This product is part of the RAND Corporation occasional paper series. RAND occasional papers may include an informed perspective on a timely policy issue, a discussion of new research methodologies, essays, a paper presented at a conference, a conference summary, or a summary of work in progress. All RAND occasional papers undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

OCCASIONAL PAPER

Moving to Outcomes

Approaches to Incorporating Child Assessments into State Early Childhood Quality Rating and Improvement Systems

Gail L. Zellman • Lynn A. Karoly

Sponsored by the David and Lucile Packard Foundation

EDUCATION and
RAND LABOR AND POPULATION

The research described in this report was conducted jointly by RAND Education and RAND Labor and Population, units of the RAND Corporation. Funding was provided by the David and Lucile Packard Foundation.

Published 2012 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

Research findings point to the importance of the period from birth to school entry for children's development and focus attention on the quality of the early care and education (ECE) experiences young children receive. Numerous studies have demonstrated that higher-quality care, defined in various ways, predicts positive developmental gains for children. However, the ECE experienced by many children is not of sufficiently high quality to achieve the potential developmental benefits, and some care may even be harmful.

In recent years, quality rating and improvement systems (QRISs)—systems that incorporate ratings based on multicomponent assessments designed to make ECE quality transparent and easily understood and that also provide feedback, technical assistance, and incentives based on those ratings to both motivate and support quality improvement—have become an increasingly popular policy tool to improve quality in ECE settings and have been adopted in many localities and states. The ultimate goal of QRISs is to raise the quality of care provided in ECE settings, which in turn is expected to improve child functioning. Yet although improved child outcomes are the ultimate goal, QRISs rarely directly assess children as a way to determine if the system is improving child outcomes. This is because it is costly to accurately measure child functioning and difficult to identify the contribution of any given ECE setting to a particular child's developmental trajectory. Despite these challenges, it is important that QRISs incorporate child assessments to at least some extent, because they can help to improve practice and do represent the ultimate goal of these systems. The purpose of this paper is to identify options for states to consider for incorporating child assessments into the design, implementation, and evaluation of their QRISs or other quality improvement (QI) efforts.

The work reported in this paper was sponsored by the David and Lucile Packard Foundation as part of its support for RAND's assistance to the State of California's efforts to develop, pilot, implement, and evaluate a QRIS. Although the paper was motivated by the agenda of California's Early Learning Advisory Council and we provide examples from California where relevant, the subject matter, analysis, and guidance are equally relevant for other states seeking to improve the quality of their child care and early learning programs. Thus, the paper should be of interest to policymakers, advocates, practitioners, and researchers seeking to identify the merits and drawbacks of alternative strategies for incorporating child assessments into state QRISs and other ECE quality improvement efforts.

This research was conducted jointly by RAND Education and RAND Labor and Population, units of the RAND Corporation. For inquiries related to RAND Education, please contact Darleen Opfer, Director, RAND Education, at Darleen_Opfer@rand.org. For inquiries related to RAND Labor and Population, please contact Arie Kapteyn, Director, RAND Labor and Population, at Arie_Kapteyn@rand.org.

Contents

Preface iii
Figure and Tables vii
Summary ix
Acknowledgments xix
Abbreviations xxi

CHAPTER ONE
Introduction 1
Defining Key Terms 2
Road Map for the Paper 3

CHAPTER TWO
The Ultimate Goal of State QRISs: Improving Child Developmental Outcomes 5
Motivation for State QRISs 6
Quality Shortfalls in Existing ECE Programs 6
Existing ECE Systems Do Not Ensure High Quality 7
Features of ECE Markets Limit Use of High-Quality Services 8
The Logic of QRISs 9
A Brief History of State QRISs 11
The QRIS Landscape 11
QRIS Design 13
The Role of Child Assessments in QRISs 14
Challenges in Assessing Young Children and Using Assessments 16
Assessment Issues 16
Assessment Objectives 18

CHAPTER THREE
Approaches to Using Assessments of Child Functioning in State ECE QI Efforts 21
A Framework for Classifying Approaches to Using Assessments of Child Functioning 21
Approach A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice 25
Current Practice 25
Resources Required 29
Expected Benefits 29
Potential Barriers to Success and Strategies for Mitigation 30
Approach B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice 30
Current Practice 30

Resources Required 31
Expected Benefits 32
Potential Barriers to Success and Strategies for Mitigation 32
Approach C: Independent Measurement of Child Outcomes to Assess Programs 32
Current Practice 34
Resources Required 36
Expected Benefits 36
Potential Barriers to Success and Strategies for Mitigation 37
Approach D: Independent Measurement of Child Outcomes to Assess QRIS Validity 37
Current Practice 38
Resources Required 40
Expected Benefits 41
Potential Barriers to Success and Strategies for Mitigation 41
Approach E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System 42
Current Practice 42
Resources Required 45
Expected Benefits 45
Potential Barriers to Success and Strategies for Mitigation 45

