special needs. The staff qualifications and staff development domain measures the level of director and teaching staff qualifications, including degree and specialization, the amount and quality of professional development, staff supervision practices, staff affiliation with early childhood professional organizations, and staffing of support staff in the center. The program management domain measures the quality of program policies, funding, program assessment and planning, structural features such as ratio and group size, teacher turnover, recruitment practices, and accessibility.

Exhibit 5.5. PQA Form B Total and Domain Scores by California QRIS Rating Level and ANOVA Results, Centers

California QRIS Rating Level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	—	—	—	—	0
Tier 2	—	—	—	—	2
Tier 3	3.84 (0.39)	4.08 (0.52)	3.52 (0.48)	3.83 (0.45)	49
Tier 4	3.80 (0.48)	3.99 (0.60)	3.50 (0.61)	3.82 (0.54)	63
Tier 5	3.98 (0.64)	4.11 (0.80)	3.80 (0.58)	4.00 (0.64)	10
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	$F[3,120] = 0.45$	$F[3,120] = 0.40$	$F[3,120] = 0.84$	$F[3,120] = 0.94$	

Cells show the mean and standard deviation for the PQA scores at each California QRIS rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

The direction of the relationships between California QRIS ratings and toddler domain scores is not consistently positive on either the CLASS or the PQA, but conclusions cannot be drawn from these results because of the very small number of toddler classrooms observed.

Very few centers in the classroom observation sample had toddler classrooms, and thus there were only small numbers of Toddler CLASS scores ($N = 14$) or Toddler PQA Form A scores ($N = 18$). These small numbers represent an insufficient sample size from which to draw valid conclusions or test the statistical significance of score differences in rating levels. The average scores by rating level are presented descriptively in Exhibit C.3 in Appendix C, but should be interpreted with caution given the small number of programs at each tier.

Relationships between California QRIS ratings and PQA Form A scores were largely positive for FCCHs, but conclusions cannot be drawn from these results because of the very small number of FCCHs observed.

Very few FCCHs ($N = 27$) in the classroom observation sample had PQA Form A scores, an insufficient sample size from which to draw valid conclusions or test the statistical significance of score differences in rating levels. Furthermore, there was little variability in rating levels among these programs. The average scores by rating level are presented descriptively in Exhibit C.2, but should be interpreted with caution given the very small number of programs at each rating level.

The relationships between California QRIS ratings and CLASS scores cannot be compared because of the very small number of FCCHs observed and the lack of variability in the rating levels of observed FCCHs.

A small number of family child care programs were observed overall ($N = 20$), and when we focus on FCCHs serving primarily preschool aged children for analyses of Preschool CLASS scores, the number decreases even further ($N = 14$). This was an insufficient sample size from which to draw valid conclusions or test the statistical significance of score differences in rating levels. Furthermore, there was little variability in rating levels among these programs. The average scores are presented descriptively in Exhibit C.1 but should be interpreted with caution given the very small number of programs at each rating level.

Among the six FCCHs in the classroom observation sample with Toddler CLASS scores (which was used where a majority of the children served in the program were toddler age), all six had California QRIS ratings of 2. The average score on the Emotional and Behavioral Support domain was 5.81, and the average score on the Engaged Support for Learning domain was 3.23.

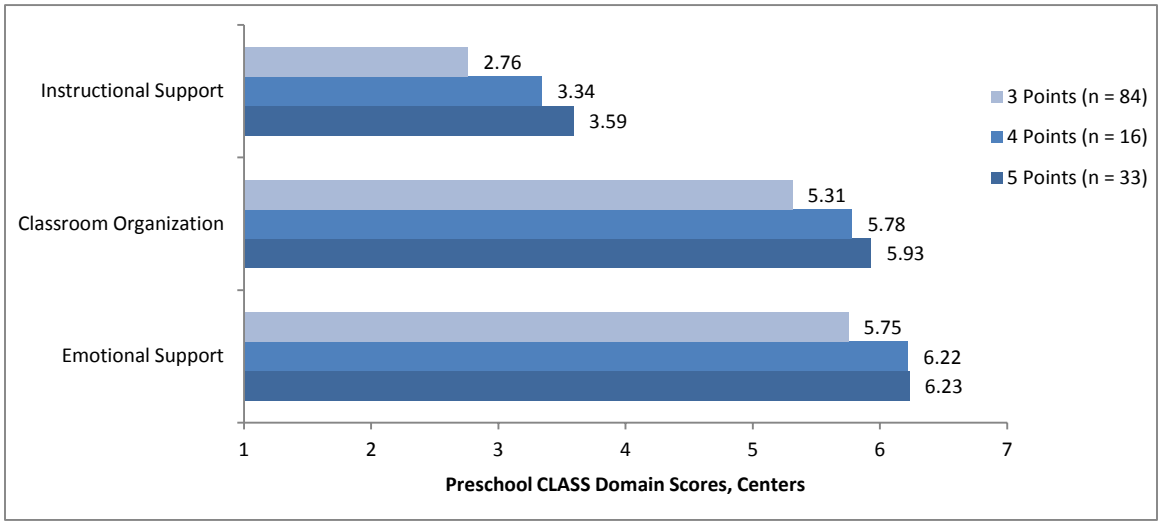
Concurrent Validity of Element Scores

Among the concurrent validity analyses using element scores, only the element scores based on the CLASS and ERS are consistent in significantly predicting classroom observation scores.

As shown in Exhibit 5.6, the element scores based on the CLASS and the ERS are the only elements that significantly predict program average CLASS and PQA scores. Among other elements, there were few significant relationships with classroom observation scores, and when the element scores did significantly predict classroom observation scores, the relationships were either negative or not consistent in direction as the element score increased.

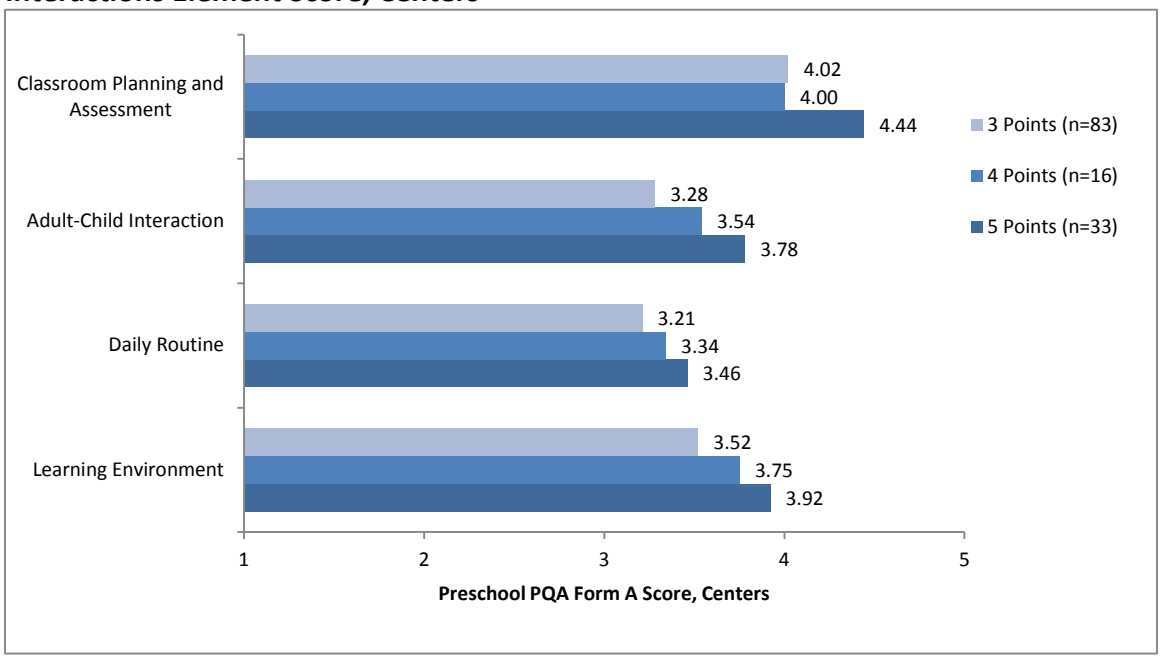
The effective teacher-child interactions element score is based the CLASS instrument. Element scores of 1 or 2 are determined by level of familiarity with the CLASS instrument, while higher scores of 3, 4, or 5 require an independent CLASS observation (see Exhibit 3.2 in Chapter 3 for the Hybrid Rating Matrix). All but two centers in the study sample received scores of 3 or higher on this element. The effective teacher-child interaction element score positively predicts all classroom-level observation scores used in the study, including CLASS and PQA Form A total scores and subscale scores, except that the average scores on the PQA curriculum planning and assessment are very slightly lower at Tier 4 than Tier 3 without statistical significance. The relationship between the effective teacher-child interaction element score and the PQA Form B scores, which measure program-level structural quality, are positive but not statistically significant. Exhibits 5.6 and 5.7 illustrate the average preschool CLASS and PQA domain scores by Effective Teacher-Child Interactions element score.

Exhibit 5.6. Average Preschool CLASS Domain Scores by Effective Teacher-Child Interactions Element Score, Centers



Note: Exhibit 5.6 excludes the two centers in the preschool CLASS sample that had Effective Teacher-Child Interactions element scores of 2 because average CLASS score data are not reliable for rating levels with fewer than five observations.

Exhibit 5.7. Average Preschool PQA Form A Domain Scores by Effective Teacher-Child Interactions Element Score, Centers



Note: Exhibit 5.7 excludes the two centers in the preschool PQA sample that had Effective Teacher-Child Interactions element scores of 2 because average PQA score data are not reliable for rating levels with fewer than five observations.

The program environment rating scale element is based on the ERS instrument and is similar to scoring of the teacher-child interactions element. Element scores of 1 or 2 are determined by level of familiarity with the ERS instrument, whereas higher scores of 3, 4, or 5 require an independent ERS observation (see Exhibit 3.1 in Chapter 3 for the Hybrid Rating Matrix). The

program environment rating scale is significantly related to the total CLASS and PQA Form A scores and most of the subscale scores, but the relationships are not consistently positive. This is largely because the average CLASS and PQA scores among programs rated at Tier 2 on the program environment rating scale element (indicating that they opted out of or were not ready for an ERS observation) are sometimes higher than the average CLASS and PQA scores of programs that scored at Tier 3 or higher (indicating that they received a CLASS observation). The program environment rating scale element scores of 3, 4, and 5 are positively associated with most CLASS and PQA Form A scores. In other words, the element is more successful at differentiating among programs that did receive an ERS observation than among programs that did not receive one. Exhibits 5.8 and 5.9 illustrate the average preschool CLASS and PQA domain scores by program environment rating score element score.

Exhibit 5.8. Average Preschool CLASS Domain Scores by Program Environment Rating Element Score, Centers

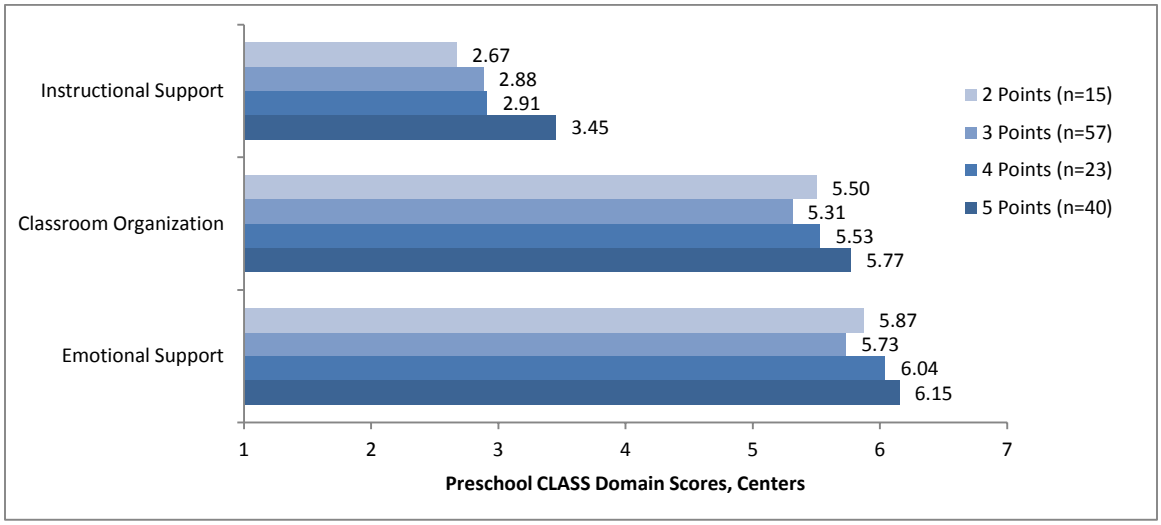


Exhibit 5.9. Average Preschool PQA Form A Domain Scores by Program Environment Rating Element Score, Centers

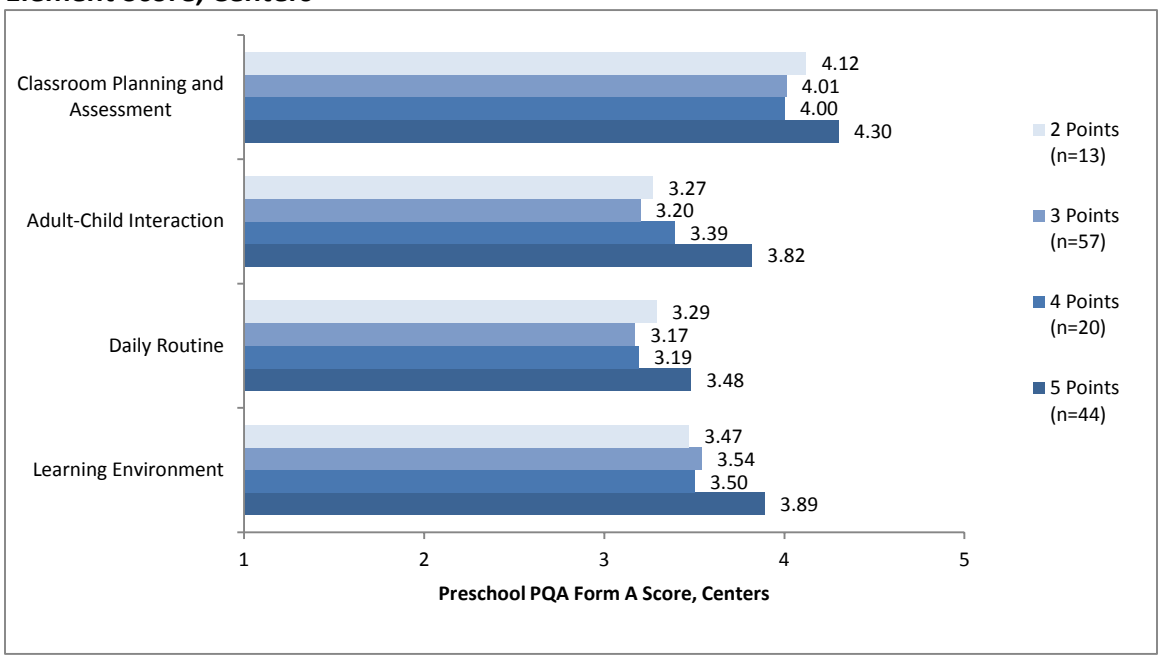


Exhibit 5.10 summarizes the element score concurrent validity analysis results.

Exhibit 5.10. Summary of Element Score Concurrent Validity Analysis Results, Centers

Concurrent Validity Dependent Variable	Analysis Results for Each Element Score						
	Child Observation	Develop-mental and Health Screenings	Minimum Qualifications for Lead Teacher	Effective Teacher-Child Interactions	Ratios and Group Sizes	Program Environment Rating Scales	Director Qualifications
CLASS Scores							
Total Score (Preschool and Toddler)	↗	↗	↗	↗*	↗	↗*	↗*
Emotional Support (Preschool)	↗	↗	↗	↗*	↗	↗*	↗
Classroom Organization (Preschool)	↗	↗	↗	↗*	↗	↗*	↗*
Instructional Support (Preschool)	↗	↗	↗	↗*	↗	↗*	↗
PQA Scores							
Form A Score (All Ages)	↗	↗	↗	↗*	↘	↗*	↗
Learning Environment (Preschool)	↗	↗	↗	↗*	↗	↗*	↗
Daily Routine (Preschool)	↗	↗	↗	↗	↘	↗*	↘
Adult-Child Interaction (Preschool)	↗	↗	↗	↗*	↗	↗	↗
Curriculum Planning and Assessment (Preschool)	↘*	↗	↗	↗*	↗	↗	↗
Form B Score (All Ages)	↗	↗*	↗	↗	↗	↗	↗
Parent Involvement and Family Services (All Ages)	↗*	↗*	↗	↗	↗	↗	↘
Staff Qualifications and Staff Development (All Ages)	↗	↗	↗	↗	↗*	↗	↗
Program Management (All Ages)	↗	↗	↗	↗	↗	↗	↗

Note: Each row references the results of a separate ANOVA model.

* indicates a statistically significant relationship, and the arrows indicate the direction of the relationship between QRIS ratings and observed classroom quality scores, for rating levels with more than five observations:

↗ indicates a consistently positive relationship; ↘ indicates a consistently negative relationship; ↗↘↗ indicates relationships that are not consistent in direction.

Summary: How Well Does the QRIS Differentiate Between Observed Quality of Programs?

Results from the concurrent validity analyses find some evidence that the California QRIS ratings differentiate between observed quality of programs, although the differences are small in magnitude in most cases. In particular, California QRIS ratings positively and significantly predict CLASS total scores, Preschool CLASS Instructional Support scores, and Preschool PQA Adult-Child Interaction scores (see Exhibit 5.11 for a summary of results). Sample sizes for toddler classrooms and FCCHs were not sufficient to produce reliable conclusions for these settings.

Exhibit 5.11 Summary of California QRIS Concurrent Validity Analysis Results, Centers

Concurrent Validity Dependent Variable	Analysis Result
CLASS Scores	
Total Score (Preschool and Toddler)	↗ *
Emotional Support (Preschool)	↗
Classroom Organization (Preschool)	↗
Instructional Support (Preschool)	↗ *
PQA Scores	
Form A Score (All Ages)	↗
Learning Environment (Preschool)	↗
Daily Routine (Preschool)	↗
Adult-Child Interaction (Preschool)	↗ *
Curriculum Planning and Assessment (Preschool)	↗
Form B Score (All Ages)	↗
Parent Involvement and Family Services (All Ages)	↗
Staff Qualifications and Staff Development (All Ages)	↗
Program Management (All Ages)	↗

Note: Each row references the results of a separate ANOVA model.

* indicates a statistically significant relationship, and the arrows indicate the direction of the relationship between QRIS ratings and observed classroom quality scores: ↗ indicates a consistently positive relationship; ↘ indicates a consistently negative relationship; ↗ indicates relationships that are not consistent in direction.

Among the concurrent validity analyses using element scores, only the element scores based on the CLASS and ERS are consistent in significantly predicting observation scores. It is not surprising that the element scores based on classroom observation tools are most strongly associated with the study's classroom observation scores. The other element scores may be thought of as indicators of structural quality; previous studies have found that structural quality measures predict classroom observation scores (Burchinal and others 2002; Goelman and others 2006; NICHD ECCRN 2002; Phillips and others 2000; Phillipsen and others 1997). However, in this study of the California QRIS, none of the structural quality element scores predict classroom observation scores. Furthermore, none of the structural quality element scores positively predict the PQA Form B, which is an independent measure of program structural quality. This lack of a relationship between structural element scores and the independent structural quality measure suggests that the element scores could be improved to ensure more variability. However, it is possible that greater variability in the structural element scores will be found if QRIS participation expands to a more diverse group of programs.

Chapter 6. Alternative Rating Approaches

This chapter presents sensitivity analyses that demonstrate how alternative approaches to calculating QRIS ratings affect the rating levels of programs participating in California's QRIS. The sensitivity analyses compare the California QRIS ratings to six alternative rating approaches, including two approaches currently used in the state as local adaptations to the statewide rating approach and four approaches that are not currently used in the state. Exhibit 6.1 provides a definition for each rating approach included in these comparisons.

In addition, this chapter presents concurrent validity analyses for each alternative rating approach.

We address each of the following research questions in this chapter.

- RQ 4. How is the hybrid rating strategy and rating outputs representative of meaningful levels of quality?
- RQ 5. How do QRIS ratings that use locally determined tiers differ from QRIS ratings calculated using recommendations in California's RTT-ELC QRIS Implementation Guide?
- RQ 6. How effective is the rating protocol at determining valid ratings versus an annual 100 percent assessment protocol?

Analysis Approaches

- Simulation of alternative rating approaches: QRIS ratings are calculated using a variety of calculation approaches, using the element score data collected for the California QRIS. The definition of each rating approach is provided in Exhibit 6.1.
- Concurrent validity: ANOVA models for centers, assessing whether average scores on independent measures of observed quality differ by each simulated QRIS rating approach. The results of the ANOVA models are compared to determine which rating approaches best differentiate observed program quality.
- Percentage of classrooms observed: Analyses examining the consistency of element scores using different protocols for the percentage of classrooms observed, in centers with multiple classrooms.

Data and Sample

- The rating simulations use QRIS element scores from 472 programs across the state with full QRIS ratings as of January 2014. The concurrent validity analyses include the programs within this sample with classroom observation data. The classroom observation data include CLASS scores in 139 centers and 20 FCCHs; PQA Form A scores in 140 centers and 27 FCCHs; and PQA Form B scores in 124 centers. Analyses of the percentage of classrooms observed use QRIS element scores, CLASS scores, and ERS scores for 26 centers with CLASS and ERS scores for every classroom in the center.
- Programs with full QRIS ratings in 2013 tend to be high in quality and differ from other programs participating in the QRIS in 2013 without full ratings; there are also some differences between sites that did and did not participate in classroom observations for both the concurrent validity analyses and the percentage of classrooms observed analyses (see Appendix B). Therefore, results of the analyses may not apply to the broader range of programs participating in QRIS.

Exhibit 6.1. Alternative Rating Approaches Examined in This Study

Rating Type	Rating Definition
Rating Approaches Currently Used in California	
California QRIS	Tier 1 is blocked: Programs must meet criteria for at least 1 point on all applicable elements for a rating of 1. Tiers 2–5 are point-based for programs meeting block criteria for Tier 1: Tier is determined by total points earned across elements. This is California's rating approach without local adaptations to the way the ratings are calculated using the element scores.
Two-Level Block	Tiers 1 and 2 are blocked, and Tiers 3–5 are point-based for programs meeting block criteria for Tier 2. This approach is used as a local adaptation to California's rating approach in some counties.
Consortia QRIS	This is not a single rating approach but instead refers to the ratings assigned by Consortia, using local adaptations to the California QRIS ratings.
Rating Approaches Under Consideration, But Not Currently Used in California	
Three-Level Block	Tiers 1–3 are blocked, and Tiers 4–5 are point-based for programs meeting block criteria for Tier 3.
Five-Level Block	Tiers 1–5 are blocked.
Element Average	Scores are determined by taking the average of all applicable rating elements (seven elements for centers, six elements for infant-only centers, five elements for FCCHs, four elements for infant-only FCCHs). Averages are rounded to whole numbers (round up for 0.5 and above, round down below 0.5).
ERS Hybrid	Tier 1 is blocked, Tiers 2–4 are point-based for programs meeting block criteria for Tier 1, and Tier 5 is based on points for programs meeting block criteria for Tier 1 and a Tier 5 block in the program environment rating scale element. Point ranges for Tiers 2–5 are adjusted to exclude the ERS element from the total points because this element is now used in a different way as a block at Tier 5.

Note: Elements are the domains of quality included in California's QRIS. All rating approaches are calculated using element scores collected by Consortia on participating programs. Scores for each element range from 1 to 5 and are determined by meeting criteria for each point level. Centers are rated on seven elements (centers serving only infants are rated on six elements), and FCCHs are rated on five of the seven elements that apply to centers (FCCHs serving only infants are rated on four elements). Some Consortia made local adaptations to element scoring rather than using the statewide criteria.[22] Blocking a tier means that programs meet all requirements for each element score at that tier (for example, blocking at Tier 2 means that programs must have a score of at least 2 on all elements in order to be rated at 2 or higher).

Sensitivity Analyses Comparing Distributions of California QRIS Ratings and Alternative Rating Approaches

The distribution of rating levels varies by rating approach, especially when blocks are used.

Program ratings are affected by the rating calculation approach both for centers and FCCHs, as shown in Exhibits 6.2 and 6.3. Among centers and FCCHs, many programs receive ratings of 1 when blocks are used for higher rating levels, although very few or no programs receive ratings

[22] The study analyses use simulated QRIS ratings that the study team calculated from element score data collected by Consortia, using the California QRIS rating guidelines without any local options to the extent possible. In most of the 11 Consortia with valid QRIS ratings, local adaptations to the rating criteria were applied after element scores were calculated. However, two Consortia (Sacramento and Ventura) incorporated local adaptations into the element scores, and the study team was not able to recalculate the element scores without these local adaptations.

of 1 using all other rating approaches. As the number of tiers that are blocked increases, fewer programs receive high ratings. This is consistent with other research that used ECLS-B data to synthesize different rating structures and found a block structure to be the most stringent rating approach (Tout and others 2014).

Exhibit 6.2. Distribution of Ratings Using Alternative Rating Approaches, Centers

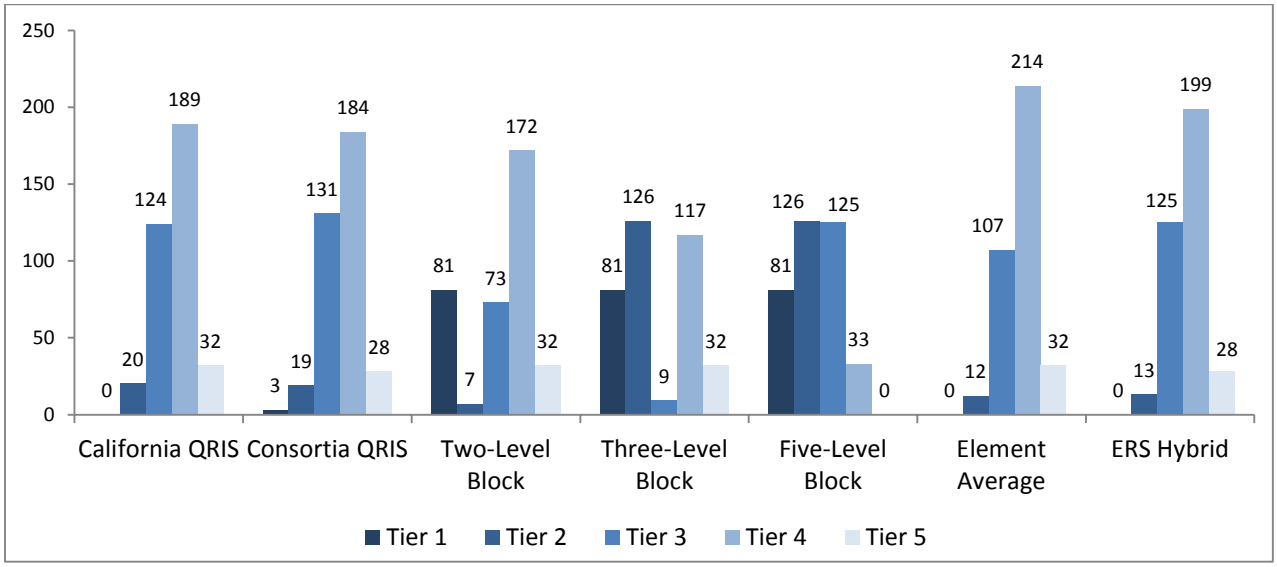
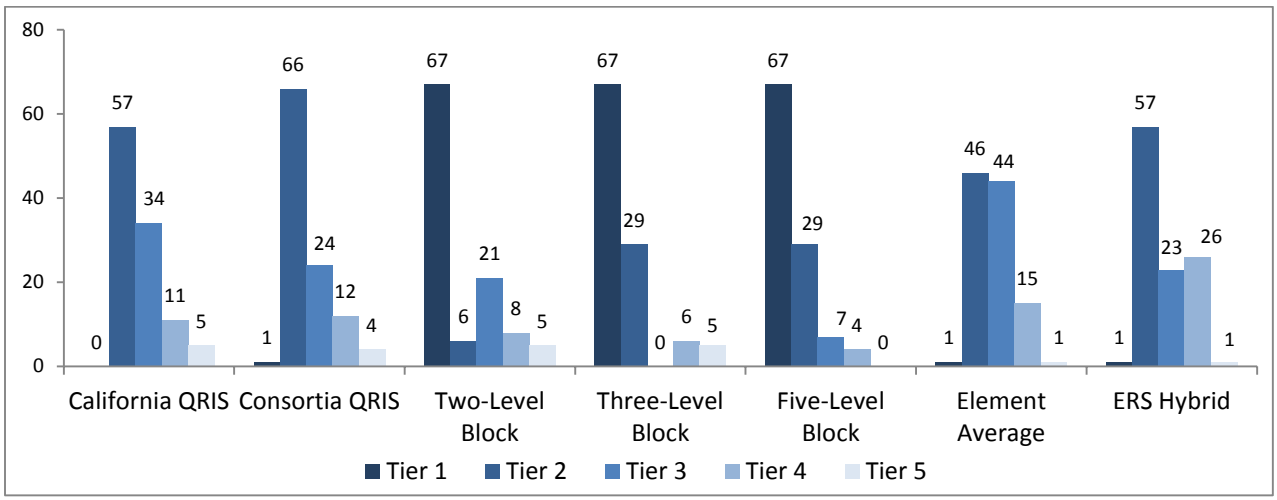


Exhibit 6.3. Distribution of Ratings Using Alternative Rating Approaches, FCCHs



The distribution of locally adapted Consortia QRIS ratings is similar to the distribution of California QRIS ratings, especially for centers. Local adaptations were used in six of the 12 counties with final QRIS ratings, representing 64 percent of the centers and 36 percent of the FCCHs. However, local adaptations did not affect all programs rated in those six counties. Across all 12 counties, 95 percent of centers and 87 percent of FCCHs had the same rating in both the Consortia QRIS ratings and the California QRIS ratings, as shown in Exhibits 6.2 and

6.3. When ratings differed, the California QRIS ratings were usually higher than the Consortia QRIS ratings.

As the number of rating levels that are blocked increases, program ratings tend to decrease, as shown in Exhibits 6.4 and 6.5. In the two-level block ratings (blocking at Tiers 1 and 2), 22 percent of centers and 63 percent of FCCHs had lower ratings in comparison to California QRIS ratings. Even programs rated as high as Tier 4 in the California QRIS ratings received a rating of 1 when Tier 2 was blocked. In the three-level block ratings (blocking at Tiers 1, 2, and 3), 55 percent of centers and 84 percent of FCCHs had lower ratings in comparison to California QRIS ratings. When all five tiers are blocked, 93 percent of centers and 94 percent of FCCHs had lower ratings in comparison to California QRIS ratings.

In contrast, the distribution of ratings skewed somewhat higher in the element average approach both for centers and for FCCHs. Among centers, 91 percent received the same ratings as in the California QRIS approach, and 9 percent received higher ratings by one rating level. Among FCCHs, 86 percent received the same rating, although 9 percent received higher ratings by one tier and 5 percent received lower ratings by one tier.