CHAPTER FOUR
Conclusions and Policy Guidance 47
Suggestion: Implement Either Approach A or Approach B, Depending on Whether a QRIS Exists 47
Suggestion: Undertake Approach D When Piloting a QRIS and Periodically Once the QRIS Is Implemented at Scale 48
Suggestion: Implement Approach E Periodically Regardless of Whether a QRIS Exists 49
Suggestion: If Approach C Is Under Consideration for Inclusion in a QRIS, Proceed with Caution 49

Bibliography 51

Figure and Tables

Figure

2.1. A Logic Model for QRISs 12

Tables

S.1. Five Approaches to Incorporating Assessments of Child Functioning into State QI Efforts xiii
S.2. Guidance for Incorporating Child Assessments into State QI Efforts xv
3.1. Five Approaches to Incorporating Assessments of Child Functioning into State QI Efforts 22
3.2. Measurement Details and Analysis Methods for Each Approach to Incorporating Child Assessments 24
3.3. Additional Features of Each Approach to Incorporating Child Assessments 26
3.4. Estimated Effects of State Preschool Programs on School Readiness Using Quasi-Experimental Designs 44
4.1. Guidance for Incorporating Child Assessments into State QI Efforts 48

Summary

In recent years, quality rating and improvement systems (QRISs) have become an increasingly popular policy tool to improve quality in early care and education (ECE) settings and have been adopted in many localities and states. QRISs incorporate ratings based on multicomponent assessments designed to make the quality of early care and education programs transparent and easily understood. Most also include feedback and technical assistance and offer incentives to both motivate and support quality improvement. The ultimate goal of QRISs is to raise the quality of care provided in ECE settings; these higher-quality settings are expected to improve child functioning across a range of domains, including school readiness. QRIS logic models focus on one set of inputs to child development—various dimensions of ECE quality—with the goal of improving system outcomes, namely, child cognitive, social, emotional, and physical development.

Yet although improved child outcomes are the ultimate goal, QRISs rarely directly assess children's developmental outcomes to determine if the system itself is improving child functioning, nor do they require child assessments for the purpose of evaluating specific programs. This is largely because it is costly to accurately measure child functioning and difficult to identify the contribution of any given ECE setting to a particular child's developmental trajectory. Despite these challenges, it is important that QRISs incorporate child assessments to at least some extent, because they can help to improve practice.

The purpose of this paper is to identify options for states to consider for incorporating child assessments into the design, implementation, and evaluation of their QRISs or other quality improvement efforts. Our analysis draws on decades of research regarding the measurement of child development and the methods available for measuring the contribution of child care and early learning settings to children's developmental trajectories. We also reference new research documenting the approaches taken in other states to include measures of child development in their QRISs and lessons learned from those experiences.

In this summary, we briefly review the motivation for QRISs and highlight some of the key challenges encountered in assessing young children and using assessment data. We then present five approaches for incorporating child assessments into state ECE quality improvement (QI) efforts. The approaches differ in terms of purpose, who conducts the assessment, and the sort of design needed to ensure that the resulting child assessment data can be used in a meaningful way. We conclude by offering guidance regarding the use of the five strategies based on our assessment of the overall strengths and weaknesses and the potential benefit relative to the cost of each approach.

The Ultimate Goal of State QRISs Is Improving Child Functioning

Research findings point to the importance of the period from birth to school entry for children's development and demonstrate that higher-quality care, defined in various ways, predicts positive developmental gains for children. Recent work has attempted to better understand how quality operates to improve child outcomes by deconstructing quality and focusing on the importance of dosage, thresholds, and quality features in promoting improved child outcomes. However, the ECE experienced by many children is not of sufficiently high quality to achieve the potential developmental benefits, and some care may even be harmful. Despite the evidence pointing to the need for improved ECE quality, there has been little policy response until the last decade. Three factors have propelled the development and implementation of QRISs in recent years:

- **Continuing gaps in quality in existing ECE programs.** Despite the evidence showing the benefits of high-quality care, the ECE experienced by many children do not meet quality benchmarks, often falling far short of even "good" care. Concerns about poor-quality care have been exacerbated by a policy focus in recent years on students' academic achievement. In particular, the K–12 accountability provisions in the No Child Left Behind (NCLB) Act of 2001 (Public Law [P. L.] 107-110) have led K–12 leaders to focus on the limited skills that many children bring to kindergarten. They argue that K–12 actors should not be expected to meet rigorous standards for students' progress in elementary school when so many enter kindergarten unprepared to learn.
- **The inability of the current ECE system to promote uniformly high quality**. Although much care is licensed, licensing represents a fairly low standard for quality, focused as it is on the adequacy and safety of the physical environment. In recent years, in response to fiscal constraints, even these minimal requirements are less likely to be monitored. Some publicly funded programs must adhere to higher quality standards, but for many providers, there is little pressure to focus on quality.
- **Features of the market for ECE that limit the consumption of high-quality services.** Research finds that parents are not very good at evaluating the quality of care settings, consistently rating program quality far higher than trained assessors do. In addition, the limited availability of care in many locations and for key age groups (particularly infants) provides ready clients for most providers, even those who do not offer high-quality services. The high cost of quality care and limited public funding to subsidize the cost of ECE programs for low-income families further constrain the demand for high-quality care.