In the ERS hybrid approach, ratings were the same for most programs, with a mix of higher and lower ratings among those that differed. Among centers, 4 percent had lower ratings and 7 percent had higher ratings in comparison to the California QRIS ratings. Among FCCHs, 8 percent had lower ratings and 15 percent had higher ratings.

Exhibit 6.4. Reclassification Rates for Alternative Rating Approaches, Centers

Rating Type	Percentage Lower Than California QRIS Rating	Percentage Same As California QRIS Rating	Percentage Higher Than California QRIS Rating
Consortia QRIS	0.6	95.1	4.4
Two-Level Block	22.2	77.8	0.0
Three-Level Block	54.8	45.2	0.0
Five-Level Block	92.6	7.4	0.0
Element Average	0.0	91.0	9.0
ERS Hybrid	4.1	89.3	6.6

N = 365 centers

Exhibit 6.5. Reclassification Rates for Alternative Rating Approaches, FCCHs

Rating Type	Percent Lower Than California QRIS Rating	Percent Same As California QRIS Rating	Percent Higher Than California QRIS Rating
Consortia QRIS	0.9	86.9	12.2
Two-Level Block	62.6	37.4	0.0
Three-Level Block	84.1	15.9	0.0
Five-Level Block	94.4	5.6	0.0
Element Average	4.7	86.0	9.4
ERS Hybrid	8.4	76.6	15.0

N = 107 FCCHs

Concurrent Validity of Alternative Rating Approaches

Element average ratings are more effective than California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores.

Element average ratings have statistically significant positive relationships with CLASS total scores and all three preschool CLASS domain scores, while the California QRIS ratings are only significantly related to instructional support scores. Element average ratings are positively associated with the learning environment domain of the preschool PQA, as well as the adult-child interaction domain, and relationships with the other PQA observation scores are positive in direction although not statistically significant. Unlike the California QRIS ratings, the direction of the relationship between element average ratings and PQA program-level Form B scores are also mostly positive, although not statistically significant.

Ratings using blocks are less effective than California QRIS ratings at differentiating centers by CLASS scores, but five-level blocks are more effective at differentiating centers according to the PQA observation scores.

Exhibit 6.6 shows that rating approaches using blocking are not positively related to CLASS domain scores in most cases, in contrast to California QRIS ratings. The relationship with CLASS scores is weakest in ratings that block at all five tiers, and CLASS scores do not consistently increase as the rating level increases. However, ratings with blocking at all five rating levels are more predictive of PQA classroom observation scores than California QRIS ratings. The five-level block ratings are positively associated with PQA Form A total scores as well as the preschool adult-child interaction domain score and are significantly related to the preschool learning environment domain score, although the relationship is not consistently positive.

Both the ERS hybrid rating approach and the Consortia QRIS ratings are similar to the California QRIS ratings in their patterns of relationships with classroom observation scores.

The ERS hybrid rating approach is similar to the California QRIS ratings in significantly predicting the CLASS total score, the preschool CLASS instructional support score, and the PQA adult-child interactions score, and having positive but nonsignificant relationships with

most other CLASS and PQA classroom observation scores. Unlike the California QRIS ratings, the relationships between the ERS hybrid ratings and the PQA program-level Form B scores are also mostly positive in direction, although also not statistically significant.

The local Consortia QRIS ratings are significantly predictive of the CLASS instructional support and PQA adult-child interaction scores, but unlike the California QRIS ratings, they are not significantly related to the CLASS total scores.

The results of the concurrent validity analyses using alternative rating approaches are specific to the sample of centers included in the study, and the relationships between alternative rating approaches and observed quality scores may differ for other programs in California.

The concurrent validity analyses using alternative rating approaches do not include FCCHs because of sample size limitations. The relationships between rating levels and observed quality scores may differ for FCCHs. Also, the sample of centers included in the study is not representative of all centers in California. A study that includes lower rated centers might find different relationships between ratings and observed quality.

Exhibit 6.6. Summary of Concurrent Validity Analysis Results for Alternative QRIS Rating Approaches, Centers

Concurrent Validity Dependent Variable	Analysis Results for Alternative QRIS Rating Approaches						
	California QRIS Rating	Consortia Rating	Two-Level Block	Three-Level Block	Five-Level Block	Element Average	ERS Hybrid
CLASS Scores							
Total Score (Preschool and Toddler)	↗ *	↗	↗*	↗	↗	↗ *	↗ *
Emotional Support (Preschool)	↗	↗	↗	↗	↗	↗ *	↗
Classroom Organization (Preschool)	↗	↗	↗	↗	↗	↗ *	↗
Instructional Support (Preschool)	↗ *	↗ *	↗*	↗*	↗	↗ *	↗ *
PQA Scores							
Form A Score (All Ages)	↗	↗	↗	↗	↗ *	↗	↗
Learning Environment (Preschool)	↗	↗	↗	↗	↗*	↗ *	↗
Daily Routine (Preschool)	↗	↗	↗	↗	↗	↗	↗
Adult-Child Interaction (Preschool)	↗ *	↗ *	↗ *	↗ *	↗ *	↗ *	↗ *
Curriculum Planning and Assessment (Preschool)	↗	↗*	↗	↗	↗	↗	↗
Form B Score (All Ages)	↗	↗	↗	↗	↗	↗	↗
Parent Involvement and Family Services (All Ages)	↗	↗	↗	↗	↗	↗	↗
Staff Qualifications and Staff Development (All Ages)	↗	↗	↗	↗*	↗*	↗	↗
Program Management (All Ages)	↗	↗	↗	↗	↗	↗	↗

Note: Each row references the results of a separate ANOVA model.

* indicates a statistically significant relationship, and the arrows indicate the direction of the relationship between QRIS ratings and observed classroom quality scores for rating levelstier with more than five observations:

↗ indicates a consistently positive relationship; ↘ indicates a consistently negative relationship; ↗↘↗ indicates relationships that are not consistent in direction.

http://www.cde.ca.gov/sp/cd/rt/aavsumvalidityanalyses.asp

Percentage of Classrooms Observed

This section of Chapter 6 presents results of analyses comparing program average observation scores, element scores, and QRIS ratings using different protocols for selecting classrooms to observe in centers with multiple classrooms. These analyses provide information about the reliability of QRIS ratings using the California QRIS protocol for selecting one third of classrooms, in comparison to alternative classroom observation protocols: 100 percent of classrooms, and one half of classrooms. Exhibit 6.7 describes these classroom observation protocols in detail.

Exhibit 6.7. Alternative Classroom Observation Protocols Examined in This Study

Classroom Observation Protocol	Detailed Description
100 percent protocol	Classroom observations are conducted in all eligible classrooms in the center. ERS element scores of 3, 4, or 5 points are determined by the average of ERS observation total scores from all infant, toddler, and preschool classrooms in the center (averaging total scores on the ECERS and the ITERS in programs with infant and toddler classrooms and preschool classrooms), and CLASS element scores of 3, 4, or 5 points are determined by Preschool CLASS domain scores averaged across all preschool classrooms in centers with preschool classrooms, as well as Toddler CLASS domain scores in centers with toddler classrooms. Infant-only classrooms are not observed with the CLASS instrument.
One third protocol	Classroom observations are conducted in one third of preschool classrooms in the center in centers with preschool classrooms, as well as one third of infant and toddler classrooms in the center in centers with these types of classrooms. Classrooms to observe are randomly selected from all eligible classrooms. The number of classrooms to observe is rounded up to the nearest whole number if one third does not equal a whole number. For example, in a center with four preschool classrooms and two toddler classrooms, one third of four classrooms is 1.33 and is rounded up to 2, and one third of two toddler classrooms is 0.67 and is rounded up to 1, so that center has two preschool observations and one toddler observation conducted. ERS and CLASS element scores are determined using the same criteria described for the 100 percent protocol, but use ERS and CLASS score averages from the randomly selected one third of classrooms.
One half protocol	Classroom observations are conducted in one half of preschool classrooms in the center in centers with preschool classrooms, as well as one half of infant and toddler classrooms in the center in centers with these types of classrooms. Classrooms to observe are randomly selected from all eligible classrooms. The number of classrooms to observe is rounded up to the nearest whole number if one third does not equal a whole number. For example, in a center with three preschool classrooms and one toddler classroom, one half of three classrooms is 1.5 and is rounded up to 2, and one half of one toddler classroom is 0.5 and is rounded up to 1, so that center has two preschool observations and one toddler observation conducted. ERS and CLASS element scores are determined using the same criteria described for the 100 percent protocol, but use ERS and CLASS score averages from the randomly selected one half of classrooms.

Note: The one third protocol is currently used to select classrooms for observations in centers with multiple classrooms for the effective teacher-child interactions element score and the program environment rating scales element score in the California QRIS.

Among the small number of centers in the sample, classroom observation scores vary more within centers than from center to center.

There is considerable variability in classroom observation scores within sites. As shown in Exhibit 6.8, more than half the variation in observation scores in the sample occurs within the centers, rather than from center to center, and almost all of the variation occurs within centers on the CLASS preschool emotional support domain and on both toddler CLASS domains. In other words, differences in CLASS scores across classrooms are more likely to occur within a center than between centers. However, these analyses include a small subset of 26 centers with classroom observation data on all classrooms and may not apply to all centers with multiple classrooms participating in the California QRIS.

Exhibit 6.8. Variance in Observation Scores and Percentage of Variance Within Centers

Classroom Observation Score	Minimum Score	Maximum Score	Grand Mean of Scores	Total Variance of Scores	Percentage of Variance Within Centers
ERS Total Score	1.76	5.62	3.95	0.63	52%
Preschool CLASS Emotional Support	4.40	6.80	5.73	0.34	99%
Preschool CLASS Classroom Organization	3.67	6.93	5.34	0.53	62%
Preschool CLASS Instructional Support	1.27	5.25	2.98	0.93	64%
Toddler CLASS Emotional and Behavioral Support	3.51	6.92	5.63	0.53	93%
Toddler CLASS Engaged Support for Learning	1.40	5.47	3.11	1.19	93%

N = 26

Among this small sample of programs, element scores based on the ERS were rarely affected by selecting different combinations of classrooms in the one third or one half classroom observation protocols, whereas element scores based on the CLASS were affected in about a third of programs.

Program Environment Rating Scale (ERS) element scores, which are based on ERS, were seldom affected by changes in the percentage of classrooms observed. Of the 26 programs in the analysis sample, just one program (4 percent) had a different score on the Program Environment Rating Scales element when the one third or one half protocols were used in comparison to the 100 percent protocol, as shown in Exhibit 6.8. The one program with a different Program Environment Rating Scales element scores had just two classrooms and had less variability in ERS scores (with a difference of just 0.31 between the two classrooms' scores) than most of the other centers in the sample. However, this small difference in scores in the center happened to cross the program environment rating scales element score threshold, from one score level to another.

In contrast, Exhibit 6.9 shows that Effective Teacher-Child Interactions (CLASS) element scores were affected more frequently by changes to the percentage of classrooms observed. More than a

third (38%) of centers in the sample had a different score on the Effective Teacher-Child Interactions element in at least some combinations of selected classrooms, using either the one third or the one half observation protocol, in comparison to the 100 percent protocol. Different scores[23] on the Effective Teacher-Child Interactions element were more prevalent in smaller programs: one half of programs with just two classrooms had different element scores using either the one half or one third protocols. Among programs with larger numbers of classrooms (three to six), 21 percent had different element scores in the one half protocol and 29 percent had different element scores in the one third protocol. Across all programs, the Effective Teacher-Child Interactions element score was the same as the 100 percent protocol 76 percent of the time, on average, using the one third protocol and 78 percent of the time, on average, using the one half protocol.

Exhibit 6.9. Frequency of Differences in QRIS Ratings or Element Scores Due to Alternative Classroom Observation Protocols

Number of Classrooms in Center	Number of Centers	Number of Programs With QRIS Ratings and Element Scores That Differ From 100 Percent Protocol					
		One Third Protocol			One Half Protocol		
		Different QRIS Rating	Different ERS Score	Different CLASS Score	Different QRIS Rating	Different ERS Score	Different CLASS Score
Two classrooms	12	0	1	6	0	1	6
Three to four classrooms	7	0	0	2	0	0	1
Five to six classrooms	7	2	0	2	2	0	2
Total	26	2	1	10	2	1	9

Although differences in the CLASS element scores due to the classroom sampling protocol were fairly common, these differences in element scores rarely resulted in different QRIS rating levels among centers in the sample.

Although differences in the CLASS element scores were fairly common, just two programs (8 percent) had different QRIS ratings as a result of either the one half or the one third observation protocol, in comparison to the 100 percent protocol. In both cases, the different QRIS ratings occurred in programs with six classrooms, in both cases because the total score happened to be close to the rating threshold from one rating level to another, and crossed the threshold when the CLASS element score changed. QRIS ratings are not necessarily affected by changes in CLASS or ERS element scores, unless the change in total points earned across all elements happens to cross a rating level threshold.

Although few programs in the analysis sample had different QRIS ratings across different sampling protocols, the results might differ in a larger sample of programs with multiple classrooms.

[23] Scores were considered different if they varied by one or more points on the element.

Summary: How Do Alternative Rating Approaches Affect the Distribution and Validity of Ratings?

There are many approaches to calculating ratings for a QRIS; California's hybrid method is one such approach. To explore how ratings and validity would change under different rating approaches, we tested six alternative rating approaches using the same element scores collected for the California QRIS ratings. To assess these approaches, we examined changes in the distribution of ratings under different rating calculation methods and then examined the concurrent validity of these different rating approaches. We also examined how ratings would change under different classroom sampling procedures for conducting the classroom assessments using the CLASS and ERS.

First, we found that the distribution of rating levels varies by rating approach. The largest changes in the distribution of ratings occur in rating approaches using blocks; 63 percent of programs have lower ratings when only Tier 2 is blocked, while 94 percent of programs have lower ratings when all five tiers are blocked.

Second, we found that element average ratings are more effective than California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores. And although ratings using blocks are less effective than California QRIS ratings at differentiating centers by CLASS scores, five-level blocks are more effective at differentiating centers according to the PQA observation scores. Both the ERS hybrid rating approach (removing the element score based on the ERS from the points-based part of the rating, but including a block requirement for an ERS score of 5.5 or higher for a Tier 5 rating) and the Consortia QRIS ratings (using the Consortia applied local options for modifying the Hybrid Rating Matrix) are similar to the California QRIS ratings in their patterns of relationships with classroom observation scores.

It is important to remember when interpreting these concurrent validity analyses using alternative rating approaches that they are specific to the sample of centers included in the study. As noted in Chapters 2 and 4, the sample of programs is small and not representative of the entire population of programs in California, and the relationships between alternative rating approaches and observed quality scores may differ for other programs in California.

Finally, we found that when we took different approaches to sampling classrooms for observations, CLASS element scores were often affected by the combination of classrooms selected, while QRIS ratings and ERS scores rarely were. Still, there is considerable variability in CLASS and ERS scores within centers with multiple classrooms, and the program average CLASS and ERS scores is affected by the combination of classrooms selected. If the CDE does not plan to use the program average CLASS and ERS scores for any purposes other than calculating QRIS ratings, the analyses with this limited sample of programs do not suggest the need to change the sampling protocol. However, if the CDE intends to publish or otherwise use the program average CLASS scores or the CLASS element scores, the analyses suggest that accuracy of these program average scores is greatly increased using 100 percent of the classrooms in the center.

Chapter 7. Discussion and Preliminary Conclusions

In this chapter, we review the findings from the Validity and Reliability Study and present preliminary conclusions. The chapter also presents some ideas that the state may want to consider as next steps in the refinement of the Hybrid Rating Matrix. As noted throughout this report and highlighted again in this chapter, what conclusions can be reached as a result of this study are limited. Because it is early in the implementation timeline, because there were only a small number of sites with full ratings at the time of this analysis, and, most importantly, because there is limited variation in ratings, the results should be interpreted with caution and with the understanding that findings could change with a different group of programs included in the system. Moreover, this phase of the study did not examine predictive validity of the ratings, which is another important piece of the validation picture. The next phase of the study will include an examination of predictive validity, which should be included in decisions about refining the system.

Summary of Findings

In this section, we summarize the findings presented in the preceding chapters as they align with each of the research questions outlined in the beginning of this report (see Exhibit 1.2 in Chapter 1). But first we begin with an overview of the context by describing the overall status of implementation of the QRIS at the time of data collection.

Context: Status of QRIS Implementation

All of the Consortia participating in the California RTT-ELC QRIS have some history of local quality improvement systems, and nearly a third had well-developed systems using tiered reimbursement (payment rates tied to the level of quality) in place prior to the RTT-ELC award because of their participation in the First 5 Power of Preschool and related initiatives. Nevertheless, the use of the Hybrid Rating Matrix and the adopted approach for calculating ratings was new to counties in 2013. As described in Chapter 2, county Consortia have made tremendous progress toward their goals, but of course, the system is not fully implemented at this time, as full implementation was not promised until the end of 2015. Interviews conducted in spring 2014 revealed that Consortia had made progress, and most were on target with respect to their implementation plans. However, few had implemented all of their planned activities for the grant. Similarly, several Consortia had reached their targets for provider participation, but most Consortia were adhering to the phased-in enrollment projected in their initial plans. Consortia also focused their efforts on conducting assessments, reviewing documents, and calculating ratings for their sites. Although many had complete ratings for sites, most of the Consortia reported that they were either holding off on publicly sharing the ratings or had had only preliminary discussions about doing so. Two Consortia specifically mentioned waiting for results of the validation study before making ratings public, so they could feel confident that the ratings appropriately reflected the quality of their programs.

One of the major barriers to full implementation of the system is conducting the classroom observations (using the CLASS and ERS) needed to calculate ratings. The most common challenge that Consortia reported facing was finding, training, and retaining qualified classroom

assessors, especially for conducting ERS observations. The cost of observations, which varied considerably among Consortia, also posed a challenge for many. Consortia took different approaches to making observations affordable and manageable, from recruiting a large pool of classroom assessors from which to draw to partnering with other Consortia or agencies in the area to pool their assessor resources. Consortia also faced other challenges, such as ensuring that local colleges offered sufficient ECE courses to meet the demand from participating providers. However, here we focus on the issues associated with developing the ratings themselves. Other challenges will be addressed in the forthcoming study on child outcomes.

The first few years of any new initiative are likely to pose implementation challenges, and the system will likely be modified on its way to full implementation; thus, it is early to be evaluating this cross-county system. However, lessons can be learned from the early ratings information currently available that could inform decisions about modifications or enhancements to the rating system. With this possibility in mind, we summarize the results presented in the previous chapters according to each of the study research questions.

RQ 1. How Effective Are the California Common Tiers' Structure and Components and Elements at Defining and Measuring Quality in Early Learning Settings?

To address this research question, we conducted (1) a content review to determine the effectiveness of the California Common Tiers' QRIS rating structure and elements at *defining* quality in early learning settings (reported in Chapter 3); and (2) an analysis of concurrent validity to determine the effectiveness of the rating structure and elements at *measuring* quality (reported in Chapter 5).

Defining Quality: Content Review

California's Hybrid Rating Matrix covers a broad range of important domains of quality in early care and education settings. Moreover, a research base and precedent exists among other QRISs for the inclusion of the current elements. For example, the Hybrid Rating Matrix includes three of the five most common indicators used across states with QRISs: staff qualifications; program environment; and program administration, management, and leadership. California's QRIS also includes child observations (as do 55 percent of other systems) and teacher-child interaction (as do 48 percent of other QRISs).

However, more than three quarters of QRISs in other states also include curriculum in their rating systems, and a growing number of states require alignment of curricula with state early learning foundations. The CAEL QIS Advisory Committee (2010) recommended aligning curricula with the *California Preschool Learning Foundations,* the *California Preschool Curriculum Framework,* and the *California Infant/Toddler Learning and Development Foundations* as an alternative to recommending a specific list of curricula. However, this recommendation was not included in the Hybrid Rating Matrix, but rather was moved to the Pathways.

In addition, although 93 percent of state QRISs in 2014 had family partnership as a separate element in their rating system, in California's cross-county system, family partnership is only

included as part of the Program Environment Rating Scale element (e.g., as a subscale for family involvement on the ERS). The RTT-ELC Continuous Quality Improvement Pathways also includes family involvement, but the elements of this document do not count toward points in the Hybrid Rating Matrix.

Other aspects of quality, such as support for DLLs, cultural competency, and support for children with special needs, are also used in other systems and worthy of attention in the broader QRIS framework and Pathways.

Measuring Quality: Concurrent Validity

The primary approach that researchers use to determine how well a rating measures quality is to assess concurrent validity results for the rating and its elements. Concurrent validity studies examine the extent to which ratings are associated with a program's average scores on other independent measures of program quality. For this study, we used the CLASS and the PQA as the independent measures of quality. Although the CLASS is also included in the QRIS rating, it is the most predictive of children's outcomes (Howes and others 2008; Mashburn and others 2008) and is widely used to validate QRIS ratings. The PQA is an instrument that measures similar constructs to California's QRIS but is not included in the rating calculation.

Comparisons of the California QRIS ratings against independent measures of program quality reveal some encouraging relationships. First, ratings are significantly and positively related to CLASS total scores in centers, suggesting that the ratings capture effective teacher-child interactions well. It is important to note, though, that differences in average CLASS score from one rating level to the next are not all statistically significant; the fact that few programs are at high and low ends of the rating scale may be contributing to this pattern. Second, and more specifically, California QRIS ratings are positively related to CLASS Instructional Support scores and PQA Adult-Child Interaction scores among centers with preschool classrooms. This finding is encouraging because the CLASS Instructional Support domain is the one most predictive of children's outcomes (Howes and others 2008), and the PQA Adult-Child Interaction subscale captures similar behaviors.

The other CLASS domains and the overall PQA scores are not significantly related to California QRIS ratings in centers. In addition, for FCCHs, relationships with California QRIS ratings were largely positive for PQA scores and mixed for CLASS scores, although conclusions cannot be drawn from these results because of the limited number of FCCHs included in the analysis.

When we examine the element scores and how they relate to the external classroom observation measures, we find that the Effective Teacher-Child Interaction (CLASS) element and the Program Environment Rating Scale (ERS) element are consistent in significantly predicting classroom observation scores. That is, programs with higher scores on each of these elements are also found to have higher independent CLASS and PQA scores. This is to be expected, given that the concurrent validity measures—the CLASS and the PQA—are most aligned with the CLASS and ERS elements. The other elements are not consistently predictive of classroom observation scores.

Thus at this early stage, with the limited number of programs with full ratings, some evidence suggests that the system produces valid ratings, especially where process quality measures are concerned, at least among the limited range of programs currently participating in the system.

RQ 2. Do Point Values of Each Element and the Final Rating Provide Meaningful Distinctions Between Programs and Program Types?

To assess the extent to which element scores and ratings distinguish programs and program types, we examined the distribution of ratings for centers and FCCHs and compared ratings for programs with different characteristics. First, as noted at the beginning of this chapter and described in more detail in Chapter 4, the distribution of QRIS ratings is constrained. In fact, among the sample of programs with full ratings, the range of ratings observed does not span all five possible QRIS rating levels—no programs are rated Tier 1 and very few are rated Tier 2 or Tier 5. This truncated range in ratings poses a critical limitation on the validation of the system. Limited variation in ratings constrains our ability to differentiate programs based on their quality.

In addition, the distribution of ratings differs markedly for centers and FCCHs, and the distribution of ratings is even more limited within each program type. The most common rating for centers is Tier 4 (52 percent), and 86 percent of centers were rated at Tiers 3 or 4. In contrast, the most common rating for FCCHs is Tier 2 (53 percent), and 85 percent of FCCHs were rated at Tiers 2 or 3. Differences in ratings between centers and FCCHs may be partially explained by differences in the percentage of centers (95.8 percent), and FCCHs (42.7 percent) that are required to meet high quality standards for State Preschool, Child Signature Program, or Head Start funding.

The distribution of scores varies by element and also shows different patterns for centers and FCCHs. Among centers, variation is particularly limited among structural quality element scores; thus, these elements may not differentiate programs well. There is more variability in structural quality element scores among FCCHs.

Although we examined a variety of program characteristics, only CSP funding and Title 5 (State Preschool, General Child Care, or CalSAFE) funding are statistically significant predictors of California QRIS rating level among centers. This is likely due to the contract requirements that go along with these funding sources; they put programs receiving these funds in higher tiers. None of the program characteristics examined significantly predicted QRIS rating among FCCHs, but as noted previously, the small number of FCCHs reduces the likelihood of finding such effects.

RQ 3. Do Element Levels Relate to Each Other in Consistent Ways (e.g., CLASS/ERS Scores and Their Relationship to Other Elements)?

To explore how elements relate to each other, we examined correlations among element scores and between element scores and the overall rating for the 472 sites with full ratings, and also the internal consistency of the ratings. As reported in Chapter 4, we find that none of the element scores were redundant, indicating that the elements capture different aspects of program quality. Indeed, some pairs of elements have very low correlations; this is true among centers and FCCHs alike. These low correlations are reflected in low internal consistency of the overall QRIS rating,

indicating that the QRIS ratings do not represent a single type of quality, but rather represent diverse types of program quality.

Elements with limited variability tend to be weakly related to the overall QRIS rating, and to other element scores. Among centers, the Ratios and Group Sizes element, which has limited variation, is weakly correlated with the overall QRIS rating. Both the Ratios and Group Sizes element and the Developmental and Health Screenings element are weakly related to most other elements. Among FCCHs, the Effective Teacher-Child Interactions (CLASS) element—again, limited in variability— is the element most weakly correlated with the overall QRIS rating and with other elements. These results indicate that, in this sample of programs in which overall ratings are fairly homogeneous, some elements contribute more to the overall rating than others.

Internal consistency of the QRIS ratings is low, particularly among centers, as expected. The low internal consistency reflects the weak relationships between some pairs of elements. Indeed, the internal consistency would increase for centers if the ratings were calculated without the two element scores that have low correlations with other elements: Ratios and Group Sizes and Developmental and Health Screenings. Low internal consistency does not suggest that the rating is flawed, but rather that the aspects of quality measured for the QRIS are not always closely related to each other. These findings confirm that the California QRIS does not represent an overarching construct of program quality that is unidimensional. Again, with a more diverse group of programs in the system, it is possible that we might see more variability among element scores, which in turn might improve the internal consistency and the value of those elements not currently contributing a great deal to overall ratings.

RQ 4. How Is the Hybrid Rating Strategy and Rating Outputs Representative of Meaningful Levels of Quality?

California's Hybrid Rating Matrix combines a building blocks approach and a points approach, with the first tier blocked and the provision of points for the remaining four tiers. In addition, participating county Consortia have the option to use a block rating structure for Tier 2 as well. As described in Chapter 3, although the building block approach—which requires programs to meet all of the criteria in one tier before they can move up and attain the next quality rating level—remains the most common rating structure among QRISs, hybrid and points approaches have gained in popularity in the last few years (QRIS Online Compendium 2014).

To examine the extent to which California's hybrid rating strategy appropriately differentiates programs in terms of quality, we conducted sensitivity analyses by comparing concurrent validity results for each of several different rating strategies, including the current Hybrid Rating Matrix approach (reported in Chapter 6). That is, we compared the extent to which the rating levels are associated with the program's average scores on other independent measures of program quality when using different approaches to calculating the ratings (e.g., the block versus points rating approach).

We found that the distribution of rating levels varies by rating approach. Although a block system has advantages—such as simplicity and transparency—the differences in the distribution of ratings compared to the recommended Hybrid Rating Matrix approach are substantial when blocks are used for more tiers than only at Tier 1. In fact, as the number of rating levels that are

blocked increases, program ratings tend to decrease. For example, when blocks are used for Tiers 1 and 2, nearly one quarter of centers receive lower ratings compared with the California QRIS rating approach. Even programs rated as high as Tier 4 in the California QRIS ratings were rated at Tier 1 when Tier 2 was blocked. More than half of centers receive lower ratings when Tiers 1, 2, and 3 are blocked, and more than 90 percent receive lower ratings when all five tiers are blocked. The differences are even more dramatic among FCCHs. Concurrent validity results for models with more blocks were also mixed. We found that ratings using blocks are less effective than California QRIS ratings at differentiating centers by CLASS scores, though five-level blocks are more effective at differentiating centers according to the PQA observation scores.

We also considered ratings calculated by averaging across the element scores. Although not a common approach to calculating a rating among QRISs across states, the element average appears to be the most effective strategy. Specifically, element average ratings are more effective than California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores.

Given the significant costs and challenges that many Consortia have faced in completing the classroom observations—especially with regard to the ERS—required to calculate the ratings, we considered how the ratings might perform if the ERS was not required except to reach Tier 5, and the Program Environment Rating Scale element was treated as a block for Tier 5. This "ERS hybrid rating" approach produced ratings very similar to the California QRIS ratings in their patterns of relationships with classroom observation scores, suggesting this approach would not dramatically alter the validity of the rating scale.

It is important to remember, though, that the results of the concurrent validity analyses using alternative rating approaches are specific to the sample of centers included in the study, and given the narrow range of programs currently participating in the system, the relationships between alternative rating approaches and observed quality scores may differ for other programs in California.

RQ 5. How Do QRIS Ratings That Use Locally Determined Tiers Differ From QRIS Ratings Calculated Using Recommendations in California's RTT-ELC QRIS Implementation Guide?

As noted in Chapter 2, few Consortia opted to make modifications to Tier 2 of the Hybrid Rating Matrix, although the majority of Consortia opted to maintain the common criteria for Tier 2. Many more made (or were planning to make) modifications to Tier 5. In total, local adaptations were used in six of the 12 counties with final QRIS ratings. Making such changes has the potential to alter how any given program might be rated. To assess the extent to which such modifications affected ratings, we compared ratings calculated by Consortia using their local adaptations with ratings for the same programs calculated using the RTT-ELC recommended approach (reported in Chapter 6). The distribution of the Consortia-calculated ratings is quite similar to the California QRIS ratings, especially for centers. In fact, across all 12 counties with full ratings, 95 percent of centers and 87 percent of FCCHs have the same rating in both the Consortia QRIS ratings and the California QRIS ratings. Where ratings differ, the California QRIS ratings are usually higher than the Consortia QRIS ratings.

The Consortia ratings are also similar to the California QRIS ratings in their patterns of relationships with classroom observation scores. The local Consortia QRIS ratings are significantly predictive of CLASS instructional support and PQA adult-child interaction scores, but unlike the California QRIS ratings, they are not significantly related to CLASS total scores. Thus, it appears that the local adaptations used to date are not having a significant impact on the distribution and validity of the ratings.

RQ 6. How Effective Is the Rating Protocol at Determining Valid Ratings Versus an Annual 100 Percent Assessment Protocol?

In addition to considering different approaches to calculating the ratings, we considered how different the ratings might look if a different protocol for conducting classroom observations were used. Currently, Consortia are asked to observe a sample of classrooms—approximately one third—across age groups. The majority of the Consortia were following these guidelines for sampling classrooms at the site level. However, as noted in Chapter 2, a few Consortia conducted observations in all classrooms to ensure that an accurate picture of quality was captured. To test the value of conducting observations in additional classrooms, we conducted observations in all classrooms for a subset of programs participating in the QRIS and compared ratings and element scores calculated using a one half of all classrooms protocol and a one third of all classrooms protocol against a 100 percent of classrooms protocol (reported in Chapter 6).

Among the small number of centers in the sample used for this analysis, we found that classroom observation scores—especially CLASS scores—vary more within centers in our sample than from center to center, which is an argument for observing more classrooms to ensure the rating reflects the program as a whole. Among this small sample of programs, Teacher-Child Interaction element scores (based on the CLASS) differ in about a third of programs depending on how many classrooms' CLASS scores are included in the element scoring. Element scores based on the ERS were rarely affected by selecting different combinations of classrooms in the one third or one half classroom observation protocols, however. That is, the Program Environment Rating Scale element score was fairly stable, regardless of how many or which combination of classrooms' ERS scores were included in the element score calculation. In addition, although CLASS element scores derived from samples of one third or one half of classrooms differed from the element scores derived from observations of *all* classrooms, QRIS ratings among centers in our sample were rarely affected by these differences. Therefore, although the evidence is limited, this result suggests that although CLASS element scores might fluctuate depending on the sampling approach taken, the overall ratings are fairly robust.

Preliminary Conclusions and Limitations

It is early to draw firm conclusions about the validity of the system, particularly because validation work continues. However, in this section we summarize the evidence to date for the system's validation, and highlight important limitations to consider when interpreting the study's findings.

There Is Some Evidence Supporting the Validity of the QRIS Ratings

An analysis of the content validity of the QRIS rating finds an evidence base for the elements of the Hybrid Rating Matrix, with some elements, such as the Teacher-Child Interactions element having a stronger evidence base than other elements. There is some evidence of concurrent validity as well. QRIS ratings for centers are significantly and positively related to independent measures of quality: CLASS total scores, CLASS Instructional Support scores, and PQA Adult-Child Interaction scores. Other measures, such as the PQA total score and other domain scores from the PQA and CLASS, were not significantly related to QRIS ratings. In addition, no statistically significant relationships between QRIS ratings and independent observation scores were observed among FCCHs, though this lack of association is likely due to the small sample size. When we examined the relationships between element scores and the independent quality observations, we found that only the two elements relying on observational tools, the Teacher-Child Interaction (CLASS) element and the Program Environment Rating Scale (ERS) element, predicted CLASS and PQA scores; this may reflect in part similarity in measurement approaches. Thus, some evidence indicates that the system produces valid ratings, especially where process quality measures are concerned.

Important Limitations of the Study Mean That Conclusions Should Be Considered Preliminary

It is important to remember several key study limitations that constrain the extent to which firm conclusions can and should be drawn. Instead, study findings should be considered preliminary. First, the system is new and still in the development and refinement stage. Moreover, in this early stage of implementation, a relatively small pool of sites have full ratings. Of the 1,273 sites in the rating system, only 472 have full nonprovisional ratings, and these sites differ from fully rated sites in terms of funding source and home language use as well as rating—with provisionally rated sites scoring lower on average. Variation in ratings is also limited. With no sites rated Tier 1 and only a handful rated Tiers 2 or 5, it is difficult to assess the extent to which the rating adequately differentiates programs based on quality. With a more diverse pool of sites, the results might look different.

Second, given the smaller pool of fully rated sites eligible for the study and delays with data collection start up, the sample of sites for concurrent validity analyses was smaller than anticipated, thus limiting inferences. Small sample sizes make it difficult to detect small relationships. That is, some analyses might miss potentially significant differences that would be detected with a larger sample size. The sample of FCCHs was especially small, making it impossible to draw conclusions about these programs.

Third, it is also important to remember that a third aspect of validity, predictive validity, will be examined through the evaluation in 2015. By examining outcomes for children participating in sites with different ratings, we can better understand how well ratings predict growth and development for children. Determinations of validity should take into account results from these analyses as well.

Policy Options for Consideration

Although study results cannot be considered conclusive at this stage, our analyses do suggest some directions that may be worth consideration by the state, at least in a preliminary way. In this section we offer some suggestions for modifications to the system that the state might want to consider in light of the evidence and other contextual factors.

Consider Ways to Increase Attention to Curriculum, Family Engagement, and Special Populations in the Hybrid Rating Matrix

As noted previously, the state may wish to consider adding to the Hybrid Rating Matrix alignment of curricula to the *Foundations* and *Frameworks*, as was recommended by the CAEL QIS Advisory Committee. This alignment can provide greater assurance that teachers and providers are aware of the *Foundations* and *Frameworks* and are learning how to incorporate them into their instruction; that children in participating Consortia are receiving instruction consistent with the frameworks developed for the state; and PK–3 alignment for teachers and students in California is being supported by the QRIS. A curriculum-alignment standard could include a tiered progression beginning with a simple requirement to have an education plan with a philosophy statement in the first tier to having a plan with all domains linked to child assessments and a professional development plan including training on the *Foundations* and *Frameworks* in the higher tiers. Adding this element would require Consortia to review program policies for compliance with the provision. Some states provide a list of curricula determined to be aligned with educational standards; employing one of these curricula meets requirements. Participating programs that prefer to use a different curriculum are required to demonstrate how their curriculum is aligned.

Adding specific reference to the *Foundations* and *Frameworks* in the Hybrid Rating Matrix would also help to address some of the other domains not fully represented in the Hybrid Rating Matrix, such as Cultural and Linguistic Diversity, DLL, Cultural Competency, Special Needs, and Health Practices. These program elements have some theoretical foundation and are important, but either the research does not exist yet to show whether they do or do not link to improvements in teaching practices or children's outcomes, or there is little consensus on how to measure the elements. For programs that serve DLLs, children from diverse cultural backgrounds, and children with special needs, the literature points to key features of successful programs, which should be reviewed when considering the addition of these elements. Adding alignment of curricula with the *Foundations* and *Frameworks* to the Hybrid Rating Matrix would offer a modest approach to addressing these issues.

In terms of family engagement, although the literature supports the importance of family engagement in early childhood programs, there is little consensus on how best to measure it (AIR and RAND 2013). The ERS, already in use, includes a subscale on *Parents and Staff,* although it is not a comprehensive measure of family engagement. A new measure—the *Family and Provider/Teacher Relationship Quality* measure (Kim and others 2014)—which has recently been developed, assesses site staff's knowledge, practices, and attitudes around family engagement. Although the measure has not yet been validated, the state might wish to explore the use of the tool as one option for addressing family engagement.

Consider Alternative Rating Strategies to Strengthen Validity or Simplify Implementation

Although some evidence supports the validity of the Hybrid Rating Matrix in its current form, our analyses shed light on ways to strengthen or simplify the rating approach that the state might consider. For example, ratings calculated by taking an average score across elements are more effective than the California QRIS ratings at differentiating centers by CLASS and PQA classroom observation scores. California's decision makers may wish to consider this as a simple alternative to the current rating strategy. However, it is important to remember that these concurrent validity results might change when programs with a wider distribution of ratings are included in the analytic sample.

The state also might consider modifying the ERS element in light of implementation challenges consistently experienced across Consortia. The ERS is a difficult and costly instrument on which to train and maintain a cadre of observers, and reducing this burden for Consortia will likely result in more fully rated programs. We explored one option for doing so in our analyses: limiting the requirement for the ERS to Tier 5 and blocking at that tier. Results suggest that this change would have minimal impact on ratings. Although the validation study results did not test whether eliminating the ERS element would improve validity, they do indicate that it might be possible to reduce its use dramatically without affecting ratings. Given the implementation challenges associated with using the ERS, it may be wise to consider ways to reduce its use, especially if the system is to be sustained long term.

Consider Options for the Presentation of Ratings Information to Parents

Given the multidimensional nature of the Hybrid Rating Matrix, the positive results for the CLASS element, and the potential value of providing parents with more specific information that they can use in making care decisions, the state might consider presenting some or all element scores or subratings along with summary ratings once the ratings become publicly available. This would enable parents to make finer distinctions between programs that might share the same or similar QRIS rating. The multidimensional nature of the rating and the fact that different rating elements measure different program components means that two programs with the same rating may actually have important underlying differences, reflecting varying strengths and weaknesses on different elements. Although the original intent of a hybrid rating system was to provide programs some flexibility in how they could reach certain levels of quality, in practice it makes comparing programs with the same ratings problematic. Moreover, parents may value some rating elements more than others; element scores would enable parents to focus their search on programs that rate highest on the elements about which they may care most.

Other Considerations Relevant for Further Expansion of the System and Its Validation

In addition, though not directly arising from the validation study results, the state may want to explore other considerations relevant to further research and validation. To support continuous quality improvement in the QRIS, the state may want to consider ways to expand the system and may also want to consider supporting another validation phase when the QRIS is more mature and more programs are in the system.

Consider Ways to Encourage or Require More Providers to Participate in the System

Perhaps the most important issue for the state to consider is whether ratings should be voluntary or required and, if required, for which programs. Although the validation analyses do not directly address this, we note frequently in this report that one of the major limitations of this research has been the relative lack of variation in the sample of programs participating in the study. The majority of programs and providers participating have been at Tiers 3 or 4, with no programs from Tier 1 and only a few at Tiers 2 or 5. Moreover, the sample is heavily skewed toward state and federally contracted programs that were already held to a set of contract standards intended to focus on quality before the implementation of the RTT-ELC QRIS. The lack of variation is not just a problem for researchers attempting to gauge the effectiveness of the system in rating quality; the narrow range of programs participating also limits the potential impact of the QRIS in providing information to families choosing care for their young children. It also forgoes an opportunity to assess the quality of the large group of private programs receiving some public funds in the form of vouchers, and makes it difficult for the public or policymakers to determine how best to direct limited resources for quality improvement.

The state might, therefore, want to consider piloting a system in one or more counties that requires all centers and FCCHs receiving state and federal subsidies to participate in the QRIS. At least nine states require programs receiving subsidies from the federal Child Care and Development Fund to participate in their QRIS, and several states, such as Illinois and Washington, make participation mandatory for school-operated early care and education programs. Another potential benefit of piloting a QRIS that requires participation by all publicly funded providers would be the information that it would give policymakers about the current quality of the programs in which taxpayers are investing and where and what type of improvements are needed. Finally, such a pilot would provide a more complete picture of the extent to which the rating system captures the distinctions between all five tiers in the Hybrid Rating Matrix.

Of course, if participation were mandatory, it would be important to ensure that programs had access to program quality assessments so that all sites could be assessed and receive a full—as opposed to provisional—rating. Establishing a process that would ensure such access to newly mandated programs would be an important part of a mandatory participation pilot before statewide implementation could be considered.

Consider Another Validation Phase Once the System Is Further Developed

As noted throughout this report, data limitations due in part to the QRIS's stage of development constrain the analyses and limit the generalizability of the results. To address this constraint, the state might consider revisiting system validation once refinements currently under discussion are made and once the system is expanded to include a more diverse array of programs. If further analyses are to be conducted, it would be essential for Consortia to collect, maintain, and share with the state additional classroom- and site-level data. Such data would enable additional analyses and suggest evidence-based refinements; this work would not be possible without these more detailed data. In particular, it would be helpful to have raw element-level data (e.g., ratios, ERS scores). In addition to being useful for accountability purposes, retaining these data would

permit the examination of element score cut points and the simulation of ratings based on modified cut points in order to refine the element scoring criteria. Such refinements would strengthen the reliability and validity of the ratings, making the QRIS a more meaningful signal of quality for parents and a more effective tool for targeting quality improvement resources.

References

AERA, APA and NCME. 1999. *Standards for educational and psychological tests*. Washington, DC: American Educational Research Association.

American Institutes for Research and RAND Corporation. 2013. *Local Quality Improvement Efforts and Outcomes Descriptive Study: Final Report*. Sacramento, CA: California Department of Education.

Barnard, W., 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.

Barnett, W. S. 2003. Better Teachers, Better Preschools: Student Achievement Linked to Teacher Qualifications. *Preschool Policy Matters, 2*. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, W. S., and others. 2007. "Two-way and Monolingual English Immersion in Preschool Education: An Experimental Comparison." *Early Childhood Research Quarterly* 22, no. 3: 277–93.

Barraclough, S. J., and A. B. Smith. 1996. "Do Parents Choose and Value Quality Child Care in New Zealand?" *International Journal of Early Years Education* 4: 5–26.

Boller, K., and others. 2010. *The Seeds of Success Modified Field Test: Findings from the Impact and Implementation Studies*. Princeton, NJ: Mathematica Policy Research.

Bromer, J., and others. 2011. "Family-sensitive Caregiving: A Key Component of Quality in Early Care and Education Settings," In *Quality Measurement in Early Childhood Settings*. Edited by M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle. Baltimore, MD: Brookes.

Bryant, D. M., and others. 2001. *Validating North Carolina's 5-Star Child Care Licensing System.* Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Center. http://fpg.unc.edu/resources/validating-north-carolinas-5-star-child-care-licensing-system (accessed March 15, 2013).

Bryant, D., and others. 2009. *The QUINCE-PFI Study: An Evaluation of a Promising Model and Delivery Approaches for Child Care Provider Training*. Draft Final Report. Chapel Hill, NC: University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Institute.

Bumgarner, E., and J. Brooks-Gunn. 2011. *Latino American Children and School Readiness: The Role of Early Care Arrangements and Caregiver Language*. New York: Teachers College, Columbia University. http://academiccommons.columbia.edu/catalog/ac%3A170988 (accessed April 3, 2015).

Burchinal, M. R., K. Kainz, and Y. Cai. 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 *Quality Measurement in Early Childhood Settings*. Edited by M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle. Baltimore, MD: Brookes.

Burchinal, M. R., and others. 1996. "Quality of Center C*hild* Care and Infant Cognitive and Language D*evelopment*." *Child Development* 67: 606–20.

Burchinal, M., and others. 2002 ."Caregiver Training and Classroom Quality in Child Care Centers." *Applied Developmental Science* 6: 2–11.

Burchinal, M., and others. 2010. "Threshold Analysis of Association Between Child Care Quality and Child Outcomes for Low-income Children in Pre-kindergarten Programs." *Early Childhood Research Quarterly* 25 no. 2: 166–76

Burchinal, M., and others. 2012. "Instruction in Spanish in Pre-kindergarten Classrooms and Child Outcomes for English Language Learners." *Early Childhood Research Quarterly* 27, no. 2: 188–97. http://eric.ed.gov/?q=%22Instruction+in+Spanish+in+pre-kindergarten+classrooms%22&id=EJ958040 (accessed April 3, 2015).

.Burchinal, M., and others. 2014. "Thresholds in the Association Between Child Care Quality and Child Outcomes in Rural Preschool Children." *Early Childhood Research Quarterly* 29, no. 1: 41–51.

Buysse, V., and others. 2013. "Effects of Early Education Programs and Practices on the Development and Learning of Dual Language Learners: A Review of the literature." *Early Childhood Research Quarterly* 29: 765–85.

California Early Learning Quality Improvement System (CAEL QIS) Advisory Committee. 2010. *Dream Big for Our Youngest Children*: *California Early Learning Quality Improvement System Advisory Committee Final Report.* Sacramento, CA: California Department of Education. http://www.cde.ca.gov/sp/cd/re/documents/fnlrpt2010.pdf (accessed April 3, 2015).

Castro, D. C. 2005. *Early Language and Literacy Classroom Observation. Addendum for English Language Learners.* Chapel Hill, NC: The University of North Carolina Frank Porter Graham Child Development Institute.

Castro, D. C., E. E. Garcia, and A. M. Markos. 2013. *Dual Language Learners: Research Informing Policy.* Chapel Hill, NC: The University of North Carolina Frank Porter Graham Child Development Institute, Center for Early Care and Education—Dual Language Learners.

Center on the Developing Child. 2007. *A Science-based Framework for Early Childhood Policy: Using Evidence to Improve Outcomes in Learning, Behavior, and Health for Vulnerable Children.* Cambridge, MA: Harvard University. http://developingchild.harvard.edu/resources/reports_and_working_papers/policy_framework/ (accessed April 3, 2015).

Clarke-Stewart, K. A., and others. 2002. "Do Regulable Features of Child-care Homes Affect Children's Development?" *Early Childhood Research Quarterly* 17, no. 1: 52–86.

Clements, D. H., and J. Sarama. J. 2007. "The Effects of Preschool Mathematics Curriculum: Summative Research on the Building Blocks Project." *Journal for Research in Mathematics Education* 38, no. 2: 136–63.

Clifford, R. M., S. S. Reszka, and H. G. Rossbach. 2009. *Reliability and Validity of the Early Childhood Environment Rating Scale—Draft Version of a Working Paper*. Chapel Hill, NC: University of North Carolina at Chapel Hill Frank Porter Graham Child Development Institute.

Cryer, D. 2014. *ECERS-3: A New, Updated Early Childhood Environment Rating Scale for the 21st Century*. The BUILD Initiative and QRIS National Learning Network "Let's Talk" webinar presentation. http://qrisnetwork.org/category/event-type/webinar (accessed April 3, 2015).

Cryer, D., and M. Burchinal. 1997. "Parents as Child Care Consumers." *Early Childhood Research Quarterly* 12: 35–58.

Cryer, D., T. Harms, and C. Riley. 2003. *All About the ECERS-R: A Detailed Guide in Words & Pictures to Be Used with the ECERS-R*. PACT House Publishing.

Cryer, D., W. Tietze, and H. Wessels. 2002. "Parents' Perceptions of Their Children's Child Care: A Cross-national Comparison." *Early Childhood Research Quarterly* 17: 259–77.

Cryer, D. and others. 1999. "Predicting Process Quality From Structural Quality In Preschool Programs: A Cross-country Comparison." *Early Childhood Research Quarterly 14*, no. 3: 339–61.

Durán, L. K., C. J. Roseth, and P. Hoffman. 2010. "An Experimental Study Comparing English-only and Transitional Bilingual Education on Spanish-speaking Preschoolers' Early Literacy Development." *Early Childhood Research Quarterly* 25, no. 2: 207–17. http://eric.ed.gov/?q=%22An+experimental+study+comparing+English-only+and+Transitional+Bilingual%22&id=EJ874836 (accessed April 3, 2015).

Early, D., and others. 2007. "Teachers' Education, Classroom Quality, and Young Children's Academic Skills: Results from Seven Studies of Preschool Programs." *Child Development* 78, no. 2: 558–80.

Elicker, J., and others. 2011. *Evaluation of Paths to QUALITY, Indiana's Child Care Quality Rating and Improvement System: Final Report*. West Lafayette, IN: Purdue University.

Farver, J. A. M., C. J. Lonigan, and S. Eppe. 2009. "Effective Early Literacy Skill Development for Young Spanish-speaking English Language Learners: An Experimental Study of Two Methods." *Child Development* 80, no. 3: 703–19. http://eric.ed.gov/?q=%22Effective+early+literacy+skill+development+for+young+Spanish-speaking%22&id=EJ840084 (accessed April 3, 2015).

Fuller, B., and others. 2004. "Child Care Quality: Centers and Home Settings That Serve Poor Families." *Early Childhood Research Quarterly 19*: 505–27.

Garcia, O., J. Kleifgen, and L. Falchi. 2008. *From English Language Learners* to *Emergent Bilinguals: Equity Matters: Research Review No. 1*. New York: Campaign for Educational Equity, Teachers College, Columbia University. http://files.eric.ed.gov/fulltext/ED524002.pdf (accessed April 3, 2015).

Goelman, H., and others. 2006. "Towards a Predictive Model of Quality in Canadian Child Care Centers." *Early Childhood Research Quarterly 21*, no.3: 280–95.

Goode, S., M. Diefendorf, and S. Colgan. 2011. *The Outcomes of Early Intervention for Infants and Toddlers with Disabilities and Their Families*. Washington, DC: The National Early Childhood Technical Assistance Center. http://ectacenter.org/~pdfs/pubs/outcomesofearlyintervention.pdf (accessed April 3, 2015).

Gordon, R. A., and others. 2013. "An Assessment of the Validity of the ECERS-R with Implications for Measures of Child Care Quality and Relations to Child Development." *Developmental Psychology 49*, no. 1: 146–60.

Gormley, W. T., and T. Gayer. 2005. "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources 40*, no. 3: 533–58.

Gormley, W. T., and others. 2005. "The Effects of Universal Pre-K on Cognitive Development." *Developmental Psychology 41*, no. 6: 872–84.

Greenwood, C., and others. 2011. "Program-level Influences on the Measurement of Early Communication of Infants and Toddlers in Early Head Start." *Journal of Early Intervention 33*, no. 2: 110–34.

Grehan, A. W., and L. J. Smith. 2004. *The Early Literacy Observation Tool.* Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Guevara, J. P., and others. 2013. Effectiveness of developmental screening in an urban setting. *Pediatrics, 131*(1), 30–37.

Halle, T., J. E. V. Whittaker, and R. Anderson. 2010. *Quality in Early Childhood Care and Education Settings: A Compendium of Measures* (2nd. ed.). Washington, DC: Child Trends. Prepared by Child Trends for the U.S. Department of Education, Administration for Children and Families, Office of Planning, Research and Evaluation.

Harms, T., R. M. Clifford, and D. Cryer. 1998. *Early Childhood Environment Rating Scale—Revised Edition*. New York: Teachers College Press.

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

Harms, T., D. Cryer, and R. M. Clifford. 2006. *Infant/Toddler Environment Rating Scale—Revised Edition* (ITERS-R). 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.

Hegland, S. M., and others. 2011. "Measuring Health-related Aspects of Quality in Early Childhood Settings," in *Quality Measurement in Early Childhood Settings*. Edited by M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle. Baltimore, MD: Brookes.

Helburn, S. W., ed. 1995. *Cost, Quality, and Child Outcomes in Child Care Centers: Technical Report*. Denver, CO: University of Colorado at Denver, Department of Economics, Center for Research in Economic and Social Policy.

Helburn, S. W., J. R. Morris, and K. Modigliani. 2002. "Family Child Care Finances and Their Effect on Quality and Incentives." *Early Childhood Research Quarterly 17*: 512–38.

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

HighScope Educational Research Foundation. 2009. *Family Child Care Program Quality Assessment (PQA) Administration Manual*. Ypsilanti, MI: HighScope Press.

Hofer, K. G. 2010. "How Measurement Characteristics Can Affect ECERS-R Scores and Program Funding." *Contemporary Issues in Early Childhood 11*, no. 2: 175–91.

Hohmann, M., S. Lockhart, and J. Montie. 2013. *Infant-Toddler Program Quality Assessment (PQA) Form A: Observation Items.* Ypsilanti, MI: HighScope Press.

Howes, C. 1988. "Relations Between Early Child Care and Schooling." *Developmental Psychology 24*: 53–7.

Howes, C., and others. 2008. "Ready to Learn? Children's Pre-academic Achievement in Pre-kindergarten Programs." *Early Childhood Research Quarterly 23*, no. 1: 27–50.

Isner, T., and others. 2011. *Coaching in Early Care and Education Programs and Quality Rating and Improvement Systems (QRIS): Identifying Promising Features*. Washington, DC: Child Trends.

Karoly, L. A. 2009. *Preschool Adequacy and Efficiency in California: Issues, Policy Options, and Recommendations*. Santa Monica, CA: RAND Corporation. http://www.rand.org/pubs/monographs/MG889.html (accessed April 3, 2015).

Karoly, L., and G. Zellman. 2012. *How Would Programs Rate Under California's Proposed Quality Rating and Improvement System*? *Evidence from Statewide and County Data on Early Care and Education Program Quality.* Santa Monica, CA: RAND Corporation.

Kim, K., and others. 2014. *Family and Provider/Teacher Relationship Quality Measures: User's Manual* (OPRE Report 2014-65). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

La Paro, K. M., B. K. Hamre, and R. C. Pianta.2012. *Classroom Assessment Scoring System (CLASS) Manual: Toddler*. Baltimore, MD: Brookes.

La Paro, K. M., and others. 2012. "Examining the Definition and Measurement of Quality in Early Childhood Education: A Review of Studies Using the ECERS-R From 2003 to 2010." *Early Childhood Research and Practice 14.* no. 1.

Lahti, M., and others. 2011. *Maine's Quality for ME–Child Care Quality Rating and Improvement System (QRIS): Final Evaluation Report*. Portland, ME: University of Southern Maine, Muskie School of Public Service, Cutler Institute for Health and Social Policy.

Lahti, M., and others. 2013. *Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States* (OPRE 2013-036). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. http://www.researchconnections.org/childcare/resources/26590/pdf (Accessed April 3, 2015).

Lambert, M. C., and others. 2008. "Are the Indicators for the Language and Reasoning Subscales of the Early Childhood Environment Rating Scales-Revised Psychometrically Appropriate for Caribbean Classrooms?" *International Journal for Early Years Education* 16, no. 1: 41–60.

Landry, S. H., and others. 2001. *Teacher Behavior Rating Scale* (unpublished research instrument). Houston, TX: Center for Improving the Readiness of Children for Learning and Education, University of Texas Health Science Center at Houston.

Le, V. N., D. D. Schaack, and C. M. Setodji. 2013. "Identifying Baseline and Ceiling Thresholds Within the Qualistar Early Learning Quality Rating and Improvement System." *Early Childhood Research Quarterly 30*, no. 2: 215–26.

Le and others. 2006. "Measuring Child-Staff Ratios in Child Care Centers: Balancing Effort and Representativeness." *Early Childhood Research Quarterly 21:* 267–69.

Leithwood, K., and others. 2004. *How Leadership Influences Student Learning.* New York: Wallace Foundation.

Li-Grining, C. P., and R. L. Coley. 2006. "Child Care Experiences in Low-income Communities: Developmental Quality and Maternal Views." *Early Childhood Research Quarterly 21*: 125–41.

Lopez, E. M. 2010. "Valuing Families as Partners." *Early Childhood News 3*, no. 1.

Lugo-Gil, J., and others. 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., 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. http://qrisnetwork.org/sites/all/files/resources/gscobb/2011-09-28%2014:15/Report.pdf (accessed March 15, 2013).

Mantzicopoulos, P. 2003. "Flunking Kindergarten After Head Start: An Inquiry into the Contribution of Contextual and Individual Variables." *Journal of Educational Psychology 95*, no. 2: 268–78.

Mashburn, A. J. 2008. "Quality of Social and Physical Environments in Preschools and Children's Development of Academic, Language, and Literacy Skills." *Applied Developmental Science 12*, no. 3: 113–27. http://eric.ed.gov/?q=%22Quality+of+social+and+physical+environments+in+preschools%22&id=EJ887977 (accessed April 3, 2015).

Mashburn, A., and others. 2008. "Measures of Classroom Quality in Pre-kindergarten and Children's Development of Academic, Language, and Social Skills." *Child Development* 79, no. 3: 732–49.

Mashburn, A. J., and R. C. Pianta. 2010. "Opportunity in Early Education: Improving Teacher-child Interactions and Child Outcomes," in *Childhood Programs and Practices in the First Decade of Life:* A *Human Capital Integration*. Edited by A. J. Reynolds, A. J. Rolnick, M. M. Englund, and J. A. Temple. Cambridge, UK: Cambridge University Press.

McWayne, C., and others. 2004. "A Multivariate Examination of Parent Involvement and the Social and Academic Competencies of Urban Kindergarten Children." *Psychology in the Schools 41,* no. 3: 363–77.

Miedel, W. T., and A. J. Reynolds. 1999. "Parent Involvement in Early Intervention for Disadvantaged Children: Does It Matter?" *Journal of School Psychology 37*, no. 4: 379–402.

Meisels, S. J., and others. 2001. "Trusting Teacher Judgments: A Validity Study of a Curriculum-embedded Performance Assessment in Kindergarten to Grade 3." *American Educational Research Journal 38*, no. 1: 73–95.

Missall, K., and others. 2007. "Examination of the Predictive Validity of Preschool Early Literacy Skills." *School Psychology Review 36*, no. 3: 433–53.

National Association for the Education of Young Children. 2009. *Quality Benchmark for Cultural Competency Project.* Washington, DC. http://www.naeyc.org/files/naeyc/file/policy/state/QBCC_Tool.pdf (accessed April 3, 2015).

National Association of Child Care Resource and Referral Agencies. 2011. *We Can Do Better, 2011 Update: NACCRRA's Ranking of State Child Care Center Regulations and Oversight.* Washington, DC: National Association of Child Care Resource and Referral Agencies. http://www.naccrra.org/publications/naccrra-publications/we-can-do-better-2011.php (accessed April 3, 2015).

National Education Goals Panel. 1998. *Principles and Recommendations for Early Childhood Assessments.* Washington, DC: National Education Goals Panel. http://govinfo.library.unt.edu/negp/reports/prinrec.pdf (accessed April 3, 2015)

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). 2000. "The Relation of *Child* Care to Cognitive and Language *Development*." *Child Development 74,* no. 4.

NICHD ECCRN. 2002. "Child-Care Structure → Process → Outcome: Direct and Indirect Effects of Child-care Quality on Young Children's Development." *Psychological Science 13*, no. 3: 199–206.

NICHD ECCRN. 2002. "Early Child Care and Children's Development Prior to School Entry: Results from the NICHD Study of Early Child Care." *American Educational Research Journal 39*, no. 1: 133–64.

National Research Council. 2001. *Eager to Learn: Educating Our Preschoolers.* Edited by B. T. Bowman, M. S. Donovan, and M. S. Burns, (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Survey of Early Care and Education Project Team. 2013. *Number and Characteristics of Early Care and Education (ECE) Teachers and Caregivers: Initial Findings from the National Survey of Early Care and Education (NSECE)* (OPRE Report #2013-38). Washington DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Norris, D. J., and L. Dunn. 2004. *Reaching for the Stars: Family Child Care Home Validation Study Final Report*. Stillwater and Norman, OK: Early Childhood Collaborative of Oklahoma.

Norris, D. J., L. Dunn, and L. Eckert. 2003. *Reaching for the Stars: Center Validation Study Final Report*. Stillwater and Norman, OK: Early Childhood Collaborative of Oklahoma.

Office of Planning, Research, and Evaluation. 2012. *Validating Quality Rating and Improvement Systems* (PowerPoint slides).

Peisner-Feinberg, E., and M. Burchinal. 1997. "Concurrent Relations Between Child Care Quality and Child Outcomes: The Study of Cost, Quality, and Outcomes in Child Care Centers." *Merrill-Palmer Quarterly 43*: 451-477.

Peisner-Feinberg, E. and others. (1999). T*he Children of the Cost, Quality, and Outcomes Study Go to School* (Technical Report). Chapel Hill: University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Center. http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/NCEDL_CQO_technical_report.pdf (accessed April 3, 2015).

Peisner-Feinberg, E. S., and others. 2001. "The Relation of Preschool Child-care Quality to Children's Cognitive and Social Developmental Trajectories Through Second Grade." *Child Development 72*, no. 5: 1534–53.

Peth-Pierce, R. 1998. *The NICHD Study of Early Child Care*. Washington, DC: National Institute of Child Health and Human Development.

Phillips, D., and others. 2000. "Within and Beyond the Classroom Door: Assessing Quality in Child Care Centers." *Early Childhood Research Quarterly, 15*. no. 4, 475-496.

Phillipsen, L. C., and others. 1997. "The Prediction of Process Quality from Structural Features of Child Care." *Early Childhood Research Quarterly 12*, no. 3: 281–303.

Pianta, R. C., K. M. La Paro, and B. K. Hamre. 2008. *Classroom Assessment Scoring System* (*CLASS*). Baltimore, MD: Brookes.

Pianta, R. C., and others. 2009. "The Effects of Preschool Education: What We Know—How Public Policy Is or Is Not Aligned with the Evidence Base, and What We Need to Know." *Psychological Science in the Public Interest* 10, no. 2: 49–88.

Ponitz, C. C., and others. 2009. "Early Adjustment, Gender Differences, and Classroom Organizational Climate in First Grade." *The Elementary School Journal 110*, no. 2: 142–62.

QRIS National Learning Network. 2013. *QRIS State Contacts & Map: Revised May 2013.* http://www.qrisnetwork.org/qris-state-contacts-map (accessed April 3, 2015).

QRIS Online Compendium. 2014. *A Catalog and Comparison of Quality Rating and Improvement Systems*.

Rigby, E., R. M. Ryan, and J. Brooks-Gunn. 2007. "Child Care Quality in Various State Policy Contexts." *Journal of Policy Analysis and Management 26*: 887–907.

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.

Sabol, T., and R. Pianta. 2014. "Validating Virginia's Quality Rating and Improvement System Among State-Funded Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 30: 183–198. http://qrisnetwork.org/sites/all/files/materials/Validating%20Virginia%E2%80%99s%20quality%20rating%20and%20improvement%20system%20among%20state-funded%20pre-kindergarten%20programs%20(Early%20Childhood%20Research%20Quarterly).pdf (accessed April 3, 2015)

Sabol, T. J., and others. 2013. "Can Rating Pre-k Programs Predict Children's Learning?" *Science 431*, no. 6148: 845–56.

Schweinhart, L. J., H. V. Barnes, and D. P. Weikart. 1993. *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27*. Ypsilanti, MI: High/Scope Press.

Scott-Little, C., S. L. Kagan, and R. M. Clifford, eds. 2003. *Assessing the State of State Assessments: Perspectives on Assessing Young Children*. Tallahassee, FL: SERVE.

Seplocha, H., and J. Strasser. 2008. *A Snapshot of Quality in Abbott Kindergarten Classrooms.* Trenton, NJ: New Jersey Department of Education. http://www.state.nj.us/education/ece/k/snapshot.pdf (accessed April 3, 2015).

Setodji, C. M., V. N. Le, and D. Schaack. 2013. "Using Generalized Additive Modeling to Empirically Identify Thresholds Within the ITERS in Relation to Toddlers' Cognitive Development." *Developmental Psychology 49*, no. 4: 632–45.

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.

Shivers, E., K. Sanders, and T. Westbrook. 2011. "Measuring Culturally Responsive Early Care and Education Programs," in *Quality Measurement in Early Childhood Settings*. Edited by M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle. Baltimore, MD: Brookes.

Shonkoff, J. P., and D. A. Phillips., eds. 2000. *From Neurons to Neighborhoods: The Science of Early Child Development.* Washington, DC: National Academy Press.

Sirinides, P. 2010. *Demonstrating Quality: Pennsylvania Keystone STARS: 2010 Program Report*. Harrisburg, PA: Office of Child Development and Early Learning.

Smith, M. W., and others. 2002. *ELLCO: User's Guide to the Early Language and Literacy Classroom Observation Toolkit.* Research ed. Baltimore, MD: Brookes.

Soukakou, E. P. 2012. "Measuring Quality in Inclusive Preschool Classrooms: Development and Validation of the Inclusive Classroom Profile (ICP)." *Early Childhood Research Quarterly* 27, no. 3: 478–88.

Spiker, D., K. Hebbeler, and L. Barton. 2011. "Measuring Quality of ECE Programs for Children with Disabilities," in *Quality Measurement in Early Childhood Settings*. Edited by M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle. Baltimore, MD: Brookes.

Sylva, K., and others. 1998. *The Early Childhood Environment Rating Scale: 4 Curricular Subscales.* London, UK: Institute of Education, University of London.

Sylva, K., and others. 1999. *The Effective Provision of Pre-school Education [EPPE] Project: An Introduction to the EPPE Project* (Technical Paper 1). London, UK: Institute of Education, University of London.

Talan, T. N., and P. J. Bloom. 2004. *Program Administration Scale: Measuring Leadership and Management in Early Childhood Programs.* New York: Teachers College Press.

Thornburg, K. R., and others. 2009. *The Missouri Quality Rating System School Readiness Study.* Columbia, MO: Center for Family Policy & Research.

Tout, K., and others. 2010a. *Compendium of Quality Rating Systems and Evaluations.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., and others. 2010b. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Year 3 Evaluation Report*. Minneapolis, MN: Child Trends.

Tout, K., and others. 2011. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Final Evaluation Report*. Minneapolis, MN: Child Trends. https://s3.amazonaws.com/Omnera/VerV/s3finder/38/pdf/Parent_Aware_Year_4_Final_Evaluation_Technical_Report_Dec_2011.pdf (accessed March 15, 2013).

Tout, K., and others. 2014. *Implications of QRIS Design for the Distribution of Program Ratings and Linkages Between Ratings and Observed Quality* (OPRE Research Brief 2014-33). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

U.S. Department of Education. 2010. *Toward the Identification of Features of Effective Professional Development for Early Childhood Educators: Literature Review*. Washington, DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

U.S. Department of Education. 2011. *Race to the Top Early Childhood (RTT-ELC) Grant Application*. Washington, DC: U.S. Department of Education. http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/applicant.html (accessed April 3, 2015).

Vandell, D. L., and B. Wolfe. 2000. *Child Care Quality: Does It Matter and Does It Need to Be Improved?* Washington, DC: U.S. Department of Health and Human Services, Office for Planning and Evaluation.

Vu, J. A., H. J. Jeon, and C. Howes. 2008. "Formal Education, Credential, or Both: Early Childhood Program Practices." *Early Education and Development 19*: 479–504.

Weiland, C., and others. 2013. "Associations Between Classroom Quality and Children's Vocabulary and Executive Function Skills in an Urban Public Prekindergarten Program." *Early Childhood Research Quarterly 28*, no. 2: 199–209. http://www.sciencedirect.com/science/article/pii/S0885200612001172 (accessed April 3, 2015).

WestEd Center for Child & Family Studies. 2007. *PITC-PARS Program Assessment Rating Scale: User's Guide for WREL Study*. San Francisco, CA: WestEd Center for Child & Family Studies.

Whitebook, M., C. Howes, and D. A. Phillips. 1989. *Who Cares? Child Care Teachers and the Quality of Care in America*. *Final report of the National Child Care Staffing Study.* Oakland, CA: Child Care Employee Project.

Whitebook, M., and L. Sakai. 2004. *By a Thread: How Child Care Centers Hold on to Teachers, How Teachers Build Lasting Careers.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Wolfe, B., and S. Scrivner. 2004. "Child Care Use and Parental Desire to Switch Care Type Among a Low-income Population." *Journal of Family and Economic Issues 25*: 139–62.

Zaslow, M., and others. 2010. *Quality Dosage, Thresholds, and Features in Early Childhood Settings: A Review of the Literature* (OPRE 2011-5). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Zellman, G. L., and R. Fiene. 2012. *Validation of Quality Rating and Improvement Systems for Early Care and Education and School-age Care* (OPRE 2012-29). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

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. and others. 2008. *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child-Care Quality*. Santa Monica, CA: RAND Corporation. http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG650.pdf (accessed April 3, 2015)

Appendix A. Literature Review

Evaluation Evidence for QRISs

In this section, our goal is to summarize what is known from empirical evaluations of existing QRISs, and to identify what we know from the published literature about effective system design and evidence of system impact. This summary updates a similar literature review provided in AIR and RAND (2013). In this discussion, we do not consider the findings from process or implementation studies of these systems. Here we summarize findings across studies and discuss findings from select studies. For more detailed information on each study, please see the tables in Appendix A1.

We differentiate between two types of evaluation evidence: validation studies and impact studies. The goal of *validation studies* is to determine if the system is well designed and operating in the ways articulated in the system's underlying logic model (whether or not it has been formulated in an explicit way). (See Zellman and Fiene 2012 for further discussion of QRIS validation.) For example, program designers need to know if the system's rating component produces accurate and meaningful program ratings: Does the system for rating program quality measure what it purports to measure? In this case, validation would come from evidence that programs receiving higher quality ratings are indeed providing higher quality care, according to one or more objective measures. Likewise, it is important to know if participating providers are able to increase their quality or their ratings over time, or if child developmental gains are stronger in programs that receive higher quality ratings. Given that many QRISs also include a public awareness campaign, it is also relevant to determine if parents know about and understand the program ratings as a result of the public engagement activities. Thus, as shown in Exhibit A.1, validation studies may be used to examine the relationship between QRIS ratings and observed program quality (V1); to measure whether program ratings or other measures of program quality improve over time (V2); to quantify the relationship between program ratings and child developmental outcomes (V3); or to measure the effectiveness of the public engagement component (V4). Addressing these questions through a validation study is relatively straightforward, as the primary focus is on the programs, teachers, parents, or children in the communities where the system is implemented, and the validation methods require measures for those stakeholders at a point in time or over time.

Exhibit A.1. Illustrative Evaluation Questions for Validation (V) and Impact (I) Studies

Number	Question
V1	Do programs with higher QRIS ratings have higher observed classroom 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?
V4	Do parents know about and understand the QRIS ratings?
I1	Does the implementation of a QRIS change the number or quality mix of providers?
I2	Does the implementation of a QRIS change parental care choice?
I3	Does the implementation of a QRIS improve teacher professional development?
I4	Does the implementation of a QRIS improve teacher performance, other measures of program quality, or program quality ratings?
I5	Does the implementation of a QRIS improve child developmental outcomes?

The aim of *impact studies* is to measure the causal effect of the QRIS on intermediate outcomes such as the provider market, parental behavior, or teacher performance, as well as measure the final outcome of interest, which is child developmental outcomes. Continuing with the evaluation questions shown in Exhibit A.1, an impact study could determine if the QRIS, through the rating component or specific QI activities, results in more high-quality providers in the market place (I1), or in parents being more likely to choose a high-quality provider for their child (I2). If the focus is on teacher outcomes, an impact evaluation might assess whether teachers are more likely to receive professional development such as classroom coaching or a postsecondary degree (I3), or whether teacher performance in the classroom improves (I4). More generally, an impact study could assess the effect of the QRIS as a whole, or specific QI components, on other measures of program quality or QRIS ratings (I4). Typically, the ultimate goal of implementing a QRIS is to improve child developmental outcomes, and this can also be the focus of an impact evaluation (I5). The impact studies required to answer questions I1 to I5 are more challenging to implement, however, because determining the causal effect of the QRIS on any of these outcomes requires measurement of the counterfactual—that is, what these outcomes would have been in the absence of the QRIS. If the QRIS itself can be considered an intervention, the gold standard impact evaluation would require an experimental design, where communities are randomly assigned to implement the QRIS or to continue with the status quo. In a more narrowly focused design, a specific component of the QRIS—such as the inclusion of provider financial incentives or specific types of technical assistance (TA)—could be tested through a randomized assignment of providers to a QRIS design with and without the financial incentive or TA component. In the absence of such experimental designs, other methods that do not include a valid control or comparison condition would be unable to provide evidence of the causal impact of the QRIS design as a whole, or of a QRIS component.

Our review of the literature identified 14 studies covering 12 QRISs in a state or specific areas within states that address one or more of the questions in Exhibit A.1.[24] See Appendix B for further descriptions of these studies. Together, 13 of the studies address one or more of the four

[24] 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 and others, 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 and others, 2011). If distinct validation studies are performed for the same QRIS, we count each study separately (e.g., Barnard and others 2006, Norris and Dunn 2004; Norris, Dunn, and Eckert 2003; and Sirinides 2010). We retain this counting convention throughout.

validation questions listed in Exhibit A.1. Only one study concerns any of the impact questions, and then only questions I3 and I4. There are no studies available to date that have addressed I1, I2, or I5. Below we summarize, in turn, the research findings for studies that address the validation and impact questions.

Evaluations Examining QRIS Ratings and Program Quality

A natural starting point for the validation of a quality rating and improvement system is to ask whether the ratings capture meaningful differences in program quality (the first validation question). We found eleven studies covering nine states that examined this question (Barnard and others 2006; Bryant and others 2001; Elicker and others 2011; Lahti and others 2011; Malone and others 2011; Norris and Dunn 2004; Norris, Dunn, and Eckert 2003; Sirinides 2010; Tout and others 2010b, 2011; Zellman, and others 2008). The evaluations typically focus exclusively on center-based programs, but family child care (FCC) homes are included in some of the validation studies as well. (See Exhibit B-2 in Appendix B for more detailed information on these studies.)

The studies generally use a common design: a program's QRIS rating is compared with an "independent" program quality measure. Ten of the 11 studies compared ratings to an ERS. Eight of the studies included other quality measures in addition to the ERS, such as the Classroom Assessment Scoring System (CLASS); Caregiver Interaction Scale (CIS); or aspects of structural quality, such as teacher education. Of the 10 studies that used an ERS as an outcome measure, all but one found that QRIS ratings were associated with observed quality, although the correlation was not always statistically significant. In many cases, the other measures of program quality—such as the CIS, the CLASS, and teacher education—were also positively correlated with QRIS ratings.

One limitation of this research is that the ERS scale or other measures of program quality (e.g., teacher education) are typically included to assess the validity of QRIS ratings. Thus, in many of these studies, the independent measure of quality against which ratings are compared is not truly independent from the rating process itself. Zellman and others (2008), one of the few studies to use quality measures not incorporated in the QRIS ratings, found that QRIS ratings in Colorado's Qualistar System 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.

Evaluations of Changes in Program Ratings or Quality Indicators

The second validation question in Exhibit A.1 relates to whether program ratings or other indicators of program quality improve over time. We found six studies that examine this issue: four examine changes in global quality as measured by the ERS (Norris, Dunn, and Eckert 2003; Shen, Tackett, and Ma 2009; Sirinides 2010; Zellman and others 2008), while the other two focus on changes in the QRIS ratings (Elicker and others 2011; Tout and others 2011). One study also examines changes in the qualifications of early educators over time (Shen, Tackett, and Ma 2009). All studies focus on providers participating in the QRIS. (See Exhibit A1-3 in Appendix A1 for more detailed information on these studies.)

A consistent finding across the six studies is that quality—as defined, measured, and incentivized in the QRIS—increased over time among participating providers. The study for Indiana (Elicker and others 2011) was the only one to rely on provider self-reports of rating changes, in this case over a short (six-month) period of time. In that evaluation, about one out of five providers had moved up one or more levels, and only a handful dropped a level. Although the studies for Colorado (Zellman and others 2008), Oklahoma (Norris, Dunn, and Eckert 2003), and Pennsylvania (Sirinides 2010) indicate that quality improvements have persisted for up to six years with the QRIS in place, the study by Shen, Tackett, and Ma (2009) for Florida suggests that quality improvements may stall after one to two years. The Florida study did find, however, that the educational attainment and credentials of providers rose over a five-year interval.

It is important to note that these studies are not measuring the *impact* of the QRIS on program ratings. In the absence of a comparison or control group of child care providers that did not participate in the QRIS, the studies cannot conclude that the QRIS as a whole—or specific components of the QRIS, such as the TA activities—produced the observed changes in quality. Another challenge in these studies is the potential attrition over time of providers in the sample. For example, the analysis by Zellman and others (2008) for Colorado is potentially compromised by the fact that lower performing centers were more likely to drop out of the study before the conclusion of data collection, so all reported correlations are based on the remaining higher quality providers.

Evaluations Examining QRIS Ratings and Child Developmental Outcomes

We identified seven studies in seven states that measured the relationship between QRIS ratings and child development outcomes (Elicker and others 2011; Sabol and Pianta 2012, 2014; Shen, Tackett, and Ma 2009; Sirinides 2010; Thornberg and others 2009; Tout and others 2010b, 2011; Zellman and others 2008). With two exceptions, the studies adopted a similar methodology that examined whether changes over time (for example, fall to spring) in an array of child developmental assessments are positively correlated with program QRIS ratings. The studies differ in terms of the care settings included, the child developmental measures deployed and method of collection, the number of time periods in which children were assessed, and the inclusion of controls for family background characteristics. In general, the seven studies provide very limited evidence that QRISs, as currently designed, give higher ratings to programs that generate larger developmental gains. Three of the seven studies found no consistent relationship between QRIS ratings and child outcomes. The four remaining studies found some evidence of a positive relationship between ratings and child outcomes, although two of the four studies have weaker designs. (See Exhibit A1-4 in Appendix A1 for more detailed information on these studies.)

Of the four studies finding associations between ratings and child outcomes, two had stronger research designs: Missouri (Thornburg and others 2009) and Virginia (Sabol and Pianta 2012, 2014). In the Missouri study, a sample of 350 preschool-age children in 38 licensed early childhood programs (32 centers and 6 FCC homes) were assessed in the fall and spring using a battery of well-validated instruments, including the Peabody Picture Vocabulary Test, the Test of Early Reading Ability, the Woodcock-Johnson III Tests of Achievement, and the Devereux Early Childhood Assessment (socio-emotional skills). The range of skills assessed with these and other instruments included vocabulary, early literacy, basic knowledge of shapes and colors,

mathematics skills, fine and gross motor skills, and socio-emotional development. Family background information was also obtained through a parent survey. Overall, the study found that children in higher rated programs, controlling for family background, had significantly higher gains in socio-emotional development compared with children in lower rated programs, but no differences were found for the array of other developmental domains. In examining children in poverty separately, the study found that children in poverty in higher rated programs also benefited in terms of early literacy and physical development, in addition to the socio-emotional gains.

The Virginia validation study relied on teacher-performed assessments of pre-literacy skills for a sample of almost 3,000 children enrolled in state-funded prekindergarten programs and participating the QRIS. The use of teacher reports rather than direct assessments is less preferred, and the study is narrow in focusing only on public prekindergarten programs. One favorable aspect of the study was the rich set of control variables included that were measured at the child, center, and community level. The evaluation showed 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. However, 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.

The two other studies that found positive associations had weaker research designs. The evaluation of Pennsylvania's Keystone STARS (Sirinides 2010) found that 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. However, the study used teacher-reported measures of proficiency in various domains rather than validated developmental assessments implemented by trained, reliable, independent assessors. Moreover, the study did not examine fall–spring changes in child development, but rather reported that participants in higher rated programs were more likely to be proficient at the time of the spring assessment compared with children in the lower rated programs.

In the evaluation of Florida's QRIS in Palm Beach County, Shen Tackett, and Ma (2009) found that readiness was higher on average for children who attended higher quality programs. However, when aggregate school readiness rates were analyzed over time using a comparison group of non-QRIS children, participating children no longer exhibited statistically significant improvement in readiness. Likewise, the evaluation of Florida's QRIS in Palm Beach County relied on a teacher-administered school readiness assessment measured only at kindergarten entry, meaning that gains over time were not measured.

It is important to note that these studies do not provide evidence for or against a causal link between participation in higher rated programs and child developmental outcomes. Without the random assignment of children to programs of varying quality, it is not possible to adequately control for the effect of unobserved factors that may influence both parental selection of programs by quality and child development. Likewise, in the absence of random assignment, these studies do not provide evidence of a causal link between the implementation of a QRIS and child developmental outcomes (question I5 in Exhibit A.1).

Nevertheless, as a validation exercise, the aim of QRIS developers is that the quality ratings denote meaningful distinctions between lower and higher quality programs, with the expectation that programs that receive a higher rating will have a greater impact on children's development compared with lower rated programs. For this reason, the mixed findings across the seven studies reviewed suggest caution about assuming that the rating scales embedded in QRISs will necessarily reflect differences in program quality that relate to child outcomes in the expected way. Only one QRIS appears to have a design that produces program ratings that are positively associated with some domains of child development. At the same time, it is important to recognize that the mixed findings from these studies, given their observational design, may arise from unobserved confounding factors (beyond the family background characteristics included in the models) that affect child development and drive selection into child care programs.

Evaluations Examining Parental Knowledge

The final validation question in Exhibit A.1 asks if parents know about and understand the QRIS ratings. Only two of the evaluation studies we identified addressed this issue (Elicker and others 2011; Tout and others 2010b). The two studies, conducted in Indiana and Minnesota, surveyed parents in QRIS-rated programs or parents in the general public with young children to assess their awareness of the rating system. In Indiana, a higher proportion of parents obtaining child care from a QRIS-rated site had heard about the rating system compared with parents of young children in the general public, as might be expected (Elicker and others 2011). For both groups, when parents had knowledge of the QRIS, their provider was the primary source of information about the rating system. The Indiana study also found that awareness among parents in the general public had increased over a two-year time period. The second study, conducted for Minnesota Parent Aware, focused only on parents in rated programs and also found that awareness of the rating system increased over a one-year interval, although just one out of four parents in rated programs had heard of the rating system by the second year of the survey (Tout and others 2010b). Across the two studies, at best no more than 4 out of 10 parents using a rated provider had knowledge of the QRIS, while just 2 out of 10 parents in the general public knew about the system. (See Exhibit A1-5 in Appendix A1 for more detailed information on these studies.)

A related impact question is whether the implementation of a QRIS changes the choices parents make about the care settings they use (question I2 in Exhibit A.1). No evaluation studies have directly addressed this question to date. It is interesting to note that the Indiana study found that two out of three parents surveyed indicated, in response to a hypothetical question, that a higher rating level would be an "important" or "very important" factor in their choice of child care in the future (Elicker and others 2011). This is suggestive—but by no means conclusive—evidence that the existence of a QRIS may influence parental care choices.

Evaluations of QRIS Impact

Only one study we identified employed an experimental design to answer any of the impact questions listed in Exhibit A.1. Boller and others (2010) focused on the effect of one component of Washington's Seeds to Success QRIS on teacher professional development (I3) and on program quality and quality ratings (I4). In particular, 52 family child care providers and 14 centers that volunteered to participate in the study were randomly assigned into treatment or

control groups. The treatment group received coaching, quality improvement grants, and funds for professional development opportunities and supports, while the control group received funds only for professional development opportunities and supports. Thus, the evaluation measured the incremental impact of including coaching and quality improvement grants in the QRIS. (See Exhibit A1-6 in Appendix A1 for more detailed information on these studies.)

The follow-up period for the Boller and others (2010) study was a relatively short six months, so it is perhaps not surprising that there were no statistically significant impacts of the added coaching and grants on teacher degree attainment for either the home- or center-based programs. However, for teachers in the center-based programs, there was a positive effect on course credits received and lead teacher turnover declined. In addition, the added QRIS components raised participation in an education or training program on the part of center leads and assistant teachers, and significantly more lead teachers in the treatment group than in the control group attended college courses at least weekly. In contrast, FCC providers in the treatment group were no more likely than their control group counterparts to be enrolled in an education or training program.

Boller and others (2010) also examined the effect of the treatment on changes over time in program quality and quality ratings. Interestingly, the study found that the added coaching and professional development significantly improved observed care quality in both home- and center-based settings, but it did not improve the QRIS ratings. The Seeds to Success rating system is based on a block design, suggesting that it may be more challenging for programs to move to higher tiers in a block system, even when some indicators of quality are increasing over time.

Although the study did not employ an experimental design, Shen, Tackett, and Ma (2009) did measure the correlation between provider training and coaching provided in the Palm Beach County QRIS and provider outcomes. The study found that the intensity of coaching (measured as total hours per month) was not associated with improvement in job skills, although skills did improve with the duration of coaching (measured in months). Shen, Tackett, and Ma (2009) also had a comparison group of non-QRIS sites against which they contrasted their QRIS sites in terms of the percentage of "low performing providers" (LPP). They found that QRIS sites showed a significantly higher growth rate in the probability of *not* being rated an LPP over a three-year period. Although these findings are informative, the study design does not provide rigorous causal evidence for any of the impact questions in Exhibit A.1 (that is, question I3 or I4). A more rigorous evaluation design would randomly assign providers to different levels of coaching intensity or duration, or would randomly assign some sites to participate in a QRIS.

This limited evidence base points to the potential for QRIS components that target professional development as part of program improvement to advance teacher participation in education and training, and perhaps eventually educational attainment. There is also some evidence to suggest that program quality may improve as a result of QRIS components that focus on professional development, although depending on the rating system structure, such improvements may not necessarily translate into higher ratings. The one experimental study discussed in this section also demonstrates the potential for using scientifically rigorous methods to evaluate the impact of QRIS components, if not the system as a whole.

The limited impact research to date has not considered the effect of the wider array of quality improvement components contained in most QRISs, such as financial incentives or forms of technical assistance beyond professional development. Particularly notable is the absence of research on the effect of financial incentives, such as improved teacher compensation, on program quality.

Summary of Evaluation Findings

Our review of QRIS evaluation studies produced the following key points regarding validation and impact findings:

- Although QRISs are being designed or implemented in nearly every state, evaluation evidence for QRISs available to date comes from just 12 states or substate areas. The 14 evaluations we identified almost exclusively consist of validation studies that address one or more questions about the effectiveness of the QRIS design. Only one study provides any evidence of QRIS impact, and only for a narrow question.
- Eleven studies examined the relationship between QRIS ratings and a measure of program quality. Ten of the 11 studies used the ERS as an outcome measure. All but one found that the system ratings were positively correlated with observed quality, although the correlation was not always statistically significant. Moreover, the ERS was generally not an independent measure of quality, as it was used to determine the ratings that were being validated.
- Five studies aimed to determine whether program ratings or other program quality measures improve over time. These studies provide consistent evidence, given the way quality is defined, measured, and incentivized in the QRIS, that programs can raise their rating and improve their quality over time.
- Seven studies examined the relationship between QRIS ratings and child developmental outcomes. The findings from these studies are mixed, at best, indicating that there is little evidence to suggest that QRIS ratings, as currently configured, are predictive of child gains for key developmental domains.
- Two studies provide validation evidence about parents' knowledge and understanding of the QRIS ratings. These studies conclude that parents in rated programs know more about the rating system than the general public, and that knowledge of the system tends to increase over time. Even so, the extent of parental awareness of the examined QRISs did not exceed 20 percent for the general public and 40 percent for those using rated providers.
- Although QRIS designers may ultimately be interested in measuring the impact of implementing key elements of a QRIS, or a QRIS as a whole, on a range of system outcomes—provider mix, parental choice, teacher professional development, program quality, or child outcomes—making such causal inferences requires experimental or quasi-experimental designs that have rarely been implemented to date. The one available experimental study demonstrates the potential for using scientifically rigorous methods to extend our understanding of the causal impacts of QRIS implementation.

Conclusions and Implications for California

QRISs constitute an ambitious policy approach to improving early care and education practices and child outcomes. There is strong consensus in the early childhood field that the discussions around QRISs have been effective in increasing awareness of the elements of quality and their importance to practice. The development of standards as part of QRISs has helped providers, parents, and other stakeholders begin to understand and develop agreement around what constitutes quality in ECE. There is also evidence from a number of studies that the combination of standards, ratings, and QI interventions that characterize QRISs improve the average quality of participating programs, at least as defined by the QRIS. However, if we are to improve QRIS implementation, maximize the effects of these systems, and target limited funds to the most promising practices in design, implementation, and quality improvement, we need to approach the design and implementation of these systems armed with far better information about what works than is currently available.

Our review suggests that all states are now engaged in discussions about QRIS design and implementation. This is a positive development because in the process of designing these systems, stakeholders develop consensual standards about quality and increased commitment to its delivery. For the most part, however, the system designers are unable to draw on empirical evidence about the best ways to rate programs, produce summary ratings, or support programs in their efforts to improve the quality of care they provide. Although state policymakers and system designers are endeavoring to learn from their own and other states' earlier QRIS efforts, and are building upon these efforts and using several common components, we do not find that QRIS efforts are yet converging on a preferred design or implementation model at this relatively early stage of their development.

Federal funding requirements have encouraged states to examine the efficacy of QRIS design and implementation practices. Certainly, the early care and education field has begun to actively build an evidence base for QRISs at this stage, and this is a noteworthy development. The research on best practices and evaluation to date primarily focuses on first-generation questions—deciding which elements should go into a well-designed QRIS, and whether design options make sense, target the right elements, and measure what is intended. Yet states are forced to make inferences about best practices in design from the rather limited evidence that is currently available (although an increased focus on validation studies should help to provide additional evidence to assist with these decisions). Furthermore, QI efforts within systems often vary intentionally by design so that they can be responsive to individual program quality improvement needs. Though useful at the program level, this practice makes it difficult to tease out which QI activities are the most effective and should be included in system development. As QRISs mature, studies that look more rigorously at the delivery of TA through quantitative and case study research, will be helpful in designing and delivering these important QI efforts.

The second generation of research should begin to focus on the causal impacts of QRISs, particularly for children, but it may be premature to attempt such studies in the current QRIS environment where change is rapidly occurring. QRISs, like all new systems, will likely need several years of steady state implementation before impact evaluations will be able to meaningfully assess changes in outcomes in a measureable way. Based on research to date, we cannot conclude whether QRISs positively affect child developmental outcomes as intended.

The RTT-ELC grants will require validation and impact studies, and this will provide additional research opportunities in this field. These validation studies, if designed well, will add to the evidence base about preferred design and implementation options. This presents an opportunity to guide the field on empirically based QRIS design and the use of data in decision making. Current QRIS expansion and evaluation also presents an opportunity to measure the impacts of systems more rigorously. However, we caution that evaluations examining the causal impacts of QRISs may not be able to conclude much within the three-year RTT-ELC grant time period. Nevertheless, the continued focus on conducting validation and impact studies to build the QRIS evidence base is a positive trend, and the growing base of evidence will improve these systems over time.

Appendix A1. Summary Tables of Studies Reviewed and Their Findings

Our review of the literature identified 14 studies covering 12 states (or specific areas within states), listed in Exhibit A1-1, that address one or more of the validation or impact questions in Exhibit A.1. (Studies are listed in order by state, with studies covering more than one state listed last.) For each study, we note the geographic coverage, the QRIS name (if applicable), and the question(s) addressed (referencing the numbering system in Exhibit A.1). Eleven of the 14 studies in Exhibit A1-1 address the first validation question by examining the relationship between the QRIS ratings and measures of program quality (V1). Second most common, with seven studies, are validation studies that assess the relationship between quality ratings and child developmental outcomes (V3). Fewer studies examine changes in quality ratings or other quality indicators over time (V2) or parent knowledge (V4)—six studies and two studies, respectively. With one exception, none of the studies provide an impact evaluation as defined in Exhibit A.1.

We note that, with few exceptions, the states listed in Exhibit A1-1 are among the leading states to implement QRISs. They include North Carolina and Oklahoma—two of the earliest adopters (1998)—as well as states that adopted QRISs soon after, between 2000 and 2003 (Colorado, Florida, Indiana, Missouri, Pennsylvania, and Tennessee). These states have had more time to undertake the research required for validation and impact studies, so they are overrepresented among those listed in Exhibit A1-1. Several more recent adopters—Maine, Minnesota, Virginia, and Washington—are also included, as these states integrated evaluation efforts into their early implementation phase or as part of a pilot. It is also worth noting that Exhibit A1-1 does not include any of the research on quality improvement initiatives in California identified in our literature review. None of the California studies to date have addressed the range of evaluation A1-1questions listed in Exhibit A.1.

Exhibit A1-1. Evaluation Questions Addressed by Identified Studies

Study	Geographic Coverage	QRIS Name	Questions Addressed
Zellman and others (2008)	Colorado	Qualistar	V1, V2, V3
Shen, Tackett, and Ma (2009)	Florida (Palm Beach County)	n.a.	V2, V3
Elicker and others (2011)	Indiana	Paths to Quality (PTQ)	V1, V2, V3, V4
Lahti and others (2011)	Maine	Quality for ME	V1
Tout and others (2010b)	Minnesota (Minneapolis, Saint Paul, Wayzata school district, Blue Earth County, and Nicollet County)	Parent Aware	V1, V3, V4
Tout and others (2011)			V1, V2, V3
Thornburg and others (2009)	Missouri (Columbia, Kansas City, and St. Joseph)	n.a.	V3
Bryant and others (2001)	North Carolina	n.a.	V1
Norris, Dunn, and Eckert (2003)	Oklahoma	Reaching for the Stars	V1, V2
Norris and Dunn (2004)	Oklahoma	Reaching for the Stars	V1
Barnard and others (2006)	Pennsylvania	Keystone STARS	V1
Sirinides (2010)	Pennsylvania	Keystone STARS	V1, V2, V3
Sabol and Pianta (2012, 2014)	Virginia	Virginia Star Quality Initiative	V3
Boller and others (2010)	Washington	Seeds to Success	I3, I4
Malone and others (2011)	Florida (Miami-Dade County) and Tennessee	n.a.	V1

Notes: All studies are statewide unless otherwise noted. Question numbers refer to Exhibit A.1.
n.a. = not applicable.

Exhibit A1-2. Evaluations of QRIS Ratings and Program Quality

Study / Location / QRIS	Methods	Key Findings
Zellman and others (2008) / Colorado / Qualistar	Compare QRIS ratings to Caregiver Interaction Scale (CIS) and Pre-Kindergarten Snapshot (Pre-K) subscales	• QRIS ratings were significantly positively related to two of the four CIS subscales (detachment and positive relationship) but not to any of the Pre-K subscales
Elicker and others (2011) / Indiana / Paths to Quality (PTQ)	Compare QRIS ratings to relevant ERS (ITERS-R, ECERS-R, and FCCERS-R) and 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 and others (2011) / Maine / Quality for ME	Compare QRIS ratings to relevant ERS (ITERS-R, ECERS-R, SACERS and FCCERS-R)	• QRIS ratings were significantly positively correlated with ERS
Tout and others (2010b) / Minnesota (see Exhibit A1-1 for sites) / Parent Aware	Compare QRIS ratings to relevant ERS (ITERS-R, ECERS-R, ECERS-E, and	• 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

Study / Location / QRIS	Methods	Key Findings
	FCCERS-R) and CLASS (for center-based programs)	programs at other levels
Tout and others (2011) / Minnesota (see Exhibit A1-1 for sites) / Parent Aware	Compare QRIS ratings to relevant ERS (ITERS-R, ECERS-R, ECERS-E, and FCCERS-R) and CLASS (for center-based programs)	• 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 and others (2001) / North Carolina / n.a.	Compare QRIS ratings to relevant ERS (ECERS-R) and 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
Norris and Dunn (2004) / Oklahoma / Reaching for the Stars	Compare QRIS ratings to relevant ERS (FDCRS) and CIS	• Two-star FCC providers had a higher ERS on average than either 1-star or 1-star plus providers • Two-star FCC providers were more sensitive in their interactions with children than 1-star providers as measured by the CIS • Sample sizes were too small to analyze 3-star (highest category) providers
Norris, Dunn, and Eckert (2003) / Oklahoma / Reaching for the Stars	Compare QRIS ratings to relevant ERS (ECCERS-R, ITERS SACERS) and CIS at two points in time (1999, 2002)	• Two-star center providers had a higher ERS on average than either 1-star or 1-star plus providers
Barnard and others (2006) / Pennsylvania / Keystone-STARS	Compare QRIS ratings to relevant ERS (ECERS-R, FDCRS) and other quality measures (teacher education, curriculum)	• QRIS ratings were positively correlated with ERS (significance not reported) • QRIS ratings for both centers and FCC homes were higher in those sites that used a defined curriculum and where teachers/caregivers had an associate's or higher degree
Sirinides (2010) / Pennsylvania / Keystone STARS	Compare QRIS ratings to relevant ERS (ECERS-R, FDCRS)	• QRIS ratings were not positively correlated with ERS
Malone and others (2011) / Tennessee and Florida (Miami-Dade County) / n.a.	Compare QRIS ratings to relevant ERS (ECERS-R)	• QRIS ratings were positively correlated with ERS

Exhibit A1-3. Evaluations of Program Ratings or Quality Indicators Over Time

Study / Location / QRIS	Methods	Key Findings
Global Quality		
Zellman and others (2008) / Colorado / Qualistar	Measurement of program quality at two points in time for QRIS-rated providers	• Program quality, primarily the ECERS-R, increased over time for providers that were retained in the study
Shen, Tackett, and Ma (2009) / Florida (Palm Beach County) / n.a.	Measurement of program quality up to four points in time for QRIS-rated providers	• ECERS-R scores improved from baseline to 13 months (all subscales) and from 13 to 26 months (4 out of 7 subscales), but not from 26 to 39 months (no subscales) • ITERS-R scores improved from baseline to 13 months (all subscales), but not from 13 to 26 months (no 39-month follow-up)
Elicker and others (2011) / Indiana / Paths to Quality (PTQ)	Provider self-reports of QRIS rating change in past six months	• 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 remained at the same level, and 5% moved or closed.
Tout and others (2011) / Minnesota (see Exhibit B-1 for sites) / Parent Aware	Measurement of program quality at two points in time for QRIS-rated providers	• 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	Measurement of program quality at two points in time for QRIS-rated providers	• ECERS-R scores were significantly higher in 2002 (6.2) than in 1999 (5.6) for the 38 centers visited at both data collection points
Sirinides (2010) / Pennsylvania / Keystone STARS	Measurement of program quality at up to six points in time for QRIS-rated providers	• Data from six years of ERS assessments (ECERS-R, ITERS, SACERS) show that the average quality of assessed sites has been steadily increasing
Other Indicators of Program Quality		
Shen, Tackett, and Ma (2009) / Florida (Palm Beach Co.) / n.a.	Measured qualifications of early educators in QRIS-rated programs at two points in time	• In 2004, 26% of QIS early educators had no high school diploma or GED, compared with 17% in 2009 • The percentage of early educators with a high school diploma or GED, associate's degree, bachelor's degree, and master's degree all increased during this period • The percentage of early educators receiving each of 17 different certificates increased between 2004 and 2009 for all but one of the 17 certificates

Exhibit A1-4. Evaluations of QRIS Ratings and Child Developmental Outcomes

Study / Location / QRIS	Methods	Key Findings
Zellman and others (2008)/ Colorado / Qualistar	Independent assessment of child development at multiple points in time, along with parent survey data, for a sample of preschool-age children enrolled in QRIS-rated centers	• QRIS scores were not associated with improvement in child outcomes • 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
Shen, Tackett, and Ma (2009) / Florida (Palm Beach County) / n.a.	Teacher-administered school readiness assessment conducted at kindergarten entry for children participating in QRIS and non-QRIS preschool sites	• QRIS ratings were found to be positively and significantly associated with the school readiness assessment • Over time, the rate of growth of school readiness rates was higher for QRIS sites, but not significantly so
Elicker and others (2011) / Indiana / Paths to Quality (PTQ)	Independent assessment of child development at one point in time for two age cohorts of children enrolled in QRIS-rated centers and FCC homes	• Infant-toddler developmental assessments were not significantly related to PTQ level, even when controlling for parental education and household income Developmental assessments for preschool-age children were not significantly related to PTQ level, even when controlling for parental education and household income
Tout and others (2010b) / Minnesota (see Exhibit A1-1 for sites) / Parent Aware	Independent assessment of child development in fall and spring, along with parent survey data, for two cohorts (2008-2009 and 2009-2010) of preschool-age children enrolled in QRIS-rated sites	• 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 Parent Aware: Tracking Learning predicted Peabody Picture Vocabulary Test change scores and Teacher Training and Education predicted Woodcock-Johnson quantitative concepts • For some measures, Parent Aware subscale scores negatively predicted child outcomes
Tout and others (2011) / Minnesota (see Exhibit A1-1 for sites) / Parent Aware	Independent assessment of child development in fall and spring, along with parent survey data, for three cohorts (2008–2009, 2009–2010, and 2010–2011) of preschool-age children enrolled in QRIS-rated sites	• Children in programs at different quality rating levels or with different scores on observed quality measures or Parent Aware quality categories 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 and others (2009) / Missouri (see Exhibit A1-1 for sites) / n.a.	Independent assessment of child development in fall and spring (2008–2009), along with parent survey data, for a sample of preschool-age children enrolled in QRIS-rated centers and FCC homes	• Children attending higher rated programs had greater gains in socio-emotional development compared with children in lower rated programs • Children in poverty experienced greater gains in socio-emotional development, early literacy, and physical development in higher rated programs compared with poor children in lower rated programs • Non-poor children in higher rated programs experienced greater gains in socio-emotional development and print awareness/comprehension compared with non-poor children in lower rated programs
Sirinides (2010) / Pennsylvania / Keystone STARS	Teacher reports on child development in fall and spring (2009-2010) for a sample of preschool-age children enrolled in STAR 3 and STAR 4 centers	• 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)

Study / Location / QRIS	Methods	Key Findings
Sabol and Pianta (2012, 2014) / Virginia / Virginia Star Quality Initiative	PreK and K teacher-administered assessment of pre-literacy skills for children participating in QRIS-rated state-funded pre-K programs	• 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 pre-K star rating

Exhibit A1-5. Evaluations of Parental Knowledge

Study / Location / QRIS	Methods	Key Findings
Elicker and others (2011) / Indiana / Paths to Quality (PTQ)	Survey of parents with children in PTQ-rated programs Survey of parents in the general public at two points in time	• 63% of parents reported they had *not* heard about PTQ before being asked to participate in the evaluation study • Of the 37% that had heard about the ratings system, 62% heard about it from the provider • 67% of parents responded that a higher PTQ level would be either an "important" or "very important" factor in their decision in choosing child care in the future • In 2009-2010, 12% of parents surveyed reported that they had heard of PTQ • In 2011, 19% of parents reported that they had heard of PTQ • Among parents who knew about PTQ, their child care provider was the most frequent source of that information
Tout and others (2010b) / Minnesota (see Exhibit A1-1 for sites) / Parent Aware	Survey of parents with children in Parent Aware-rated programs at two points in time	• 20% of surveyed parents reported that they had heard of Parent Aware in the fall of 2008 • 25% of surveyed parents reported that they had heard of Parent Aware in the fall of 2009

Exhibit A1-6. Evaluations of QRIS Impact

Study / Location / QRIS	Methods	Key Findings
Boller and others (2010) / Washington / Seeds for Success	Random assignment of FCC providers and centers to a treatment group that received coaching, quality improvement grants, and funds for professional development opportunities and supports versus a control group that received funds only for professional development opportunities and supports	Impacts on teacher professional development: • For FCC providers, no treatment-control difference in enrollment in an education or training program or in educational attainment • For center lead and assistant teachers, enrollment in an education or training program and in college courses was higher for the treatment group • More center-based teachers in the treatment group than in the control group earned three credits in the past six months, but there was no impact on completion of a postsecondary degree • Lead teacher turnover was lower in the treatment group Impacts on program quality and quality ratings: • For both FCC providers and centers in the treatment group, the ERS total score and most of the ERS subscale scores were significantly higher than control group scores at follow-up • There was no treatment-control difference in Seeds scores

Bibliography for Literature Review

Administration for Children and Families. *Justification of Estimates for Appropriations Committees,* n.d. http://www.acf.hhs.gov/sites/default/files/assets/CCDF%20final.pdf (accessed March 15, 2013).

American Institutes for Research. *Evaluation of Preschool for All (PFA) Implementation in San Mateo and San Francisco Counties: Year 2 Report*. San Mateo, CA: American Institutes for Research, 2007.

American Institutes for Research. *Evaluation of Preschool for All (PFA) Implementation in San Francisco County: Year 5 Report.* San Mateo, CA: American Institutes for Research, 2010.

Applied Survey Research and First 5 Santa Cruz County. *Annual Evaluation Report: July 1, 2011–June 30, 2012*. Santa Cruz, CA: First 5 Santa Cruz County, 2012.

Austin, L., and P. Scroggins. *Pilot of Tracking & Reporting Training Participants & Training Activities.* Sacramento, CA: CDE/CDD Quality Improvement Office, Quality Improvement Professional Development, 2011–12.

Avila de Lima, J. "Thinking More Deeply About Networks in Education." *Journal of Educational Change* 11, no. 2 (2010): 1–21.

Barnard, W., and others. *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, 2006.

Bernzweig, J. *Quality Counts: Consultation to Family Child Care*. Alameda, CA: First 5 Alameda County, 2011.

Boller, K., and others. *The Seeds of Success Modified Field Test: Findings from the Impact and Implementation Studies*. Princeton, NJ: Mathematica Policy Research, Inc., 2010.

Bowman, B. T., and others, eds. *Eager to Learn: Educating Our Preschoolers*. Washington, DC: National Academy Press, 2000.

Bowman, B.T., Donovan, M.S., and Burns, M.S. eds. *Eager to Learn: Educating Our Preschoolers*. Washington, D.C.: National Academy Press, 2001.

Brandon, R. *Financing to Promote Quality in Early Care and Education and School-Age Care: Incentives, Supports and Affordability*. Research-to-Policy, Research-to-Practice Brief. Washington, DC: Office of Planning, Research and Evaluation, Administration for

Children and Families, U.S. Department of Health and Human Services, 2011 www.acf.hhs.gov/programs/opre (accessed March 15, 2013).

Bromer, J., M. ,and others. *Staffed Support Networks and Quality in Family Child Care: Findings From the Family Child Care Network Impact Study*. Chicago, IL: Local Initiatives Support Corporation, 2009.

Bryant, D. M., and others. *Validating North Carolina's 5-Star Child Care Licensing System.* Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Center, 2001. http://fpg.unc.edu/resources/validating-north-carolinas-5-star-child-care-licensing-system (accessed March 15, 2013).

Bryant, D., and others. *The QUINCE-PFI Study: An Evaluation of a Promising Model and Delivery Approaches for Child Care Provider Training*. Draft final report. Chapel Hill, NC: University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Institute, 2009.

Burchinal, M., and others. "Quality of Center Child Care and Infant Cognitive and Language Development." *Child Development* 67 (1996): 606–620.

CAEL QIS. *Dream Big for Our Youngest Children,* 2010. http://www.cde.ca.gov/sp/cd/re/documents/fnlrpt2010.pdf (accessed March 15, 2013).

Cassidy, D. J., and others. "The Effect of Education on Child Care Teachers' Beliefs and Classroom Quality: Year One Evaluation of the TEACH Early Childhood Associate Degree Scholarship Program." *Early Childhood Research Quarterly* 10, no. 2 (1995): 171–183.

Cassidy, D .J., and others. "Measurement of Quality in Preschool Child Care Classrooms: An Exploratory and Confirmatory Factor Analysis of the Early Childhood Environment Rating Scale-Revised." *Early Childhood Research Quarterly* 20, no.3 (2005): 345–360.

Center on the Developing Child, National Forum on Early Childhood Program Evaluation, and National Scientific Council on the Developing Child. *A Science-Based Framework for Early Childhood Policy: Using Evidence to Improve Outcomes in Learning, Behavior, and Health for Vulnerable Children,* 2007.

Child Care Aware of America. *Parents and the High Cost of Child Care,* 2012. http://www.naccrra.org/sites/default/files/default_site_pages/2012/cost_report_2012_final_081012_0.pdf (accessed June 4, 2013).

Clarke-Stewart, K., and others. "Do Regulable Features of Child-Care Homes Affect Children's Development?" *Early Childhood Research Quarterly* 17, no. 1 (2002): 52–86.

Constantine, W., D. S. Gomby, and B. Mitchell. *A Review of the First 5 Contra Costa Services for: Mental Health Consultation, Inclusion Facilitation, and Parents and Caregivers of Children with Special Needs*. Concord, CA: First 5 Contra Costa, 2008.

Davis Consultant Network. *Evaluation Findings for Fiscal Year 2011–2012*. Woodland, CA: First 5 Yolo, 2012.

Dickinson, D. K., and L. Caswell. "Building Support for Language and Early Literacy in Preschool Classrooms Through In-service Professional Development: Effects of the Literacy Environment Enrichment Program (LEEP)." *Early Childhood Research Quarterly* 22 no. 2 (2007): 243–260.

Donegan, M. M., M. M. Ostrosky, and others. "Peer Coaching: Teachers Supporting Teachers." *Young Exceptional Children* 3, no. 3 (2000): 9.

Elicker, J., and others. *Evaluation of Paths to QUALITY, Indiana's Child Care Quality Rating and Improvement System: Final Report*. West Lafayette, IN: Purdue University, 2011.

Felix, E., A., and others. *Evaluation Report 2010–2011: A report to the Santa Barbara County First 5 Commission*. Santa Barbara, CA: University of California–Santa Barbara, 2011.

Fiene, R. "Improving Child Care Quality Through an Infant Caregiver Mentoring Project." *Child & Youth Care Forum* 31, no. 2 (2002): 79–87.

First 5 Fresno. *Fresno QRIS Pilot Project*. Unpublished report, 2011.

First 5 Los Angeles. *Universal Preschool Child Outcomes Study (UPCOS),* 2012.

First 5 San Francisco. *A Report to the Community on Strategic Plan Progress 2010–11*. San Francisco: Author, 2012.

Franke, T., L. Espinosa, and L. Hanzlicek. *Power of Preschool Program Evaluation Report December 2011*. Sacramento, CA: First 5 California, 2011.

Fuller, B., and S. L. Kagan. *Remember the Children: Mothers Balance Work and Child Care Under Welfare Reform: Growing Up in Poverty Project 2000, Wave 1 Findings—California, Connecticut, Florida*. Berkeley, CA: Graduate School of Education—Policy Analysis for California Education, University of California, 2000.

Girolametto, L., and others. "The Effects of In-Service Education to Promote Emergent Literacy in Child Care Centers: A Feasibility Study." *Language, Speech, and Hearing Services in Schools* 38 (2007):72–83.

Hamre, B. K., S. G. Goffin, and M. Kraft-Sayre. *Classroom assessment scoring system (CLASS) implementation guide,* 2009. http://www.teachstone.org/about-the-class/ (accessed March 15, 2013).

Harder+Company. *2006–2007 First 5 San Joaquin Preschool for All Evaluation Report*. Stockton, CA: First 5 San Joaquin, 2007.

Harder+Company. *CARES Statewide Retention Study: Final Report.* Sacramento, CA: First 5 California, 2008.

Harder+Company. *CARES for the Early Learning Workforce: Evaluation Report, Round Four, 2006–06 to 2007–08*. Sacramento, CA: First 5 California, 2009.

Harder+Company. *First 5 Contra Costa Annual Evaluation Report: 2008–2009*. Concord, CA: First 5 Contra Costa, 2010a.

Harder+Company. *First 5 San Joaquin Evaluation Report 2008–2010,* Stockton, CA: First 5 San Joaquin, 2010b.

Harder+Company. *First 5 San Joaquin School Readiness Longitudinal Evaluation Study*, *Year 3 Progress Report*. Stockton, CA: First 5 San Joaquin, 2010c.

Harder+Company. *First 5 San Diego Annual Evaluation Report FY 10–11* San Diego, CA: First 5 San Diego, 2012.

Harder+Company. *First 5 San Joaquin Longitudinal School Readiness Evaluation Report*. Stockton, CA: First 5 San Joaquin, 2013.

Helburn, S. W., ed. *Cost, Quality, and Child Outcomes in Child Care Centers: Technical Report*. Denver, CO: Department of Economics, Center for Research in Economic and Social Policy, University of Colorado at Denver, 1995.

Helburn, S. W., ed. *Cost, Quality, and Child Outcomes in Child Care Centers: Technical Report*. Denver, CO: Department of Economics, Center for Research in Economic and Social Policy, University of Colorado at Denver, 1995.

Howes, C. "Relations Between Early Child Care and Schooling." *Developmental Psychology* 24 (1988): 53–57.

Isner, T., and others. *Coaching in Early Care and Education Programs and Quality Rating and Improvement Systems (QRIS): Identifying Promising Features*. Washington, DC: Child Trends, 2011.

jdcPartnerships. *Corps AA 2006–2010: Summary of Findings and Recommendations.* San Rafael, CA: jdcPartnerships, 2010.

Karoly, L. A. *Preschool Adequacy and Efficiency in California: Issues, Policy Options, and Recommendations*. MG-889. Santa Monica, CA: RAND Corporation, 2009.

Karoly, L. A. *A Golden Opportunity: Advancing California's Early Care and Professional Development System*. MG-1188. Santa Monica, CA: RAND Corporation, 2012.

Karoly, L. A., and others. *Prepared to Learn: The Nature and Quality of Early Care and Education for Preschool-Age Children in California.* TR-539. Santa Monica, CA: RAND Corporation, 2008.

Keys, T. D., and others. "Preschool Center Quality and School Readiness: Quality Effects and Variation by Demographic and Child Characteristics." *Child Development* 84, no. 4 (2013): 1171–1190.

Kipnis, F., M. Whitebook, and others. *Learning Together: A Study of Six B.A. Completion Cohort Programs in Early Care and Education: Year 4*. Berkeley, CA: Center for the Study of Child Care Employment, 2012.

Koh, S., and S. B. Neuman. "The Impact of Professional Development in Family Child Care: A Practice-Based Approach." *Early Education & Development* 20, no.3 (2009): 537–562.

Kohler, F. W., K. M. McCullough, and others. "Using Peer Coaching to Enhance Preschool Teachers' Development and Refinement of Classroom Activities." Early Education & Development 6, no.3 (1995): 215–239.

Kretlow, A. and others. "Using In-Service and Coaching to Increase Kindergarten Teachers' Accurate Delivery of Group Instructional Units." *The Journal of Special Education* 44, no. 4 (2009): 234–436.

Lahti, M., and others. *Validation of Quality Rating and Improvement Systems (QRIS): Examples from Four States* (OPRE 2013-036). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2013. http://www.researchconnections.org/childcare/resources/26590/pdf (Accessed April 3, 2015).

Lam, I., and S. Muenchow. *First 5 Power of Preschool: Lessons from an Experiment in Tiered Reimbursement*. San Mateo, CA: American Institutes for Research, 2009.

Landry, S. H., and others. "Effectiveness of Comprehensive Professional Development for Teachers of At-Risk Preschoolers." *Journal of Educational Psychology* 102, no. 2 (2009): 448–465.

Lipsey, M. W., and D. B. Wilson. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications, 2001.

Love, J., S. and others. *Los Angeles Universal Preschool Programs, Children Served, and Children's Progress in the Preschool Year: Final Report of the First 5 LA Universal Preschool Child Outcomes Study*. Princeton, NJ: Mathematica Policy Research, 2009.

Malone, L., and others. *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, 2011. http://qrisnetwork.org/sites/all/files/resources/gscobb/2011-09-28%2014:15/Report.pdf (accessed March 15, 2013).

Mitchell, A. *Financial Incentives in Quality Rating and Improvement Systems: Approaches and Effects*. QRIS National Learning Network, 2012. http://www.qrisnetwork.org/sites/all/files/resources/gscobb/2012-05-24%2015:13/Approaches%20to%20Financial%20Incentives%20in%20QRIS.pdf (accessed March 15, 2013).

Mitchell, A., and others. *Comparison of Financial Incentives in States' Quality Rating and Improvement Systems*. Albany, NY: Early Childhood Policy Research, Alliance for Early Education Finance, 2008. http://www.earlychildhoodfinance.org/downloads/2008/MitchQRISfinIncentives_2008.pdf (accessed March 15, 2013).

Moiduddin, E., Y. Xue, and S. Atkins-Burnett. *Informing the Performance-Based Contract Between First 5 LA and LAUP: Child Progress in the 2010–2011 Program Year*. Princeton, NJ: Mathematica Policy Research, 2011.

National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network (ECCRN). "The Relation of Child Care to Cognitive and Language Development." *Child Development* 74, no. 4 (2000).

Norris, D. J., and L. Dunn. *Reaching for the Stars: Family Child Care Home Validation Study Final Report*. Stillwater and Norman, OK: Early Childhood Collaborative of Oklahoma, 2004.

Norris, D. J., L. Dunn, and L. Eckert. *Reaching for the Stars: Center Validation Study Final Report*. Stillwater and Norman, OK: Early Childhood Collaborative of Oklahoma, 2003.

Neuman, S. B., and L. Cunningham. "The Impact of Professional Development and Coaching on Early Language and Literacy Instructional Practices." *American Educational Research Journal* 46, no. 2 (2009): 532–566.

Parsons, G., and S. LaFrance. *Evaluation of Every Child Counts First 5 Alameda Children and Families Commission's Every Director Counts Project*. San Francisco: LaFrance Associates, 2006.

Peisner-Feinberg, E., and others. "The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Development Trajectories through Second Grade." *Child Development* 72, no. 5 (2001): 1534–1553.

Pianta, R. C., and others. "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre- Kindergarten Classrooms." *Early Childhood Research Quarterly* 23. no. 4 (2008): 431–451.

Prayaga, R. B. *Power of Preschool Program Evaluation Report*. Sacramento, CA: First 5 California, 2009.

Ramey, S., and C. Ramey. *The Right from Birth study: An Evidence-informed Training Model to Improve the Quality of Early Child Care and Education*. Washington, DC: Presentation at the Child Care Policy Research Consortium Meeting, 2008.

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

Sabol, T., and R. Pianta. "Validating Virginia's Quality Rating and Improvement System Among State-Funded Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 30 (2014): 183–198. http://qrisnetwork.org/sites/all/files/materials/Validating%20Virginia%E2%80%99s%20quality%20rating%20and%20improvement%20system%20among%20state-funded%20pre-kindergarten%20programs%20(Early%20Childhood%20Research%20Quarterly).pdf (accessed April 3, 2015)

Sanchez, M. *Educational Outcomes for Preschool for All Participants in Redwood City School District—Update*. Stanford, CA: John W. Gardner Center for Youth and Their Communities, 2012.

Shen, J., W. Tackett, and X. Ma. *Second Evaluation Report for Palm Beach County Quality Improvement System.* Palm Beach, CA: Children's Services Council of Palm Beach County, 2009.

Shonkoff, J. P. and Phillips, D. A., eds. *From Neurons to Neighborhoods: The Science of Early Child Development*. Washington, D.C.: National Academy Press, 2000.

Sirinides, P. *Demonstrating Quality: Pennsylvania Keystone STARS: 2010 Program Report*. Harrisburg, PA: Office of Child Development and Early Learning, 2010.

Thornburg, K. R., and others. *The Missouri Quality Rating System School Readiness Study.* Columbia, MO: Center for Family Policy & Research, 2009.

Tout, K., and others. *Compendium of Quality Rating Systems and Evaluations.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2010a.

Tout, K., and others. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Year 3 Evaluation Repor*t. Minneapolis, MN: Child Trends, 2010b. https://s3.amazonaws.com/Omnera/VerV/s3finder/38/pdf/Parent_Aware_Year_3_Evaluation_Report_Nov_2010.pdf (accessed March 15, 2013).

Tout, K., and others. *Evaluation of Parent Aware: Minnesota's Quality Rating System Pilot: Final Evaluation Report*. Minneapolis, MN: Child Trends, 2011. https://s3.amazonaws.com/Omnera/VerV/s3finder/38/pdf/Parent_Aware_Year_4_Final_Evaluation_Technical_Report_Dec_2011.pdf (accessed March 15, 2013).

Tschanz, J. M. and C. O. Vail. "Effects of Peer Coaching on the Rate of Responsive Teacher Statements During a Child-directed Period in an Inclusive Preschool Setting." *Teacher Education and Special Education* 23, no. 3 (2000): 189–201.

Valcasti, M. R., and others. *First 5 Merced County Annual Evaluation Report FY 2010–2011.* Merced, CA: First 5 Merced County, 2011.

Vandell, D. L., and B. Wolfe. *Child Care Quality: Does It Matter and Does It Need to be Improved?* Report prepared for the U.S. Department of Health and Human Services, Office for Planning and Evaluation, 2000.

WestEd E3 Institute. *5 Years of Learning: A Report on the First Five Years of Santa Clara CARES 2002–2006*. San Jose, CA: WestEd, 2007.

WestEd E3 Institute. *Power of Preschool in Santa Clara County: Making the First Five Years Count*. San Jose, CA: WestEd, 2011.

The White House. *Fact Sheet: President Obama's Plan for Early Education for all Americans,* 2013. http://www.whitehouse.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans

Whitebook, M., and others. *Learning Together: A Study of Six B.A. Completion Cohort Programs in Early Care and Education: Year 3*. Berkeley, CA: Center for the Study of Child Care Employment, 2011.

Whitebook, M., and others. *Two Years in Early Care and Education: A Community Portrait of Quality and Workforce Stability*. Berkeley, CA: Center for the Study of Child Care Employment, 2004.

Whitebook, and others. *Learning Together: A Study of Six B.A. Completion Cohort Programs in Early Care and Education. Year 1 Report.* Berkeley, CA: Center for the Study of Child Care Employment, University of California at Berkeley, 2008.

Xue, Y, S. Atkins-Burnett, and E. Moiduddin. *Informing the Performance-Based Contract Between First 5 LA and LAUP: Child Progress in the 2011–2012 Program Year.* Princeton, NJ: Mathematica Policy Research, 2012.

Yoon, K. S., and others. *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, 2007.

Zaslow, M., and others. *Toward the Identification of Features of Effective Professional Development for Early Childhood Educators: Literature Review*. Washington, DC: Child Trends, 2010.

Zellman, G. L., and R. Fiene. *Validation of Quality Rating and Improvement Systems for Early Care and Education and School-age Care,* Research-to-Practice Brief, OPRE. Washington, DC: Office of Planning Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2012.

Zellman, G. L., and A. S. Johansen. *Examining the Implementation and Outcomes of the Military Child Care Act of 1989.* MR-665-OSD. Santa Monica, CA: RAND Corporation, 1988.

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

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

Zellman, G. L., and others. *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child-Care Quality*. Santa Monica, CA: RAND Corporation, 2008. http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG650.pdf (accessed March 15, 2013)

Zwart, R.C., and others. "Teacher Learning Through Reciprocal Peer Coaching: An Analysis of Activity Sequences." *Teaching and Teacher Education* 24, no. 4 (2008): 982–1002.

Appendix B. Validation Study Methods

This appendix describes in detail the methods—the study samples, the data sources, and the procedures—used for the validation study.

Study Samples

Sampling of programs for the study drew from a comprehensive list of all programs participating in California's QRIS in January 2014 (N = 1,272), which was compiled from separate lists of programs enrolled in each of the 17 Consortia's locally administered QRIS. Three separate samples were used for the study analyses, including a sample of all programs with full QRIS ratings as of January 2014 (N = 472), a subsample of the programs with full ratings that received classroom observations for the concurrent validity analyses (N = 175), and a subsample of centers with classroom observations in 100 percent of their classrooms (N = 26). Descriptions of these samples follow.

Programs With Full QRIS Ratings

Study analyses that only require existing data collected for QRIS ratings use the sample of programs with full QRIS ratings as of January 2014. Programs with full ratings were identified by each Consortium and include those programs with complete rating data on all required elements. Data files received from the Consortia were initially reviewed for completeness, and some back and forth with Consortia was necessary to ensure that data were completed or corrected where needed. Of the 1,272 programs participating in the QRIS, 472 programs (365 centers and 107 FCCHs) in 12 of the 17 Consortia had full ratings. The relatively low percentage of participating programs with full ratings (37 percent) reflects the early stage of implementation of California's RTT-ELC QRIS during the study period. Of the remaining 800 participating programs without full ratings, 552 had provisional ratings based on incomplete or estimated element scores and 248 did not yet have any assigned rating. Programs with provisional ratings could not be included in the study analyses because the provisional rating data are not reliable, and they are not comparable to the rating data for programs with full ratings.

The sample of programs with full ratings in January 2014, early in QRIS implementation, is not representative of all programs participating in California's QRIS. The programs with and without full ratings differ significantly according some program characteristics, as shown in Exhibit B.1. Programs with full ratings are more likely than programs without full ratings to have standards-based public funding requiring programs to meet specific quality standards for State Preschool, the Child Signature Program (CSP), or Head Start, although the prevalence is quite high among both categories of programs participating in the QRIS[25]. Because programs without full ratings are less likely to receive this type of standards-based funding (particularly CSP), there may be greater diversity in the quality ratings of these programs when they are finalized, in comparison to programs that already have full ratings. Fully rated programs are also more likely to receive child care subsidy vouchers as well as private pay. Fully rated programs are less likely to use a

[25] In the early phases of RTT-ELC implementation, California prioritized enrollment of programs receiving public funding in the QRIS, in response to RTT-ELC guidelines on the inclusion of programs serving high-needs children.

language other than English during the program day with children compared with non-fully rated programs. There are also differences in terms of the distribution of programs across Consortia, in part because five Consortia have no fully rated programs. In addition, the programs with full ratings are concentrated in four Consortia located in three large counties that had existing QRISs in place prior to RTT-ELC (Los Angeles, San Diego, and San Francisco). In contrast, programs without full ratings are spread more evenly across the 17 Consortia. There are no significant differences in the percentage of programs that are centers and FCCHs, or in the average total enrollment of either program type. Still, the differences in program characteristics indicate limited generalizability of the validation study results presented in this report.

Exhibit B.1. Characteristics of Programs Participating in California QRIS, With and Without Full Ratings

	Programs With Full Ratings (*N* = 472)		Programs Without Full Ratings (*N* = 800)		
	N	**Percentage**	***N***	**Percentage**	***p***
Program Type	472		800		.660
Center-Based	365	77%	610	76%	
FCCH	107	23%	190	24%	
Funding Sources (Programs May Have Multiple Sources)	452		733		
Standards-Based Public Funding (CSP, Title 5, or Head Start)	382	85%	504	67%	< .0001
First 5 California CSP 1 or CSP 2 funding	222	49%	62	8%	< .0001
California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	249	55%	381	52%	.297
Federal Head Start or Early Head Start funding	149	33%	157	21%	< .0001
State/Federally Funded Child Care Subsidy Vouchers	169	37%	123	17%	< .0001
Private Pay	192	42%	258	35%	.012
Language Spoken With Children	445		674		
Non-English Language Spoken With Children	256	58%	506	75%	< .0001
Spanish Spoken With Children	249	56%	501	74%	< .0001
Consortia	472		800		< .0001
Alameda	17	4%	0	0%	
Contra Costa	8	2%	54	7%	
El Dorado	0	0%	32	4%	
Fresno	5	1%	45	6%	
LA OCC	52	11%	126	16%	
LAUP	97	21%	44	6%	
Merced	0	0%	48	6%	

	Programs With Full Ratings (*N* = 472)		Programs Without Full Ratings (*N* = 800)		
	N	**Percentage**	***N***	**Percentage**	***p***
Orange	8	2%	60	8%	
Sacramento	27	6%	106	13%	
San Diego	89	19%	12	2%	
San Francisco	102	22%	9	1%	
San Joaquin	13	3%	60	8%	
Santa Barbara	0	0%	97	12%	
Santa Clara	13	3%	6	1%	
Santa Cruz	0	0%	40	5%	
Ventura	41	9%	34	4%	
Yolo	0	0%	27	3%	
Full or Provisional QRIS Rating With Local Adaptations	472		552		<.0001
Tier 1	4	1%	103	19%	
Tier 2	85	18%	162	29%	
Tier 3	155	33%	179	32%	
Tier 4	196	42%	104	19%	
Tier 5	32	7%	4	1%	
	N	**Mean (SD)**	***N***	**Mean (SD)**	***p***
Total Enrollment					
Average Total Enrollment, Centers	362	52.9 (32.3)	462	50.6 (42.2)	.396
Average Total Enrollment, FCCHs	107	9.1 (4.1)	161	8.3 (4.5)	.161

Source: Common Data Elements 2014.

Notes: *p* values are based on χ^2 tests for all comparisons except average total enrollment, which is based on a *t* test. QRIS ratings presented in this table are those reported by Consortia using local adaptations, and the distribution shown in this table differs from the distribution of ratings without local adaptations that were calculated for study analyses. The QRIS ratings of programs with full ratings are not directly comparable to the provisional ratings available for programs without full ratings.

In addition, the distribution of QRIS ratings with local adaptations differs significantly across the two groups, although this comparison should be interpreted with caution because the ratings are by definition not comparable. The programs without full ratings have been assigned provisional ratings that do not use complete or verified data. The provisional ratings assigned to programs without full ratings skews lower than the full ratings assigned to programs in the fully rated group. Programs with provisional ratings are far more likely to receive low ratings of 1 or 2. This difference could suggest lower quality among programs without full ratings, but other explanations are possible. For example, programs may wait to finalize their ratings until they have met requirements for the next rating level above their provisional rating. Also, some of the programs with provisional ratings were in that category because they were waiting to receive the independent CLASS or ERS observations and had provisional ratings of 2 points on these

elements while waiting for the observations to be completed (these programs will receive a minimum of 3 points and as many as 5 points on each element once the observations are completed, depending on the observation score, and thus may earn enough additional points for a higher QRIS rating).

Concurrent Validity Sample

The concurrent validity sample includes all programs with complete CLASS or PQA data for classrooms selected for the concurrent validity analyses and represents a subsample of programs with full ratings. The number of fully rated programs is smaller than the original planned sample size for the study, which was based on rating projections from 2013 estimating that more than 1,000 programs would have full ratings by 2014. With fewer than expected fully rated programs, we did not draw a random sample of programs for the concurrent validity sample. Instead, all fully rated programs were invited to participate in the study. Recruitment procedures and the resultant sample are described in this section.

Study recruitment

We invited all programs with full ratings ($N = 472$) to participate in the study. In order to recruit programs to participate in the study, we first worked with the Consortia to help develop buy-in for the study. We began by holding a webinar for the Consortia where we provided an overview of the study and addressed questions and concerns expressed by the Consortia. We then asked the Consortia to send an introductory e-mail to all sites. Our study team followed up with an additional e-mail addressed to the sites and then began calling sites to invite them to participate in the study. We e-mailed and called all eligible sites and followed up with additional phone calls and e-mails as needed to gain participation agreements with the sites. As a part of this process, we also provided a website where sites could review a webinar overview of the study, available in both English and Spanish. In addition, we provided written materials to answer sites' questions and address potential concerns about the study.

Once sites agreed to participate, we collected basic information on their classrooms (e.g., number and ages of children). This information was used to sample classrooms for observations according to the implementation guide procedures. For a subsample of sites, the "100 percent subsample," all classrooms were selected for observation in order to test different classroom sampling approaches and to compare resultant ratings.

Challenges associated with recruitment and gaining sites' participation agreements

Two main challenges hampered our ability to recruit the number of sites that we estimated were necessary for our planned analyses: (1) fewer sites than anticipated had full ratings, and (2) the short timeline, compounded by delays caused by concerns about the study expressed by the Consortia, made recruiting sites and completing data collection prior to the end of the program year a further challenge.

First, based on information gathered from the *Descriptive Study* (AIR and RAND 2013), we anticipated that there would be more than 1,000 rated sites from which to draw a sample for inclusion in the study. We learned at the initial meeting with the implementation team that not all

sites had "full" ratings. That is, some sites had not had the opportunity to receive their CLASS or ERS observation and were assigned a temporary "provisional" rating. Once we collected the data from all of the Consortia and removed the provisionally rated sites, we found that there were only 472 sites with full ratings. This meant that it would not be possible to achieve the sample size originally planned.

Second, after the initial webinar for the Consortia, several Consortia expressed concerns about the design of the study. Questions were raised about the feasibility of conducting the study and the appropriateness of evaluating so early in the implementation of the RTT-ELC QRIS. In addition, several Consortia raised concerns about the burden placed on sites by the study. In particular, the Consortia were concerned about the number of classroom observations that the study would be conducting on top of the multiple observations that sites were already receiving.

Several Consortia preferred that we wait until their concerns had been addressed before we invited sites in their counties to participate in the study. We worked with the CDE to develop a plan for accepting some extant data in lieu of conducting additional classroom observations in sites that had recent Consortia-conducted observations. This reduced the burden on these sites. Unfortunately, this process also caused a significant delay in conducting the initial recruitment, which, in turn, inhibited our ability to get sites on board in time to collect all the data (i.e., the program or school year ended before we were able to collect data in some programs). Response rates also varied by Consortia and suggest that the fact that several Consortia were apprehensive about the study may have filtered down to the sites and reduced buy-in for the study, limiting our ability to collect participation agreements.

Extant CLASS and ERS data obtained from Consortia

To reduce burden on sites, we accepted extant CLASS and ERS data from Consortia to supplement our independently conducted CLASS and ERS observations for the study. To maintain consistency with the primary data collected for the study, some restrictions were applied to the inclusion of data from the Consortia.

1. CLASS and ERS score data had to be collected recently—in August of 2013 or later.
2. The CLASS and ERS data had to be collected using the instruments as published without any local adaptation of the measures.
3. Only data on classrooms sampled by AIR were used.
4. Consortia had to be able to provide AIR with raw data for every item on the observation measure in addition to the domain scores, overall score, and the date of the observation for each observed classroom in the site.
5. The data provided to AIR had to be complete, with plausible values provided for each variable needed.

Six Consortia provided data that met these criteria and could be included in the analyses.

Sample size and response rates

From the 472 sites with full ratings, we determined that 50 sites were ineligible because of one or more of the following reasons:

- The site had closed or was planning to close prior to the spring data collection ($N = 12$).
- The site used a language other than English or Spanish in the classroom (and therefore classroom observations could not be conducted by our Spanish/English bilingual observers) ($N = 15$).
- The site was part of a 12th Consortium that provided full ratings in time to be included in the reliability and sensitivity analyses but not in time to recruit their programs to participate in the concurrent validity analyses ($N = 8$).
- The site had recent staffing disruptions that meant observations could not occur ($N = 4$).
- There were no age eligible children for the child assessments ($N = 2$).
- Other reasons ($N = 9$).

In addition, 16 sites were unresponsive to our communications and never provided a final response to our invitation to participate.

We were able to secure participation agreements from 195 sites out of the total of 422 eligible sites, or 46 percent. However, because of the recruitment delays, we were unable to schedule and conduct data collection at eight of these sites. Observation data were obtained on at least some classrooms (either through independent observations conducted by study field staff or through extant data provided by the Consortia) from 187 sites (44 percent of eligible sites). Complete data on the CLASS or PQA were obtained for 175 sites, which compose the concurrent validity sample. Exhibit B.2 provides an overview of the sample size and response rates.

Exhibit B.2. Sample Size and Response Rates

	N	Percentage
Number of sites with full ratings	472	
Ineligible sites	50	11%
Number of eligible sites	422	
Number of sites that agreed to participate	195	46%
Number of sites with some observation data	187	44%
Number of sites with complete CLASS or PQA data	175	41%

As anticipated, given the time required for participation and the multiple demands on staff's time, many sites declined to participate in the study. Given an overall site level participation rate of less than 50 percent, it is important to examine the extent to which the characteristics of the participating sites differ from the characteristics of nonparticipating sites with full ratings. Because our analyses focus on the 175 sites with complete data, we compare these to all fully rated sites that did not have complete data ($n = 297$).

Exhibit B.3 presents details on the comparison of characteristics of sites with and without complete data. The rate of study participation was significantly higher among centers than FCCHs, and the total number of FCCH participants that received observations for the concurrent validity study was too small to support concurrent validity analyses (FCCHs cannot be included with centers because the ratings are calculated differently for centers and FCCHs); therefore, the concurrent validity results do not apply to FCCHs. The prevalence of standards-based public funding was similar among programs that did and did not participate in the concurrent validity study, although there were differences in specific funding types and in acceptance of vouchers and private pay. A much higher percentage of participating programs used languages other than English as well as Spanish specifically with children. Enrollment was slightly larger, on average, in participating centers, but there were no significant differences in FCCH enrollment. There were no significant differences between groups in the distribution of QRIS ratings with local adaptations. There were significant differences in participation by Consortia, with participation rates ranging from 11 percent of programs to 100 percent of programs.

Exhibit B.3. Characteristics of Fully Rated Programs, With and Without Complete Concurrent Validity Data

	Sites With Complete Observation Data (*N* = 175)		Sites Without Complete Observation Data (*N* = 297)		
	N	**Percentage**	***N***	**Percentage**	***p***
Program Type	175		297		.004
Center-Based	148	85%	217	73%	
FCCH	27	15%	80	27%	
Funding Sources (Programs May Have Multiple Sources)	167		285		
Standards-Based Public Funding (CSP, Title 5, or Head Start)	148	89%	234	82%	.065
First 5 California CSP 1 or CSP 2 funding	65	39%	157	55%	.001
California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	87	52%	162	57%	.327
Federal Head Start or Early Head Start fundingFederal Head Start or Early Head Start funding	70	42%	79	28%	.002
State/Federally Funded Child Care Subsidy Vouchers	35	21%	134	47%	<.0001
Private Pay	47	28%	145	51%	<.0001
Language Spoken With Children	163		282		
Non-English Language Spoken with Children	113	69%	143	51%	<.0001
Spanish Spoken With Children	112	69%	137	49%	<.0001
Consortia	175		297		<.0001
Full QRIS Rating With Local Adaptations	175		297		.231
Tier 1	2	1%	2	1%	
Tier 2	23	13%	62	21%	
Tier 3	62	35%	93	31%	
Tier 4	78	45%	118	40%	
Tier 5	10	6%	22	7%	
	N	**Mean (SD)**	***N***	**Mean (SD)**	***p***
Total Enrollment					
Average Total Enrollment, Centers	146	57.6 (33.6)	216	49.7 (31.2)	.023
Average Total Enrollment, Family Child Care Homes	27	8.4 (3.4)	80	9.3 (4.3)	.332

Source: Common Data Elements 2014.

Notes: *p* values are based on χ^2 tests for all comparisons except average total enrollment, which is based on a *t* test. QRIS ratings presented in this table are those reported by Consortia using local adaptations, and the distribution shown in this table differs from the distribution of ratings without local adaptations that were calculated for study analyses.

100 Percent Sample

The 100 percent sample includes all multi-classroom centers with CLASS and ERS data available on every eligible classroom in the center and represents a subsample of programs with full ratings. The 100 percent sample is used only for analyses comparing the element scores and QRIS ratings of centers using different combinations of classrooms. Initially, a sample of 59 programs was randomly drawn from the subset of centers with full ratings that had two or more classrooms. Recruitment of programs for the 100 percent sample occurred as part of recruitment for the concurrent validity sample. The acceptance rate was quite low among the sampled centers (36 percent), so an additional five centers with complete existing data available from Consortia were added to the sample. Given the low participation rate, it is important to examine the extent to which the characteristics of the participating centers differ from the characteristics of centers that were selected for the 100 percent sample but did not participate. These analyses do not include FCCHs because multiple classrooms only occur in centers; therefore, the results of analyses using the 100 percent sample analyses do not apply to FCCHs.

Exhibit B.4 presents details on the comparison of characteristics of centers with and without complete data for the 100 percent sample analyses. Characteristics between programs that did and did not participate in the 100 percent sample analyses have few significant differences; however, the lack of statistical significance in differences partly reflects the small number of programs in each group, so less attention should be paid to statistical significance in comparing the programs that did and did not participate. Overall, the pattern of differences in program characteristics between programs that did and did not participate in the 100 percent sample appears quite similar to the pattern of differences between programs that did and did not participate in the concurrent validity sample. The prevalence of standards-based public funding was similar among programs that did and did not participate in the 100 percent sample study, although there did appear to be differences in specific funding types. A much higher percentage of participating programs used Spanish with children. There were no significant differences between groups in the distribution of QRIS ratings with local adaptations, although the ratings appeared to skew lower among participating than nonparticipating programs. There were significant differences in participation by Consortia, with participation rates ranging from none of the selected programs to 100 percent of the selected programs.

Exhibit B.4. Characteristics of Fully Rated Centers, With and Without Complete 100 Percent Data

	Sites with Complete Observation Data (*N* = 26)		Sites Without Complete Observation Data (*N* = 38)		
	N	**Percentage**	***N***	**Percentage**	***p***
Funding Sources (Programs May Have Multiple Sources)	25		38		
Standards-Based Public Funding (CSP, Title 5, or Head Start)	23	92%	37	97%	.328
First 5 California CSP 1 or CSP 2 funding	9	36%	21	55%	.134
California Title 5 (State Preschool, General Child Care, or CalSAFE) funding	15	60%	22	58%	.868
Federal Head Start or Early Head Start funding	16	64%	13	34%	.020
State/Federally Funded Child Care Subsidy Vouchers	5	20%	12	32%	.311
Private Pay	9	36%	15	39%	.781
Language Spoken With Children	25		38		
Non-English Language Spoken With Children	18	72%	19	50%	.083
Spanish Spoken With Children	18	72%	16	42%	.020
Consortia	26		38		.002
Full QRIS Rating With Local Adaptations	26		38		.257
Tier 1	0	0%	0	0%	
Tier 2	0	0%	2	5%	
Tier 3	14	54%	12	32%	
Tier 4	11	42%	22	58%	
Tier 5	1	4%	2	5%	
	N	**Mean (SD)**	***N***	**Mean (SD)**	***p***
Total Enrollment					
Average Total Enrollment, Centers	26	75.4 (34.9)	38	64.6 (37.0)	.245

Source: Common Data Elements 2014.

Notes: *p* values are based on χ^2 tests for all comparisons except average total enrollment, which is based on a *t* test. QRIS ratings presented in this table are those reported by Consortia using local adaptations, and the distribution shown in this table differs from the distribution of ratings without local adaptations that were calculated for study analyses.

Statistical Power

In our study planning phase, we conducted statistical power analysis to determine sample sizes for the main concurrent validity analyses. Statistical power analysis is a method of calculating the minimum sample size needed to conduct meaningful statistical analyses when comparing

group differences (such as differences between programs at each QRIS rating level). The goal of statistical power analysis is to ensure that a study is appropriately designed to answer the research questions and also to control study expenses by selecting only the number of programs needed for the analyses.

In study planning, we estimated that concurrent validity analyses would require a sample size of 350 programs to detect differences between groups that are small to medium in size (*f* effect sizes of 0.17 to 0.19) and that concurrent validity analyses using 150 programs would permit us to detect differences that are medium to large in magnitude (*f* effect sizes of 0.26 to 0.29). To calculate these sample size estimates, we used power criteria of 0.80, which means that the analysis has an 80 percent chance of correctly rejecting the null hypothesis when it is false. Power of 0.80 is widely considered to be an acceptable level.

As described previously, the number of programs that received complete CLASS and PQA observations for the concurrent validity analyses is lower than expected. Furthermore, because of the small number of FCCHs that participated in the study, concurrent validity analyses are only conducted for centers. Analyses could not combine data on centers and FCCHs because the QRIS ratings are calculated differently for each program type and do not represent the same measure of quality.

Complete CLASS scores are available for 139 centers, and retrospective power analyses find that the adjusted power for the ANOVA analyses with the CLASS total scores is 0.58, lower than the desired level of 0.80. Complete PQA Form A scores are available for 140 centers, and retrospective power analyses find that the adjusted power for the ANOVA analyses with the PQA Form A total scores is 0.35, far lower than desired. The low power estimates indicate that some study analyses have sample sizes too small to make a definitive determination about whether statistically significant differences exist between QRIS rating levels. In other words, some analyses might miss potentially significant differences that would be detected with a larger sample size. This does not mean that the analysis results are necessarily incorrect, but it does indicate that the analyses that find no significant differences are not conclusive in their findings. Analyses that do find significant differences are not affected by the low power estimates.

The sample of 472 programs with complete California QRIS ratings and the subsamples of 365 centers and 107 FCCHs were adequate for all descriptive, predictive, and comparative analyses conducted using that data.

Data Limitations

Limitations of the data available for the study analyses include a low sample size for the planned analyses (as described in the section on statistical power), and some indication that programs participating in the study differ from programs in the QRIS that did not participate in the study, in particular in terms of funding source, language of instruction, and, to some extent, rating. These limitations mean that the study results should be interpreted with some caution. First, concurrent validity ANOVA analyses that have nonsignificant results are not conclusive because the small sample size increases the chance of a false negative result in the analyses. Second, the analysis results apply to the programs that participated in the study but may not apply to other programs in the QRIS. In particular, given the extent of differences between programs with and

without full QRIS ratings among those participating in the QRIS, the results of study analyses may not apply to other programs that enrolled in the QRIS but did not have a final rating at the time of the study. Also, results of analyses using classroom observation data collected for the study are not applicable to FCCHs and also may not apply to other programs in the QRIS that did not participate in classroom observations.

Measures

The validation study draws on two primary sources of data: the extant data provided by the Consortia (the "Common Data Elements") and classroom observations conducted primarily by the study team. Descriptions of these sources and their measures follow.

Existing State Data on Programs Participating in California's QRIS

For the study analyses, AIR collected extant data on the program characteristics and QRIS ratings of programs participating in the QRIS as of January 2014. Each of the 17 Consortia in the state collected data on their local participating programs separately, using different procedures and database systems but following specific statewide requirements for QRIS reporting. The data submitted to the state using the QRIS reporting requirements are referred to as the Common Data Elements and include data on program type, enrollment, funding sources, languages spoken in the program, element scores, the total of the element scores, the QRIS rating, and the program average CLASS scores used to calculate the CLASS element scores. In addition, as noted previously, six Consortia also provided some classroom-level CLASS and ERS data to supplement the concurrent validity sample.

California's QRIS permits participating Consortia to make local adaptations to the QRIS rating criteria for tiers 2 and 5. To ensure comparability of the QRIS ratings for the study analyses further, AIR used the element score data for each program to simulate QRIS ratings for programs in all Consortia using the California QRIS rating criteria without local adaptations, to the extent possible. In most cases, Consortia used the same criteria for element scores, but two of the Consortia added unique local criteria to the California QRIS criteria for element scores and could not provide raw data to determine element scores without the local criteria. In those two counties, the simulated California QRIS ratings are not perfectly comparable to other counties.

Classroom Observation Measures

To measure classroom quality, we conducted observations within the settings using seven different data collection protocols:

1. The CLASS Pre-K (Pianta, La Paro, and Hamre 2008)—used in all sampled classrooms where the majority of children were preschool-aged.

2. CLASS Toddler (La Paro, Hamre, and Pianta 2012)—used in all sampled classrooms where the majority of children were toddlers.

3. The PQA (HighScope Educational Research Foundation 2003)—used in all sampled classrooms where the majority of children were preschool aged.

4. The Infant-Toddler PQA (Hohmann, Lockhart, and Montie 2013)—used in all sampled classrooms where the majority of children were infants or toddlers.

5. The Family Child Care PQA (High/Scope Educational Research Foundation 2009)—used in all sampled FCCHs.

6. The ECERS-R (Harms, Clifford, and Cryer 2005)—used in all classrooms in the 100 percent subsample where the majority of children were preschool aged.

7. The ITERS-R (Harms, Cryer, and Clifford 2006)—used in all classrooms in the 100 percent subsample where the majority of children were infants or toddlers.

A description of each measure follows.

The CLASS

The CLASS was developed by the Center for Advanced Study in Teaching and Learning at the University of Virginia and has been used widely for research and professional development purposes. The CLASS Pre-K organizes teacher and student interactions into three broad domains: Emotional Support, Classroom Organization, and Instructional Support, which are further subdivided into 10 dimensions that describe the complex classroom environment, as shown in Exhibit B.5.

Research suggests that for healthy social-emotional development, children need to feel safe with their caregiver or educator and in their early education and care environment. CLASS examines how teachers interact with children to create warm relationships and a positive climate in the classroom. CLASS also looks at how teachers interact with their students to promote cognitive development—for example, how they foster higher-level thinking (Pianta, La Paro, and Hamre 2008).

The internal consistency of CLASS Pre-K dimension scores across four cycles ranges from 0.79 for Instructional Learning Formats to 0.91 for Emotional Support. Internal consistency is somewhat higher among the Emotional Support dimensions than among the Classroom Organization or Instructional Support dimensions. The CLASS Pre-K also has sound validity. It was evaluated during a 10-year period as part of the National Center for Early Development and Learning (NCEDL) Multi-State Study of Prekindergarten and Study of State-Wide Early Education Programs (SWEEP) and the NICHD Study of Early Child Care and Youth Development. Together, these studies conducted observations in more than 3,000 early childhood classrooms and found that children in classes with higher CLASS scores go on to make higher academic and social gains than children in classrooms with lower CLASS scores. CLASS was also found to be valid at different ages (Pianta, La Paro, and Hamre 2008) and correlated with other measures of classroom quality.

Exhibit B.5. Description of CLASS Pre-K Domains and Dimensions

Domain	Dimensions
Emotional Support	**Positive Climate.** Positive Climate reflects the emotional connection between the teacher and students and among students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.
	Negative Climate. Negative Climate reflects the overall level of expressed negativity in the classroom; the frequency, quality, and intensity of teacher and peer negativity are key to this scale.
	Teacher Sensitivity. Teacher Sensitivity encompasses the teacher's awareness of and responsiveness to students' academic and emotional needs; high levels of sensitivity facilitate students' ability to actively explore and learn because the teacher consistently provides comfort, reassurance, and encouragement.
	Regard for Student Perspectives. Regard for Student Perspectives captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests, motivations, and points of view, and encourage student responsibility and autonomy.
Classroom Organization	**Behavior Management.** Behavior Management encompasses the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior.
	Productivity. Productivity considers how well the teacher manages instructional time and routines and provides activities for students so that they have the opportunity to be involved in learning activities.
	Instructional Learning Formats. Instructional Learning Formats focus on the ways in which the teacher maximizes students' interest, engagement, and ability to learn from lessons and activities.
Instructional Support	**Concept Development.** Concept Development measures the teacher's use of instructional discussions and activities to promote students' higher-order thinking skills and cognition, and the teacher's focus on understanding rather than on rote instruction.
	Quality of Feedback. Quality of Feedback assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation.
	Language Modeling. Language Modeling captures the quality and amount of the teacher's use of language-stimulation and language-facilitation techniques.

Source: CLASS Manual, Pre-K (Pianta, La Paro, and Hamre 2008).

The CLASS Toddler tool was adapted from the CLASS Pre-K tool and also incorporates best practices for toddler development from the literature (La Paro, Hamre, and Pianta 2012). The CLASS Toddler organizes teacher and student interactions into two broad domains: Emotional and Behavioral Support and Engaged Support for Learning, which are further subdivided into eight dimensions that describe the complex classroom environment, as shown in Exhibit B.6. The CLASS Toddler has been used in some pilot studies, and the authors currently are in the process of conducting further validation work on the tool.

Exhibit B.6. Descriptions of CLASS Toddler Domains and Dimensions

Domain	Dimensions
Emotional and Behavioral Support	**Positive Climate.** Positive Climate reflects the connection between the teacher and children and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.
	Negative Climate. Negative Climate reflects the overall level of expressed negativity in the classroom. The frequency, quality, and intensity of teacher and child negativity are the key to this scale.
	Teacher Sensitivity. Teacher Sensitivity encompasses the teacher's responsiveness of children's individual needs and emotional functioning. The extent to which the teacher is available as a secure base (being there to provide comfort, reassurance, and encouragement) is included in this rating.
	Regard for Child Perspectives. Regard for Child Perspectives captures the degree to which the teacher's interactions with children and classroom activities emphasize children's interests, motivations, and points of view and encourage children's responsibility and independence.
	Behavior Guidance. Behavior Guidance encompasses the teacher's ability to promote behavioral self-regulation in children by using proactive approaches, supporting positive behavior, and guiding and minimizing problem behavior.
Engaged Support for Learning	**Facilitation of Learning and Development.** Facilitation of Learning and Development considers how well the teacher facilitates activities to support children's learning and developmental opportunities. How the teacher connects and integrates learning into activities and tasks should be included in this rating.
	Quality of Feedback. Quality of Feedback assesses the degree to which the teacher provides feedback (in response to what children say and/or do) that promotes learning and understanding and expands children's participation.
	Language Modeling. Language Modeling captures the quality and amount of the teacher's use of language-stimulation and language-facilitation techniques to encourage children's language development.

Source: CLASS Manual, Toddler (La Paro, Hamre, and Pianta 2012).

The PQA

The PQA is a rating instrument designed to evaluate the quality of early childhood programs and identify staff training needs. The three versions of the PQA were developed by HighScope Educational Research Foundation. The measures identify the structural characteristics and dynamic relationships that effectively promote the development of young children, encourage involvement of families and communities, and create supportive working environments for staff. The PQA examines multiple dimensions of program implementation, from the physical characteristics of the setting to the nature of adult-child interaction to program staffing and management.

The Preschool PQA measures seven areas of program quality: learning environment, daily routine, adult-child interaction, curriculum planning and assessment, parent involvement and family services, staff qualifications and development, and program management. Observers rate a number of items for each of the seven areas based on observation and answers to interview questions. More details on the items in the Preschool PQA are included in Exhibit B.7.

Exhibit B.7. Preschool PQA Sections and Items

Section	Item	
I. Learning Environment	■ Safe and healthy environment ■ Defined interest areas ■ Logically located interest areas ■ Outdoor space, equipment, materials ■ Organization and labeling of materials	■ Varied and open-ended materials ■ Plentiful materials ■ Diversity-related materials ■ Displays of child-initiated work
II. Daily Routine	■ Consistent daily routine ■ Parts of the day ■ Appropriate time for each part of day ■ Time for child planning ■ Time for child-initiated activities ■ Time for child recall	■ Small-group time ■ Large-group time ■ Choices during transition times ■ Cleanup time with reasonable choices ■ Snack or meal time ■ Outside time
III. Adult-Child Interaction	■ Meeting basic physical needs ■ Handling separation from home ■ Warm and caring atmosphere ■ Support for child communication ■ Support for non-English speakers ■ Adults as partners in play ■ Encouragement of child initiatives	■ Support for child learning at group times ■ Opportunities for child exploration ■ Acknowledgment of child efforts ■ Encouragement of peer interactions ■ Independent problem solving ■ Conflict resolution
IV. Curriculum Planning and Assessment	■ Curriculum model ■ Team teaching ■ Comprehensive child records	■ Anecdotal note taking by staff ■ Use of child observation measure
V. Parent Involvement and Family Services	■ Opportunities for involvement ■ Parents on policy-making committees ■ Parent participation in child activities ■ Sharing of curriculum information ■ Staff-parent informal interactions	■ Extending learning at home ■ Formal meetings with parents ■ Diagnostic/special education services ■ Service referrals as needed ■ Transition to kindergarten
VI. Staff Qualifications and Staff Development	■ Program director background ■ Instructional staff background ■ Support staff orientation and supervision ■ Ongoing professional development	■ Inservice training content and methods ■ Observation and feedback ■ Professional organization affiliation
VII. Program Management	■ Program licensed ■ Continuity in instructional staff ■ Program assessment ■ Recruitment and enrollment plan	■ Operating policies and procedures ■ Accessibility for those with disabilities ■ Adequacy of program funding

The Preschool PQA has been used extensively as a research tool by trained independent raters in more than 800 preschool classrooms and child care centers. The authors report a high level of internal consistency and evidence of validity for the overall measure. The authors report that the Preschool PQA is significantly correlated with other measures of program quality and child outcomes, such as the ECERS and the Caregiver Interaction Scale. The national Training for Quality study also showed that the PQA total score and all the subscales were positively and significantly associated with concurrent measures on the language scale of the Developmental Indicators for the Assessment of Learning Revised (DIAL-R).

The Infant-Toddler PQA was developed for use in center-based classrooms serving children aged 0–36 months. The instrument measures seven domains of curriculum implementation and program operations in child care settings: Learning Environment; Schedules and Routines; Adult-Child Interaction; Curriculum Planning and Child Observation; Parent Involvement and Family Services; Staff Qualifications and Staff Development; and Program Management. More details on the items in each domain of the Infant-Toddler PQA are included in Exhibit B.8. The agency items or sections V–VII are the same as on the Preschool PQA and are only measured once if a center has both preschool and infant-toddler classrooms.

Exhibit B.8. Infant-Toddler PQA Domains and Items

Section	Item	
I. Learning Environment	■ Safe and healthy environment ■ Spaces for sleeping, eating, and bodily care ■ Spaces for play and movement	■ Accessible sensory materials ■ Children's photos, creations ■ Accessible, safe, outdoor space
II. Schedules and Routines	■ Flexible, predictable schedule ■ Comfortable arrivals/departures ■ Child-initiated choice times ■ Bodily care choices ■ Smooth transitions	■ Child-centered feedings/meals ■ Fluid, dynamic group times ■ Nature-based outside times ■ Individualized naptimes
III. Adult-Child Interaction	■ Long-term adult-child relationships ■ Child-adult trust ■ Child-adult partnerships	■ Children's intentions ■ Children's social relationships ■ Children's conflict resolution
IV. Curriculum Planning and Child Observation	■ Comprehensive curriculum ■ Child observation and planning ■ Assessing developmental progress	■ Individualized planning by caregivers
V. Parent Involvement and Family Services	■ Opportunities for involvement ■ Parents on policy-making committees ■ Parent participation in child activities ■ Sharing of curriculum information ■ Staff-parent informal interactions	■ Extending learning at home ■ Formal meetings with parents ■ Diagnostic/special education services ■ Service referrals as needed ■ Transition to kindergarten
VI. Staff Qualifications and Staff Development	■ Program director background ■ Instructional staff background ■ Support staff orientation and supervision ■ Ongoing professional development	■ Inservice training content and methods ■ Observation and feedback ■ Professional organization affiliation
VII. Program Management	■ Program licensed ■ Continuity in instructional staff ■ Program assessment ■ Recruitment and enrollment plan	■ Operating policies and procedures ■ Accessibility for those with disabilities ■ Adequacy of program funding

The Family Child Care PQA measures four domains of quality for family child care programs: Daily Schedule, Learning Environment, Provider-Child Interaction, and Safe and Healthy Environment. More details on the items included in each domain are presented in Exhibit B.9.

Exhibit B.9. Family Child Care PQA Domains and Items

Section	Item	
I. Daily Schedule	■ Consistent daily schedule ■ Child-initiated activities ■ Adult-initiated group activities ■ Cleanup time with choices	■ Snacks or meals ■ Outside play ■ Nap, rest, or quiet time ■ Child planning
II. Learning environment	■ Space for play ■ Logically arranged interest areas, with easy access ■ Outside space with equipment and materials ■ Materials are systematically stored and labeled ■ Materials are accessible to children	■ Materials are varied, manipulative, open ended, and appeal to multiple senses ■ Materials are plentiful ■ Materials reflect human diversity and the positive aspects of children's lives ■ Adult and child work is on display
III. Provider-Child Interaction	■ Supportive arrivals and departures ■ Warm and caring atmosphere ■ Encouragement and support for child language, verbal and nonverbal ■ Support for non-English speakers ■ Adults participate as partners in play ■ Support for child learning during group activities ■ Opportunities for child exploration at own pace	■ Acknowledgement of child efforts ■ Encouragement of peer interactions ■ Opportunities for self-help and solving problems with materials ■ Encouragement of conflict resolution ■ Use of television and computers
IV. Safe and Healthy Environment	■ Spaces are free of physical hazards ■ Healthy hand-washing routines are in place ■ Safe and healthy toileting and diapering routines are in place ■ Food preparation practices are healthy and safe	■ Resting/napping equipment and routines are safe ■ Animals and pets are healthy ■ Emergency equipment and procedures are in place

The ERS

The ERS are the most commonly used measure of quality in early childhood classrooms (Pianta, La Paro, and Hamre 2008). The scales are designed to assess process quality in an early childhood or school age care group. Process quality consists of the various interactions that go on in a classroom between staff and children, staff, parents, and other adults, among the children themselves, and the interactions children have with the many materials and activities in the environment, as well as those features, such as space, schedule, and materials that support these interactions. Process quality is assessed primarily through observation and has been found to be

more predictive of child outcomes than structural indicators, such as staff-to-child ratio, group size, cost of care, and even type of care, for example child care center or FCCH (Whitebook, Howes, and Phillips 1995).

The ECERS-R is designed to be used with one room or one group at a time, for children 2½ through 5 years of age in center-based programs. The 43 items of the ECERS-R comprise seven subscales (see Exhibit B.10). The authors report strong internal consistency and evidence of predictive validity for the original versions of the scales (e.g., Peisner-Feinberg and Burchinal 1997; Whitebook, Howes, and Phillips 1989) and indicate that the revised version would be expected to maintain that form of validity.

Exhibit B.10. ECERS-R Subscales and Items

Subscale	Item	
Space and Furnishings	■ Indoor Space ■ Furniture for routine care, play and learning ■ Furnishings for relaxation and comfort ■ Room arrangement for play	■ Space for privacy ■ Child-related display ■ Space for gross motor play ■ Gross motor equipment
Personal Care Routines	■ Greeting/departing ■ Meals/snacks ■ Nap/rest	■ Toileting/diapering ■ Health practices ■ Safety practices
Language-Reasoning	■ Books and pictures ■ Encouraging children to communicate	■ Using language to develop reasoning skills ■ Informal use of language
Activities	■ Fine motor ■ Art ■ Music/movement ■ Blocks ■ Sand/water	■ Dramatic play ■ Nature/science ■ Mathematics/number ■ Use of TV, video, and/or computer ■ Promoting acceptance of diversity
Interaction	■ Supervision of gross motor activities ■ General supervision of children (other than gross motor)	■ Discipline ■ Staff-child interactions ■ Interactions among children
Program Structure	■ Schedule ■ Free play ■ Group time	■ Provisions for children with disabilities
Parents and Staff	■ Provisions for parents ■ Provisions for personal needs of staff ■ Provisions for professional needs of staff	■ Staff interaction and cooperation ■ Supervision and evaluation of staff ■ Opportunities for professional growth

The ITERS-R is designed to be used with one room or one group at a time, for children from birth through 30 months of age. The 39 items of the ITERS-R comprise seven subscales (see Exhibit B.11). The authors report high levels of internal consistency and evidence of concurrent and predictive validity for the original versions of the environment rating scales, citing associations with structural measures of quality as caregiver-child ratios and caregiver education level (Cryer and others 1999; Phillipsen and others 1997) and evidence of predictive validity in relation to child development (Burchinal and others 1996; Peisner-Feinberg and others 1999). The authors also report that because the current revisions maintain the basic properties of the original instruments, the revised scales are expected to maintain validity (Harms, Cryer, and Clifford 2006; Harms, Cryer, and Clifford, 2007).

Exhibit B.11. ITERS-R Subscales and Items

Subscale	Item	
Space and Furnishings	■ Indoor Space ■ Furniture for routine care and play ■ Provision for relaxation and comfort	■ Room arrangement ■ Display for children
Personal Care Routines	■ Greeting/departing ■ Meals/snacks ■ Nap	■ Diapering/toileting ■ Health practices ■ Safety practices
Language and Talking	■ Helping children understand language ■ Helping children use language	■ Using books
Activities	■ Fine motor ■ Active physical play ■ Art ■ Music and movement ■ Blocks	■ Dramatic play ■ Sand and water play ■ Nature/science ■ Use of TV, video, and/or computer ■ Promoting acceptance of diversity
Interaction	■ Supervision of play and learning ■ Peer interaction	■ Staff-child interaction ■ Discipline
Program Structure	■ Schedule ■ Free play	■ Group play activities ■ Provisions for children with disabilities
Parents and Staff	■ Provisions for parents ■ Provisions for personal needs of staff ■ Provisions for professional needs of staff ■ Staff interaction and cooperation	■ Staff continuity ■ Supervision and evaluation of staff ■ Opportunities for professional growth

Observation Data Collection Procedures

In the spring of 2014 (April through June) all programs were observed using the PQA and CLASS. A subsample of center-based programs was also observed using the ERS. Wherever possible, up to two to three observers (i.e., one PQA, one CLASS, and one ERS as needed) observed in a classroom at the same time to minimize the number of days a classroom was being observed. Observation data collected by the study team were supplemented by data provided by a few Consortia in order to reduce the burden on sites. In total, data were collected from a total of 640 observations: 281 CLASS, 214 PQA, and 145 ERS observations conducted through the spring of 2014.

Training and Certification of Observers. A total of 107 staff members were trained to conduct observations for the study in the spring of 2014. The training and certification process for each observation protocol was slightly different; the certification rates varied by measure as well. See Exhibit B.12 for the certification rates for each tool.

For the CLASS, staff members participated in a two-day classroom-based training led by a certified CLASS trainer. To certify as reliable on the CLASS measure staff members had to pass the CLASS online certification test by rating videos of classrooms. In order to pass the online certification test, staff members had to have 80 percent of all codes within one point of the master codes and at least two out of five codes within one point of the master codes within each dimension. For the CLASS measure, the authors allow users three attempts to certify by passing the test.

Exhibit B.12. Number of Field Staff Who Were Trained, Were Certified, and Conducted Observations.

	Number of Staff Trained	Number of Staff Certified	Number of Staff Who Conducted Observations
CLASS Pre-K	30	23	20
CLASS Toddler	15	12	12
Preschool PQA	32	27	26
Infant-Toddler PQA	8	8	8
Family Child Care PQA	6	4	4
ECERS-R	11	4	4
ITERS-R	5	3	2

For the PQA, staff members participated in a two-day classroom-based training led by a certified PQA trainer. The certification process for the PQA is slightly different depending on the version (e.g., Preschool, Infant-Toddler, Family Child Care). For the Infant-Toddler and Family Child Care versions of the PQA, certification required staff to watch and score a video on the final day of training and receive a passing score. To certify on the Preschool PQA, staff had to watch and score an online video in the weeks after training was completed and receive a passing score.

For the ERS, first staff members needed to complete two online courses successfully (i.e., ERS 101 and either ITERS 101 or ECERS 101) by passing the brief quiz at the end of each course.

Next, staff members participated in at least four to five days of live practice observations (one observation per day) and reliability discussions with an ERS trainer.[26] In order to be certified as a reliable ERS observer, staff members needed to complete a minimum of three consecutive practice observations scoring 85 percent of all items within one point of the ERS trainer's scores across the three days.

The authors of the ERS recommend that ERS observers check their reliability by conducting paired observations after every 10 observations to ensure that observers continue to reliably score.[27] To this end, for every 10 ITERS-R or ECERS-R observations completed by a single observer, a paired observation with a second trained ITERS-R or ECERS-R observer was conducted. After the observation, observers compared scores and recorded their scores on an Interrater Reliability Sheet just as they had for reliability discussions during training. Observers discussed scores they selected for each item. The score booklets and Interrater Reliability Sheet were return to the data collection manager for review and determination of continued reliability. Reliability was achieved on all peer-observed observations for the spring 2014 QRIS ERS observations.

Classroom Observation Data Collection Challenges

Scheduling and conducting the observations prior to the end of the program year was the main data collection challenge we faced. As much as possible, we tried to complete observations of all sampled classrooms prior to when programs closed before summer break. In some cases, this meant we were conducting observations in classrooms in less than ideal circumstances (e.g., the teacher had begun to pack up her classroom in preparation for the program closing in the coming days). In other cases, this meant we were unable to complete observations in all sampled classrooms at a particular site.

Summary of Data Challenges and Limitations

The results presented in the body of this report should be interpreted within the context of several data challenges and limitations. First, a little more than a third of participating programs had a full, nonprovisional rating and could therefore be included in study analyses. The programs with provisional ratings appear to differ from the fully rated programs in several ways, suggesting that fully rated programs are not representative of the entire population of programs in the QRIS, thus limiting the generalizability of the validation study results presented in this report.

Second, we obtained a smaller than anticipated sample for the concurrent validity analyses, and there is some indication that programs participating in the classroom observations differ from programs in the QRIS that did not participate. These limitations mean that the study results should be interpreted with some caution. In particular, concurrent validity analyses that have nonsignificant results are not conclusive because the small sample size limits our ability to detect small differences. In addition, the analysis results apply to the programs that participated in the

[26] Debby Cryer, one of the authors of the ERS, trained a total of six staff. Other ERS trainers included Environment Rating Scales Institute (ERSI) staff as well as ERSI approved state anchors.

[27] The authors of the CLASS and the PQA do not require that observers conduct paired observations to check for drift once the observer is certified.

classroom observations, and results might differ if a broader group of programs participated in the study.

Third, the concurrent validity analyses using the CLASS measure, and also the analyses using the 100 percent sample, included a combination of data collected by the study team and existing data collected by independent observers for Consortia's QRIS ratings. Using data collected for the QRIS ratings has several potential limitations. First, the study team was not able to verify the reliability of the classroom observation data collected by Consortia. However, the Consortia are required to follow stringent requirements for training and certification of classroom observers, and the study team is confident that the data can be considered reliable. Second, the study team's observation data was collected in spring of 2014. In contrast, the Consortia observation data was collected prior to January 2014, the cutoff date for inclusion in this study, and could have been collected as early as August 2013. It may be that results using earlier data would be biased in comparison to analyses using current data. However, sensitivity analyses for the concurrent validity analyses found no differences in findings when analyses were run separately for each data source.

Finally, it is also important to remember that the QRIS is relatively new and not fully implemented (as evidenced by the large number of provisionally rated programs). Thus, results presented in the report should be interpreted within the context of the system's stage of development and current participants.

Analysis Methods

Exhibit B.13 summarizes the analysis methods used in the study and also presents the sample and type of data used for each analysis.

Exhibit B.13. Methods, samples, and measures used for each analysis reported.

	Sample			Measures				
	Programs With Full QRIS Ratings	Concurrent Validity Sample	100 Percent Sample	QRIS Ratings and Element Scores	Program Characteristics	PQA Observations	CLASS Observations	ERS Observations
Chapter 4: Distribution and Reliability of QRIS Ratings								
Rating Distributions: distribution of QRIS ratings and elements describing the number of programs at each tier, for all programs and by program type	x			x				
Predictors of Ratings: Ordinal logistic regression models indicating which program characteristics predicting QRIS ratings and element scores	x			x	x			
Internal Consistency: Cronbach's alpha, assessing the extent to which the QRIS rating measures a single construct	x			x				
Measurement Properties: Percentages of programs that had the same score on each pair of elements	x			x				
Relationships Between Element Scores and Ratings: Spearman's rho, measuring the correlations between pairs of element scores and between element scores and the QRIS rating	x			x				
Relationships Between Element Scores and Ratings: Average element scores at each QRIS rating level	x			x				
Chapter 5: Concurrent Validity								
Concurrent Validity: Analysis of Variance models describing the average scores on independent measures of observed quality by QRIS rating level		x		x		x	x	
Concurrent Validity: Analysis of Variance models describing the average scores on independent measures of observed quality by each element score level		x		x		x	x	

	Sample			Measures				
	Programs With Full QRIS Ratings	Concurrent Validity Sample	100 Percent Sample	QRIS Ratings and Element Scores	Program Characteristics	PQA Observations	CLASS Observations	ERS Observations
Chapter 6: Alternative Rating Approaches								
Simulation of Alternative Rating Approaches: distribution of California QRIS ratings with and without local adaptations, and of simulated ratings using alternative calculation approaches (see Exhibit 6.1 in Chapter 6 for detailed descriptions of the alternative rating approaches)	x			x				
Rating Comparisons: cross-tabulations of California QRIS ratings without local adaptations and simulated ratings using alternative calculation approaches, and of California QRIS ratings with and without local adaptations	x			x				
Concurrent Validity: Analysis of Variance models describing the average scores on independent measures of observed quality by each alternative rating approach tier		x		x		x	x	
Percentage of Classrooms Observed: Analyses examining the consistency of element scores and QRIS ratings of centers with multiple classrooms, comparing scores and ratings using program averages of 100 percent of classrooms (100 percent protocol) to all possible scores and ratings based on randomly selecting all possible combinations of classrooms in one third of classrooms as well as in one half of classrooms (rounding up)			x	x			x	x

Appendix C. Descriptive Comparisons of Classroom Observation Scores in Small Samples

Descriptive Average Scores by Rating Level, FCCHs

Exhibit C.1. CLASS Preschool Domain Scores by California QRIS Rating Level, FCCHs

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	0	—	—	—	0
2	4.52 (0.85)	16	5.56 (0.74)	4.81 (1.17)	2.20 (0.74)	10
3	—	3	—	—	—	3
4	—	1	—	—	—	1
5	—	0	—	—	—	0
All scores	4.67 (0.91)	20	5.76 (0.74)	4.95 (1.11)	2.67 (1.45)	14

Cells show the mean and standard deviation for the indicated CLASS scores at each element score. Results are descriptive only; the number of CLASS scores among FCCHs was too small to permit statistical comparison of the mean scores by rating level. Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit C.2. PQA Form A Preschool Domain Scores by California QRIS Rating Level, FCCHs

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Schedule	Provider-Child Interaction	Safe and Health Environment	*N*
1	—	0	—	—	—	—	0
2	4.00 (0.63)	19	3.85 (0.83)	4.02 (0.68)	3.78 (0.68)	4.58 (0.47)	19
3	4.59 (0.43)	5	4.52 (0.38)	4.62 (0.37)	4.54 (0.58)	4.69 (0.36)	5
4	—	2	—	—	—	—	2
5	—	1	—	—	—	—	1
All scores	4.20 (0.64)	27	4.07 (0.78)	4.23 (0.67)	4.03 (0.74)	4.64 (0.43)	27

Cells show the mean and standard deviation for the indicated PQA scores at each element score. Results are descriptive only; the number of PQA scores among FCCHs was too small to permit statistical comparison of the mean scores by rating level. Note that average PQA score data are not presented for rating levels with fewer than five observations.

Descriptive Average Scores by Rating Level, Toddler Classrooms in Centers

Exhibit C.3. CLASS and PQA Form A Toddler Domain Scores by California QRIS Rating Level, Centers

Element Score	CLASS Toddler Domain Scores			PQA Toddler Domain Scores				
	Engaged Support for Learning	Emotional and Behavioral Support	*N*	Learning Environ-ment	Schedules and Routines	Adult-Child Interaction	Curriculum Planning and Child Observation	*N*
1	—	—	0	—	—	—	—	0
2	—	—	2	—	—	—	—	2
3	2.96 (0.44)	6.02 (0.20)	6	2.87 (0.43)	3.37 (0.88)	3.05 (0.78)	3.20 (0.68)	8
4	2.71 (1.51)	5.20 (0.77)	5	2.79 (0.44)	3.16 (0.74)	2.86 (1.05)	3.41 (0.39)	7
5	—	—	1	—	—	—	—	1
All scores	2.84 (0.97)	5.59 (0.67)	14	2.87 (0.53)	3.45 (0.84)	3.15 (0.94)	3.43 (0.67)	18

Cells show the mean and standard deviation for the indicated CLASS and PQA scores at each element score. Results are descriptive only; the number of toddler scores among centers was too small to permit statistical comparison of the mean scores by rating level. Note that average CLASS and PQA score data are not presented for rating levels with fewer than five observations.

Appendix D. Element Score Analysis Results

Internal Consistency Results

Exhibit D.1. Internal Consistency of Element Scores, Centers

Element Score	Correlation Of Element score and Overall Scale (Item-Test)	Correlation of Element Score and Scale With Other Six Scores (Item-Rest)	Internal Consistency Without Element
Child Observation	0.605	0.423	0.446
Developmental and Health Screenings	0.533	0.151	0.583
Minimum Qualifications for Lead Teacher or FCCH	0.644	0.419	0.431
Effective Teacher-Child Interactions: CLASS	0.507	0.303	0.488
Ratios and Group Sizes	0.160	-0.039	0.585
Program Environment Rating Scales	0.533	0.300	0.486
Director Qualifications	0.611	0.395	0.447
Internal Consistency of All 7 Element Scores (Cronbach's α)			0.537

N = 363 centers, excluding 2 centers serving only infants.

Exhibit D.2. Internal Consistency of Element Scores, FCCHs

Element Score	Correlation of Element Score and Overall Scale (Item-Test)	Correlation of Element Score and Scale With Other 4 Scores (Item-Rest)	Internal Consistency Without Element
Child Observation	0.666	0.399	0.564
Developmental and Health Screenings	0.742	0.487	0.512
Minimum Qualifications for Lead Teacher or FCCH	0.609	0.305	0.618
Effective Teacher-Child Interactions: CLASS	0.410	0.241	0.631
Program Environment Rating Scales	0.706	0.503	0.515
Internal Consistency of All Five Domain Scores (Cronbach's α)			0.627

N = 107 FCCHs

Element Score Descriptive Results

Exhibit D.3. Correlations Between Element Scores and California QRIS Ratings, Centers

Element	Correlation (ρ)
Child Observation	0.549*
Developmental and Health Screenings	0.459*
Minimum Qualifications for Lead Teacher or FCCH	0.569*
Effective Teacher-Child Interactions: CLASS	0.449*
Ratios and Group Sizes	0.160*
Program Environment Rating Scales	0.481*
Director Qualifications	0.529*

N = 363 centers, excluding two centers serving infants only. Correlations are calculated using Spearman's ρ, a nonparametric correlation coefficient that can be interpreted in a similar way to Pearson's *r*.

* $p < .05$

Exhibit D.4. Correlations Between Element Scores and California QRIS Ratings, FCCHs

Element	Correlation (ρ)
Child Observation	0.534*
Developmental and Health Screenings	0.620*
Minimum Qualifications for Lead Teacher or FCCH	0.610*
Effective Teacher-Child Interactions: CLASS	0.355*
Program Environment Rating Scales	0.624*

N = 107 FCCHs. Correlations are calculated using Spearman's ρ, a nonparametric correlation coefficient that can be interpreted in a similar way to Pearson's *r*.

* $p < .05$

Exhibit D.5. Average Element Scores by California QRIS Rating Level, Centers

California QRIS Rating	*N*	Mean Element Score						
		CO	DHS	MQ	CLASS	RGS	ERS	DQ
Tier 1	0	—	—	—	—	—	—	—
Tier 2	20	2.20	2.25	2.10	2.35	4.05	2.10	1.95
Tier 3	123	3.43	2.90	3.09	3.27	3.83	3.33	3.37
Tier 4	188	4.09	4.02	4.16	3.74	4.17	3.85	4.24
Tier 5	32	4.66	4.94	4.72	4.47	4.31	4.84	4.66

Element score name abbreviations are as follows: CO = Child Observation; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher or FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; RGS = Ratios and Group Sizes; ERS = Program Environment Rating Scales; DQ = Director Qualifications

Exhibit D.6. Average Element Scores by California QRIS Rating Level, FCCHs

California QRIS Rating	*N*	Mean Element Score				
		CO	DHS	MQ	CLASS	ERS
Tier 1	0	—	—	—	—	—
Tier 2	57	1.81	1.53	1.84	3.14	2.16
Tier 3	34	3.00	2.74	3.38	3.29	2.97
Tier 4	11	3.55	3.91	3.82	3.82	4.27
Tier 5	5	3.80	5.00	4.20	4.40	4.80

Element score name abbreviations are as follows: CO = Child Observation; DHS = Developmental and Health Screenings; MQ = Minimum Qualifications for Lead Teacher or FCCH; CLASS = Effective Teacher-Child Interactions: CLASS; RGS = Ratios and Group Sizes; ERS = Program Environment Rating Scales; DQ = Director Qualifications

Element Score Concurrent Validity Results

Exhibit D.7. CLASS Total and Preschool Domain Scores by Child Observation Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	4	—	—	—	3
2	—	4	—	—	—	3
3	4.52 (0.59)	8	5.66 (0.30)	5.09 (0.38)	2.65 (1.27)	8
4	4.96 (0.62)	100	5.93 (0.60)	5.57 (0.71)	3.09 (0.84)	98
5	4.90 (0.62)	23	5.95 (0.70)	5.40 (0.70)	2.95 (0.88)	23
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 0.98		F[4,130] = 0.42	F[4,130] = 1.26	F[4,130] = 0.76	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average CLASS score data are not presented for element score levels with fewer than five observations.

Exhibit D.8. PQA Form A Total and Preschool Domain Scores by Child Observation Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	3	—	—	—	—	2
2	—	3	—	—	—	—	2
3	3.32 (0.37)	8	3.43 (0.35)	2.95 (0.53)	3.18 (0.71)	4.24 (0.58)	8
4	3.54 (0.51)	101	3.68 (0.53)	3.31 (0.61)	3.46 (0.71)	4.24 (0.51)[e]	97
5	3.48 (0.59)	25	3.63 (0.40)	3.27 (0.6)	3.44 (0.83)	3.64 (0.65)[d]	25
All scores	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.6)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,135] = 0.56		F[4,129] = 1.09	F[4,129] = 0.77	F[4,129] = 0.31	F[4,129] = 6.27***	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.9. PQA Form B Total and Domain Scores by Child Observation Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	3
2	—	—	—	—	4
3	3.79 (0.31)	4.13 (0.46)[e]	3.43 (0.46)	3.68 (0.41)	9
4	3.91 (0.42)[e]	4.16 (0.49)[e]	3.54 (0.52)	3.91 (0.46)	85
5	3.55 (0.60)[d]	3.55 (0.71)[c, d]	3.47 (0.76)	3.64 (0.65)	23
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 3.19*	F[4,119] = 6.3***	F[4,119] = 0.64	F[4,119] = 1.82	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.10. CLASS Total and Preschool Domain Scores by Developmental and Health Screening Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	4.95 (0.45)	17	5.92 (0.40)	5.39 (0.73)	3.09 (0.87)	15
2	4.88 (0.59)	28	5.88 (0.62)	5.56 (0.75)	2.96 (0.73)	28
3	—	4	—	—	—	4
4	4.53 (0.71)	13	5.67 (0.85)	5.26 (0.93)	2.52 (0.75)	13
5	4.99 (0.62)	77	5.99 (0.58)	5.57 (0.62)	3.12 (0.90)	75
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 1.66		F[4,130] = 0.95	F[4,130] = 0.92	F[4,130] = 1.54	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average CLASS score data are not presented for element score levels with fewer than five observations.

Exhibit D.11. PQA Form A Total and Preschool Domain Scores by Developmental and Health Screening Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	3.43 (0.58)	17	3.70 (0.51)	3.18 (0.78)	3.12 (0.73)	4.06 (0.83)	15
2	3.45 (0.48)	29	3.61 (0.46)	3.20 (0.57)	3.36 (0.66)	4.19 (0.64)	28
3	—	3	—	—	—	—	3
4	3.28 (0.50)	11	3.34 (0.72)	3.14 (0.48)	3.26 (0.75)	3.97 (0.59)	11
5	3.59 (0.52)	80	3.70 (0.48)	3.36 (0.6)	3.56 (0.73)	4.10 (0.55)	77
All scores	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.6)	3.44 (0.72)	4.11 (0.6)	134
ANOVA results	F[4,135] = 1.26		F[4,129] = 1.83	F[4,129] = 0.74	F[4,129] = 1.58	F[4,129] = 0.67	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.12. PQA Form B Total and Domain Scores by Developmental and Health Screening Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	3.74 (0.43)	3.87 (0.60)	3.60 (0.54)	3.72 (0.45)	16
2	3.67 (0.52)	3.83 (0.65)[e]	3.37 (0.62)	3.74 (0.57)	26
3	—	—	—	—	3
4	3.67 (0.43)	3.86 (0.47)	3.36 (0.60)	3.69 (0.45)	13
5	3.95 (0.41)	4.20 (0.52)[b]	3.63 (0.51)	3.91 (0.50)	66
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 2.63*	F[4,119] = 2.95*	F[4,119] = 1.54	F[4,119] = 1.11	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.13. CLASS Total and Preschool Domain Scores by Minimum Qualifications for Lead Teacher or FCCH Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	3	—	—	—	3
2	4.79 (0.60)	41	5.81 (0.54)	5.45 (0.60)	2.83 (0.98)	41
3	4.77 (0.58)	19	5.85 (0.52)	5.43 (0.73)	2.78 (0.63)	18
4	4.97 (0.65)	41	5.99 (0.67)	5.49 (0.68)	3.13 (0.85)	38
5	5.11 (0.55)	35	6.05 (0.62)	5.66 (0.75)	3.28 (0.83)	35
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 1.70		F[4,130] = 1.11	F[4,130] = 0.75	F[4,130] = 1.81	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average CLASS score data are not presented for element score levels with fewer than five observations.

Exhibit D.14. PQA Form A Total and Preschool Domain Scores by Minimum Qualifications for Lead Teacher or FCCH Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	2	—	—	—	—	2
2	3.40 (0.54)	39	3.54 (0.51)	3.24 (0.64)	3.16 (0.72)[e]	4.21 (0.47)	39
3	3.67 (0.48)	18	3.73 (0.57)	3.54 (0.61)	3.63 (0.56)	4.24 (0.58)	17
4	3.50 (0.48)	45	3.66 (0.50)	3.25 (0.57)	3.49 (0.65)	4.11 (0.53)	40
5	3.61 (0.54)	36	3.70 (0.51)	3.27 (0.60)	3.64 (0.79)[b]	3.99 (0.78)	36
All scores	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,135] = 1.56		F[4,129] = 0.67	F[4,129] = 0.94	F[4,129] = 3.24*	F[4,129] = 1.55	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.15. PQA Form B Total and Domain Scores by Minimum Qualifications for Lead Teacher or FCCH Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	2
2	3.82 (0.36)	4.10 (0.46)	3.38 (0.39)	3.84 (0.43)	35
3	3.79 (0.36)	4.07 (0.41)	3.35 (0.46)	3.83 (0.46)	18
4	3.88 (0.54)	4.08 (0.60)	3.63 (0.67)	3.85 (0.59)	40
5	3.80 (0.53)	3.87 (0.76)	3.72 (0.55)	3.78 (0.55)	29
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 0.24	F[4,119] = 0.95	F[4,119] = 2.57*	F[4,119] = 0.19	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.16. CLASS Total and Preschool Domain Scores by Effective Teacher-Child Interactions Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	1	—	—	—	0
2	—	2	—	—	—	2
3	4.69 (0.59)[d, e]	85	5.75 (0.60)[d, e]	5.31 (0.67)[d, e]	2.76 (0.83)[d, e]	84
4	5.22 (0.38)[c]	16	6.22 (0.41)[c]	5.78 (0.47)[c]	3.34 (0.71)[c]	16
5	5.35 (0.42)[c]	35	6.23 (0.52)[c]	5.93 (0.58)[c]	3.59 (0.71)[c]	33
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 11.16***		F[3,131] = 7.60***	F[3,131] = 9.77***	F[3,131] = 9.84***	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average CLASS score data are not presented for element score levels with fewer than five observations.

Exhibit D.17. PQA Form A Total and Preschool Domain Scores by Effective Teacher-Child Interactions Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	1	—	—	—	—	0
2	—	2	—	—	—	—	2
3	3.40 (0.52)[e]	84	3.52 (0.51)[e]	3.21 (0.59)	3.28 (0.75)[e]	4.02 (0.59)[e]	83
4	3.61 (0.50)	17	3.75 (0.53)	3.34 (0.64)	3.54 (0.65)	4.00 (0.64)	16
5	3.77 (0.43)[c]	35	3.92 (0.40)[c]	3.46 (0.61)	3.78 (0.58)[c]	4.44 (0.49)[b, c]	33
All scores	3.52 (0.52)	139	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,134] = 3.6**		F[3,130] = 5.74***	F[3,130] = 1.42	F[3,130] = 4.14**	F[3,130] = 5.74***	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.18. PQA Form B Total and Domain Scores by Effective Teacher-Child Interactions Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	1
2	—	—	—	—	2
3	3.76 (0.48)	3.97 (0.61)	3.46 (0.56)	3.75 (0.54)	75
4	3.93 (0.50)	4.13 (0.59)	3.63 (0.58)	3.91 (0.48)	13
5	3.98 (0.39)	4.17 (0.52)	3.69 (0.54)	3.99 (0.43)	32
All scores	3.84 (0.46)	4.05 (0.58)	3.54 (0.56)	3.83 (0.51)	123
ANOVA results	F[4,118] = 1.56	F[4,118] = 0.92	F[4,118] = 1.37	F[4,118] = 1.54	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not presented for element score levels with fewer than five observations.

Exhibit D.19. CLASS Total and Preschool Domain Scores by Ratios and Group Size Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	0	—	—	—	0
2	4.78 (0.31)	8	5.83 (0.34)	5.38 (0.42)	2.79 (0.53)	8
3	5.29 (0.56)	13	6.27 (0.58)	5.73 (0.75)	3.39 (0.93)	12
4	4.91 (0.58)	82	5.93 (0.55)	5.50 (0.65)	3.05 (0.85)	80
5	4.84 (0.69)	36	5.80 (0.72)	5.46 (0.81)	2.91 (0.94)	35
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 1.96		F[3,131] = 1.95	F[3,131] = 0.54	F[3,131] = 1.12	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.20. PQA Form A Total and Preschool Domain Scores by Ratios and Group Size Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	0	—	—	—	—	0
2	3.74 (0.40)	7	3.67 (0.63)	3.59 (0.51)	3.53 (0.44)	4.44 (0.36)	7
3	3.53 (0.49)	13	3.61 (0.43)	3.36 (0.70)	3.61 (0.66)	3.93 (0.60)	12
4	3.52 (0.52)	82	3.68 (0.49)	3.28 (0.59)	3.42 (0.74)	4.16 (0.60)	80
5	3.45 (0.54)	38	3.57 (0.56)	3.22 (0.61)	3.41 (0.76)	4.01 (0.62)	35
All scores	3.52 (0.52)	139	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 0.62		F[3,130] = 0.36	F[3,130] = 0.76	F[3,130] = 0.31	F[3,130] = 1.59	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.21. PQA Form B Total and Domain Scores by Ratios and Group Size Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	0
2	3.75 (0.31)	3.96 (0.47)	3.49 (0.46)	3.70 (0.33)	7
3	4.17 (0.53)	4.21 (0.62)	4.13 (0.64)[d, e]	4.13 (0.68)	7
4	3.82 (0.46)	4.03 (0.60)	3.52 (0.57)[c]	3.80 (0.49)	76
5	3.82 (0.45)	4.04 (0.56)	3.45 (0.45)[c]	3.87 (0.54)	34
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 1.38	F[3,120] = 0.24	F[3,120] = 3.12*	F[3,120] = 1.15	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.22. CLASS Total and Preschool Domain Scores by Program Environment Rating Scale Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	0	—	—	—	0
2	4.76 (0.58)[e]	15	5.87 (0.63)	5.50 (0.66)	2.67 (0.71)[e]	15
3	4.73 (0.66)[e]	60	5.73 (0.64)[e]	5.31 (0.67)[e]	2.88 (0.96)[e]	57
4	4.94 (0.54)	23	6.04 (0.54)	5.53 (0.79)	2.91 (0.82)	23
5	5.24 (0.42)[b, c]	41	6.15 (0.46)[c]	5.77 (0.59)[c]	3.45 (0.66)[b, c]	40
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 7.14***		F[3,131] = 4.55**	F[3,131] = 3.67*	F[3,131] = 5.11**	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.23. PQA Form A Total and Preschool Domain Scores by Program Environment Rating Scale Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	0	—	—	—	—	0
2	3.41 (0.28)	13	3.47 (0.43)[e]	3.29 (0.37)	3.27 (0.49)	4.12 (0.36)	13
3	3.35 (0.52)[e]	61	3.54 (0.51)[e]	3.17 (0.62)	3.20 (0.74)[e]	4.01 (0.62)	57
4	3.47 (0.57)	21	3.50 (0.54)[e]	3.19 (0.68)	3.39 (0.84)	4.00 (0.57)	20
5	3.79 (0.43)[c]	45	3.89 (0.43)[b, c, d]	3.48 (0.57)	3.82 (0.55)[c]	4.30 (0.62)	44
All scores	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 7.46***		F[3,130] = 5.83***	F[3,130] = 2.48	F[3,130] = 7.1***	F[3,130] = 2.21	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.24. PQA Form B Total and Domain Scores by Program Environment Rating Scale Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	0
2	3.74 (0.47)	3.98 (0.57)	3.41 (0.54)	3.74 (0.49)	12
3	3.78 (0.48)	4.01 (0.64)	3.50 (0.59)	3.74 (0.49)	56
4	3.85 (0.47)	4.04 (0.55)	3.63 (0.59)	3.80 (0.58)	19
5	3.93 (0.43)	4.10 (0.53)	3.59 (0.49)	4.00 (0.48)	37
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 0.91	F[3,120] = 0.23	F[3,120] = 0.6	F[3,120] = 2.17	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Note that average PQA score data are not reliable for element score levels with fewer than five observations.

Exhibit D.25. CLASS Total and Preschool Domain Scores by Director Qualifications Element Score and ANOVA Results, Centers

Element Score	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
1	—	0	—	—	—	0
2	4.80 (0.54)	26	5.83 (0.48)	5.37 (0.64)	2.90 (0.85)	25
3	5.16 (0.40)[e]	33	6.08 (0.50)	5.72 (0.58)[e]	3.30 (0.60)	32
4	4.98 (0.61)	45	6.00 (0.53)	5.63 (0.71)	3.11 (0.94)	44
5	4.70 (0.72)[c]	35	5.73 (0.78)	5.24 (0.71)[c]	2.78 (0.94)	34
All scores	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 4.09**		F[3,131] = 2.43	F[3,131] = 3.73*	F[3,131] = 2.34	

Cells show the mean and standard deviation for the CLASS scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.26. PQA Form A Total and Preschool Domain Scores by Director Qualifications Element Score and ANOVA Results, Centers

Element Score	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
1	—	0	—	—	—	—	0
2	3.45 (0.53)	22	3.59 (0.57)	3.33 (0.64)	3.15 (0.65)	4.15 (0.53)	21
3	3.63 (0.45)	34	3.74 (0.47)	3.33 (0.57)	3.58 (0.68)	4.30 (0.48)	33
4	3.56 (0.52)	46	3.73 (0.48)	3.30 (0.63)	3.52 (0.70)	4.10 (0.64)	44
5	3.41 (0.54)	38	3.48 (0.52)	3.21 (0.60)	3.38 (0.80)	3.94 (0.67)	36
All scores	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 1.3		F[3,130] = 2.18	F[3,130] = 0.29	F[3,130] = 1.84	F[3,130] = 2.18	

Cells show the mean and standard deviation for the PQA scores at each element score level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Exhibit D.27. PQA Form B Total and Domain Scores by Director Qualifications Element Score and ANOVA Results, Centers

Element Score	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
1	—	—	—	—	0
2	3.82 (0.26)	4.17 (0.33)	3.41 (0.36)	3.74 (0.44)	19
3	3.92 (0.41)	4.13 (0.48)	3.55 (0.54)	3.97 (0.50)	31
4	3.86 (0.43)	4.06 (0.62)	3.61 (0.45)	3.82 (0.47)	42
5	3.72 (0.61)	3.85 (0.71)	3.51 (0.76)	3.76 (0.59)	32
All scores	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 1.03	F[3,120] = 1.75	F[3,120] = 0.59	F[3,120] = 1.22	

Cells show the mean and standard deviation for the PQA scores at each element score level. For ANOVA *F* test results (indicating significant differences across element score levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual element score levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows: [a] differs from level 1; [b] differs from level 2; [c] differs from level 3; [d] differs from level 4; [e] differs from level 5.

Appendix E. Alternative Rating Approach Analysis Results

Cross-Tabulations of California QRIS Ratings and Alternative Rating Approaches

Exhibit E.1. Comparison of California QRIS and Consortia QRIS Ratings, Centers

California QRIS Rating	Consortia QRIS Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	0	19	1	0	0	20
Tier 3	2	0	121	1	0	124
Tier 4	1	0	9	179	0	189
Tier 5	0	0	0	4	28	32
Total	3	19	131	184	28	365

Exhibit E.2. Comparison of California QRIS and Consortia QRIS Ratings, FCCHs

California QRIS Rating	Consortia QRIS Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	1	56	0	0	0	57
Tier 3	0	10	23	1	0	34
Tier 4	0	0	1	10	0	11
Tier 5	0	0	0	1	4	5
Total	1	66	24	12	4	107

Exhibit E.3. Comparison of California QRIS and Two-Level Block Ratings, Centers

California QRIS Rating	Two-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	13	7	0	0	0	20
Tier 3	51	0	73	0	0	124
Tier 4	17	0	0	172	0	189
Tier 5	0	0	0	0	32	32
Total	81	7	73	172	32	365

Exhibit E.4. Comparison of California QRIS and Two-Level Block Ratings, FCCHs

California QRIS Rating	Two-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	51	6	0	0	0	57
Tier 3	13	0	21	0	0	34
Tier 4	3	0	0	8	0	11
Tier 5	0	0	0	0	5	5
Total	67	6	21	8	5	107

Exhibit E.5. Comparison of California QRIS and Three-Level Block Ratings, Centers

California QRIS Rating	Three-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	13	7	0	0	0	20
Tier 3	51	64	9	0	0	124
Tier 4	17	55	0	117	0	189
Tier 5	0	0	0	0	32	32
Total	81	126	9	117	32	365

Exhibit E.6. Comparison of California QRIS and Three-Level Block Ratings, FCCHs

California QRIS Rating	Three-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	51	6	0	0	0	57
Tier 3	13	21	0	0	0	34
Tier 4	3	2	0	6	0	11
Tier 5	0	0	0	0	5	5
Total	67	29	0	6	5	107

Exhibit E.7. Comparison of California QRIS and Five-Level Block Ratings, Centers

California QRIS Rating	Five-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	13	7	0	0	0	20
Tier 3	51	64	9	0	0	124
Tier 4	17	55	106	11	0	189
Tier 5	0	0	10	22	0	32
Total	81	126	125	33	0	365

Exhibit E.8. Comparison of California QRIS and Five-Level Block Ratings, FCCHs

California QRIS Rating	Five-Level Block Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	51	6	0	0	0	57
Tier 3	13	21	0	0	0	34
Tier 4	3	2	6	0	0	11
Tier 5	0	0	1	4	0	5
Total	67	29	7	4	0	107

Exhibit E.9. Comparison of California QRIS and Element Average Ratings, Centers

California QRIS Rating	Element Average Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	0	12	8	0	0	20
Tier 3	0	0	99	25	0	124
Tier 4	0	0	0	189	0	189
Tier 5	0	0	0	0	32	32
Total	0	12	107	214	32	365

Exhibit E.10. Comparison of California QRIS and Element Average Ratings, FCCHs

California QRIS Rating	Element Average Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	1	46	10	0	0	57
Tier 3	0	0	34	0	0	34
Tier 4	0	0	0	11	0	11
Tier 5	0	0	0	4	1	5
Total	1	46	44	15	1	107

Exhibit E.11. Comparison of California QRIS and ERS Hybrid Ratings, Centers

California QRIS Rating	ERS Hybrid Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	0	13	7	0	0	20
Tier 3	0	0	107	17	0	124
Tier 4	0	0	11	178	0	189
Tier 5	0	0	0	4	28	32
Total	0	13	125	199	28	365

Exhibit E.12. Comparison of California QRIS and ERS Hybrid Ratings, FCCHs

California QRIS Rating	ERS Hybrid Rating					
	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Total
Tier 1	0	0	0	0	0	0
Tier 2	0	53	4	0	0	57
Tier 3	0	4	18	12	0	34
Tier 4	0	0	1	10	0	11
Tier 5	0	0	0	4	1	5
Total	0	57	23	26	1	107

Alternative Rating Approach Concurrent Validity Results

Exhibit E.13. CLASS Total and Preschool Domain Scores by Consortia QRIS Rating Level and ANOVA Results, Centers

Consortia QRIS Rating Level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	—	2	—	—	—	2
Tier 2	—	3	—	—	—	2
Tier 3	4.85 (0.51)	55	5.83 (0.49)	5.39 (0.64)	2.95 (0.87)[e]	54
Tier 4	4.94 (0.68)	69	5.97 (0.69)	5.58 (0.75)	3.00 (0.85)[e]	67
Tier 5	5.38 (0.38)	10	6.17 (0.49)	5.78 (0.52)	3.85 (0.62)[c, d]	10
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 2.26		F[4,130] = 1.03	F[4,130] = 1.47	F[4,130] = 2.92*	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit E.14. PQA Form A Total and Preschool Domain Scores by Consortia QRIS Rating Level and ANOVA Results, Centers

Consortia QRIS Rating Level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	—	2	—	—	—	—	—
Tier 2	—	2	—	—	—	—	—
Tier 3	3.43 (0.49)	53	3.55 (0.53)	3.29 (0.60)	3.19 (0.62)[d, e]	4.09 (0.62)	52
Tier 4	3.54 (0.51)	73	3.68 (0.49)	3.27 (0.59)	3.57 (0.74)[c]	4.13 (0.56)	69
Tier 5	3.85 (0.61)	10	3.93 (0.46)	3.47 (0.69)	3.86 (0.75)[c]	4.10 (0.85)	10
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,135] = 1.62		F[4,129] = 1.61	F[4,129] = 0.51	F[4,129] = 3.51**	F[4,129] = 0.09	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.15. PQA Form B Total and Domain Scores by Consortia QRIS Rating Level and ANOVA Results, Centers

Consortia QRIS Rating Level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	—	—	—	—	2
Tier 2	—	—	—	—	2
Tier 3	3.84 (0.39)	4.07 (0.52)	3.53 (0.44)	3.83 (0.44)	49
Tier 4	3.78 (0.50)	3.97 (0.64)	3.48 (0.60)	3.80 (0.56)	63
Tier 5	4.11 (0.47)	4.22 (0.55)	3.93 (0.56)	4.13 (0.44)	8
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 1.03	F[4,119] = 0.75	F[4,119] = 1.22	F[4,119] = 1.20	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.16. CLASS Total and Preschool Domain Scores by Two-Level Block Rating Level and ANOVA Results, Centers

Two-Level Block Rating Level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	4.91 (0.44)	22	5.88 (0.39)	5.33 (0.74)	3.07 (0.76)	20
Tier 2	—	1	—	—	—	1
Tier 3	4.77 (0.51)[e]	38	5.77 (0.48)	5.37 (0.59)	2.86 (0.89)[e]	38
Tier 4	4.93 (0.70)	66	5.97 (0.70)	5.57 (0.74)	3.00 (0.87)[e]	64
Tier 5	5.39 (0.34)[c]	12	6.23 (0.50)	5.88 (0.54)	3.74 (0.70)[c, d]	12
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 3.12*		F[4,130] = 1.60	F[4,130] = 1.76	F[4,130] = 2.75*	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit E.17. PQA Form A Total and Preschool Domain Scores by Two-Level Block Rating Level and ANOVA Results, Centers

Two-Level Block Rating Level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	3.42 (0.53)	21	3.63 (0.5)	3.20 (0.71)	3.11 (0.68)[e]	4.00 (0.78)	19
Tier 2	—	1	—	—	—	—	1
Tier 3	3.42 (0.47)	36	3.52 (0.53)	3.29 (0.55)	3.20 (0.57)[e]	4.12 (0.56)	36
Tier 4	3.55 (0.51)	70	3.66 (0.50)	3.29 (0.60)	3.58 (0.75)	4.17 (0.51)	66
Tier 5	3.81 (0.59)	12	3.95 (0.43)	3.43 (0.65)	3.85 (0.71)[a, c]	3.93 (0.91)	12
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,135] = 1.61		F[4,129] = 1.73	F[4,129] = 0.48	F[4,129] = 4.06**	F[4,129] = 0.64	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.18. PQA Form B Total and Domain Scores by Two-Level Block Rating Level and ANOVA Results, Centers

Two-Level Block Rating Level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	3.80 (0.40)	3.96 (0.58)	3.58 (0.51)	3.78 (0.43)	20
Tier 2	—	—	—	—	1
Tier 3	3.86 (0.38)	4.14 (0.47)	3.49 (0.44)	3.84 (0.45)	32
Tier 4	3.81 (0.49)	4.00 (0.61)	3.51 (0.62)	3.82 (0.55)	61
Tier 5	3.98 (0.64)	4.11 (0.80)	3.80 (0.58)	4.00 (0.64)	10
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 0.45	F[4,119] = 0.42	F[4,119] = 0.82	F[4,119] = 0.61	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.19. CLASS Total and Preschool Domain Scores by Three-Level Block Rating Level and ANOVA Results, Centers

Three-Level Block Rating Level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	4.91 (0.44)	22	5.88 (0.39)	5.33 (0.74)	3.07 (0.76)	20
Tier 2	4.87 (0.60)	64	5.87 (0.61)	5.52 (0.69)	2.91 (0.88)[e]	64
Tier 3	—	1	—	—	—	1
Tier 4	4.87 (0.72)	40	5.93 (0.68)	5.47 (0.70)	3.00 (0.89)	38
Tier 5	5.39 (0.34)	12	6.23 (0.50)	5.88 (0.54)	3.74 (0.70)[b]	12
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[4,134] = 2.13		F[4,130] = 0.95	F[4,130] = 1.32	F[4,130] = 2.52*	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit E.20. PQA Form A Total and Preschool Domain Scores by Three-Level Block Rating Level and ANOVA Results, Centers

Three-Level Block Rating Level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	3.42 (0.53)	21	3.63 (0.50)	3.20 (0.71)	3.11 (0.68)[d, e]	4.00 (0.78)	19
Tier 2	3.44 (0.50)	63	3.56 (0.48)	3.25 (0.60)	3.27 (0.69)[d]	4.19 (0.55)	62
Tier 3	—	1	—	—	—	—	1
Tier 4	3.59 (0.49)	43	3.68 (0.55)	3.34 (0.56)	3.74 (0.66)[a, b]	4.09 (0.48)	40
Tier 5	3.81 (0.59)	12	3.95 (0.43)	3.43 (0.65)	3.85 (0.71)[a]	3.93 (0.91)	12
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[4,135] = 1.80		F[4,129] = 1.77	F[4,129] = 0.41	F[4,129] = 5.11***	F[4,129] = 0.89	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.21. PQA Form B Total and Domain Scores by Three-Level Block Rating Level and ANOVA Results, Centers

Three-Level Block Rating Level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	3.80 (0.40)	3.96 (0.58)	3.58 (0.51)	3.78 (0.43)	20
Tier 2	3.77 (0.47)	3.99 (0.58)	3.37 (0.53)	3.84 (0.50)	56
Tier 3	—	—	—	—	2
Tier 4	3.88 (0.42)	4.11 (0.52)	3.67 (0.55)	3.78 (0.54)	36
Tier 5	3.98 (0.64)	4.11 (0.80)	3.80 (0.58)	4.00 (0.64)	10
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[4,119] = 1.36	F[4,119] = 0.89	F[4,119] = 2.85*	F[4,119] = 0.82	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.22. CLASS Total and Preschool Domain Scores by Five-Level Block Rating Level and ANOVA Results, Centers

Five-Level Block Rating Level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	4.91 (0.44)	22	5.88 (0.39)	5.33 (0.74)	3.07 (0.76)	20
Tier 2	4.87 (0.60)	64	5.87 (0.61)	5.52 (0.68)	2.91 (0.88)	64
Tier 3	4.90 (0.70)	43	5.95 (0.68)	5.48 (0.68)	3.04 (0.85)	41
Tier 4	5.36 (0.42)	10	6.22 (0.43)	5.88 (0.58)	3.69 (0.91)	10
Tier 5	—	0	—	—	—	0
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 1.95		F[3,131] = 1.06	F[3,131] = 1.44	F[3,131] = 2.45	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Exhibit E.23. PQA Form A Total and Preschool Domain Scores by Five-Level Block Rating Level and ANOVA Results, Centers

Five-Level Block Rating Level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	3.42 (0.53)[d]	21	3.63 (0.50)	3.20 (0.71)	3.11 (0.68)[c, d]	4.00 (0.78)	19
Tier 2	3.44 (0.50)[d]	63	3.56 (0.48)[d]	3.25 (0.60)	3.27 (0.69)[d]	4.19 (0.55)	62
Tier 3	3.56 (0.52)	45	3.68 (0.54)	3.29 (0.58)	3.69 (0.68)[a]	4.02 (0.58)	43
Tier 4	3.95 (0.33)[a, b]	11	4.04 (0.37)[b]	3.63 (0.47)	4.06 (0.45)[a, b]	4.26 (0.67)	10
Tier 5	—	0	—	—	—	—	0
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 3.61*		F[3,130] = 2.85*	F[3,130] = 1.28	F[3,130] = 7.69***	F[3,130] = 1.17	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Exhibit E.24. PQA Form B Total and Domain Scores by Five-Level Block Rating Level and ANOVA Results, Centers

Five-Level Block Rating level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	3.80 (0.40)	3.96 (0.58)	3.58 (0.51)	3.78 (0.43)	20
Tier 2	3.77 (0.47)	3.99 (0.58)	3.37 (0.53)[d]	3.84 (0.50)	56
Tier 3	3.88 (0.46)	4.09 (0.57)	3.65 (0.57)	3.79 (0.57)	38
Tier 4	4.11 (0.48)	4.29 (0.64)	3.91 (0.46)[b]	4.04 (0.48)	10
Tier 5	—	—	—	—	0
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 1.77	F[3,120] = 0.94	F[3,120] = 4.11**	F[3,120] = 0.72	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Exhibit E.25. CLASS Total and Preschool Domain Scores by Element Average Rating level and ANOVA Results, Centers

Element Average Rating level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	—	0	—	—	—	0
Tier 2	—	1	—	—	—	0
Tier 3	4.76 (0.48)[e]	46	5.77 (0.41)[e]	5.32 (0.59)[e]	2.84 (0.83)[e]	46
Tier 4	4.94 (0.67)	80	5.96 (0.68)	5.56 (0.74)	3.03 (0.86)[e]	77
Tier 5	5.39 (0.34)[c]	12	6.23 (0.50)[c]	5.88 (0.54)[c]	3.74 (0.70)[c, d]	12
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 3.81*		F[2,132] = 3.32*	F[2,132] = 3.88*	F[2,132] = 5.38**	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit E.26. PQA Form A Total and Preschool Domain Scores by Element Average Rating level and ANOVA Results, Centers

Element Average Rating level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	—	0	—	—	—	—	0
Tier 2	—	1	—	—	—	—	0
Tier 3	3.39 (0.50)	41	3.52 (0.54)[e]	3.24 (0.59)	3.14 (0.60)[d, e]	4.12 (0.58)	41
Tier 4	3.53 (0.50)	86	3.66 (0.49)	3.29 (0.60)	3.53 (0.74)[c]	4.14 (0.56)	81
Tier 5	3.81 (0.59)	12	3.95 (0.43)[c]	3.43 (0.65)	3.85 (0.71)[c]	3.93 (0.91)	12
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 2.28		F[2,131] = 3.59*	F[2,131] = 0.46	F[2,131] = 6.77**	F[2,131] = 0.61	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.27. PQA Form B Total and Domain Scores by Element Average Rating level and ANOVA Results, Centers

Element Average Rating level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	—	—	—	—	0
Tier 2	—	—	—	—	1
Tier 3	3.77 (0.34)	4.05 (0.46)	3.41 (0.43)	3.73 (0.42)	39
Tier 4	3.84 (0.49)	4.02 (0.61)	3.56 (0.60)	3.86 (0.53)	74
Tier 5	3.98 (0.64)	4.11 (0.80)	3.80 (0.58)	4.00 (0.64)	10
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 0.68	F[3,120] = 0.33	F[3,120] = 1.76	F[3,120] = 1.16	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating level): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating level, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.28. CLASS Total and Preschool Domain Scores by ERS Hybrid Rating level and ANOVA Results, Centers

ERS Hybrid Rating level	Preschool and Toddler		Preschool Domain Scores			
	CLASS Total Score	*N*	Emotional Support	Classroom Organization	Instructional Support	*N*
Tier 1	—	0	—	—	—	0
Tier 2	—	2	—	—	—	1
Tier 3	4.80 (0.50)[e]	53	5.82 (0.45)	5.39 (0.64)	2.87 (0.82)[e]	53
Tier 4	4.92 (0.69)	72	5.95 (0.70)	5.54 (0.73)	3.02 (0.89)[e]	69
Tier 5	5.39 (0.34)[c]	12	6.23 (0.50)	5.88 (0.54)	3.74 (0.70)[c, d]	12
All tiers	4.92 (0.61)	139	5.92 (0.60)	5.51 (0.69)	3.03 (0.87)	135
ANOVA results	F[3,135] = 3.24*		F[3,131] = 1.68	F[3,131] = 1.78	F[3,131] = 3.47*	

Cells show the mean and standard deviation for the CLASS scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating level): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tiert3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average CLASS score data are not presented for rating levels with fewer than five observations.

Exhibit E.29. PQA Form A Total and Preschool Domain Scores by ERS Hybrid Rating level and ANOVA Results, Centers

ERS Hybrid Rating level	All Ages		Preschool Domain Scores				
	PQA Form A Total Score	*N*	Learning Environment	Daily Routine	Adult-Child Interaction	Curriculum Planning and Assessment	*N*
Tier 1	—	0	—	—	—	—	0
Tier 2	—	1	—	—	—	—	0
Tier 3	3.45 (0.50)	48	3.56 (0.52)	3.30 (0.61)	3.22 (0.61)[e]	4.15 (0.57)	48
Tier 4	3.51 (0.51)	79	3.64 (0.50)	3.26 (0.60)	3.52 (0.76)	4.12 (0.57)	74
Tier 5	3.81 (0.59)	12	3.95 (0.43)	3.43 (0.65)	3.85 (0.71)[c]	3.93 (0.91)	12
All tiers	3.52 (0.52)	140	3.64 (0.51)	3.29 (0.60)	3.44 (0.72)	4.11 (0.60)	134
ANOVA results	F[3,136] = 1.65		F[2,131] = 2.9	F[2,131] = 0.43	F[2,131] = 4.84**	F[2,131] = 0.64	

Cells show the mean and standard deviation for the PQA scores at each rating level. The preschool domain scores have a smaller *N* because some participating centers did not have any preschool classrooms. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.

Exhibit E.30. PQA Form B Total and Domain Scores by ERS Hybrid Rating level and ANOVA Results, Centers

ERS Hybrid Rating level	All Ages				
	PQA Form B Total Score	Parent Involvement and Family Services	Staff Qualifications and Staff Development	Program Management	*N*
Tier 1	—	—	—	—	0
Tier 2	—	4.56 (—)	4.00 (—)	3.43 (—)	1
Tier 3	3.80 (0.36)	4.07 (0.45)	3.45 (0.47)	3.78 (0.44)	44
Tier 4	3.83 (0.49)	4.00 (0.63)	3.55 (0.6)	3.85 (0.54)	69
Tier 5	3.98 (0.64)	4.11 (0.8)	3.80 (0.58)	4.00 (0.64)	10
All tiers	3.83 (0.46)	4.04 (0.58)	3.54 (0.56)	3.83 (0.51)	124
ANOVA results	F[3,120] = 0.48	F[3,120] = 0.44	F[3,120] = 1.38	F[3,120] = 0.75	

Cells show the mean and standard deviation for the PQA scores at each rating level. For ANOVA *F* test results (indicating significant differences across rating levels): * $p < .05$; ** $p < .01$; *** $p < .001$. When ANOVA is significant, significant ($p < .05$) differences between individual rating levels, after Tukey-Kramer adjustment for multiple comparisons, are indicated as follows:

[a] differs from Tier 1; [b] differs from Tier 2; [c] differs from Tier 3; [d] differs from Tier 4; [e] differs from Tier 5.

Note that average PQA score data are not presented for rating levels with fewer than five observations.