Given these issues, policymakers and the public have turned to QRISs as a mechanism to improve ECE quality, starting with the first system launched in 1998 in Oklahoma. QRISs are essentially accountability systems centered around quality ratings that are designed to improve ECE quality by defining quality standards, making program quality transparent, and providing supports for quality improvement. Although consistent with accountability efforts in K–12 education, QRISs differ in a key way in their almost exclusive focus on inputs into caregiving and caregiving processes rather than on outcomes of the process, which for K–12 accountability systems are measures of student performance on standardized assessments. QRISs have proved popular with state legislatures in recent years because they represent a conceptually

straightforward way to improve quality that appeals both to child advocates—because of the promise of support for improvements—and to those who support market-based solutions—because QRISs incentivize improvement. Indeed, the number of states that are implementing some form of rating system, including system pilots, has increased from 14 in early 2006 to 35 as of early 2011.

There are, of course, good reasons why QRISs focus on the input side of the logic model: The use of child assessments to improve programs or assess how well QRISs are working presents many challenges, including young children's limited attention spans, uneven skills development, and discomfort with strangers and strange situations. One effect of these challenges is that reliability (i.e., consistent measurement) is more difficult to achieve. Validity is also an issue; validity is attached not to measures but to the use of a specific instrument in a specific context. Often, assessments used in QRISs were designed for use in low-stakes settings such as research studies and program self-assessments. But QRISs increasingly represent high-stakes settings, where the outcomes of assessments affect public ratings, reimbursement rates, and the availability of technical assistance.

The choice of which child assessment tool to use depends on the purpose of the assessments and the way in which the resulting data are to be used. Child assessments may be formal or informal and may take a number of forms, including standardized assessments, home inventories, portfolios, running records, and observation in the course of children's regular activities. They are generally understood to have three basic purposes: screening individual children for possible handicapping conditions, supporting and improving teaching and learning, and evaluating interventions. Because screening individual children for handicapping conditions is not a program-related issue, we do not discuss screening in detail. Assessments for improving practice are designed to determine how well children are learning so that interactions with children, curricula, and other interventions can be modified to better meet children's learning needs, at the levels of the individual child, the classroom, and the program. These assessments may be formal or informal. Key to these assessments is the establishment of a plan for using the data that are collected to actually improve programs and interventions. Assessments used for evaluation must meet a higher standard: They should be imbedded in a rigorous research design that increases the likelihood of finding effects, if they exist, to the greatest extent possible. In selecting instruments to use, it also is critical to select tools and use them in ways that meet the guidelines for reliability and validity.

Given these assessment challenges, QRIS designs consistently have focused on measuring inputs to quality rather than outputs such as children's level of school readiness, literacy, or numeracy or noncognitive skills such as self-regulation or the ability to follow instructions or get along with peers. This input focus was considered a necessary concession to the reality that the performance and longer-term outcomes of young children are difficult and costly to measure and that measures of these attributes are less reliable and less accurate than those for older children. Yet advocates understood that the ultimate goal of these systems was to improve children's functioning through the provision of higher-quality ECE programs.

There Are Multiple Approaches for Incorporating Child Assessments into State QI Efforts

As QRISs have developed and been refined over time, assessments of child developmental outcomes have increasingly found their way into QRISs, although they generally are designed to improve inputs to care by clarifying children's progress in developing key skills.[1] Efforts to use child assessments as outcomes that contribute to a determination about how well a QRIS is working are relatively rare. To frame our discussion, we define five strategies for using assessments of child functioning to improve ECE quality, three of which are predicated on the existence of a QRIS.

Table S.1 summarizes the purpose of each approach and its relationship to a QRIS. The strategies are arrayed in Table S.1, from those that focus on assessments of child functioning at the micro level—the developmental progress of an individual child or group of children in a classroom—to those that have a macro focus—the performance of the QRIS at the state level or the effect of a specific ECE program or the larger ECE system on children's growth trajectories at the state level. Given the different purposes of these assessments, the assessment tools used and the technical requirements involved in the process are likely to be quite different. Our review of each strategy considers current use in state systems, lessons learned from prior experience, the resources required for implementing the strategy, the benefits of the approach, and possible barriers to success and strategies for mitigation of these barriers. In brief, the five approaches are as follows:

- With Approach A, labeled **Caregiver/Teacher- or Program-Driven Assessments to Improve Practice**, individual caregivers or teachers are trained as part of their formal education or ongoing professional development to use developmentally appropriate assessments to evaluate each child in his or her care. Program leadership may aggregate the assessment results to the classroom or program level to improve practice and identify needs for professional development or other quality enhancements. This approach does not assume a formal link to a QRIS but rather that the use of child assessments is part of standard practice as taught in teacher preparation programs or other professional development programs and as reinforced through provider supervision. The practice of using child assessments is currently endorsed in the National Association for the Education of Young Children (NAEYC) accreditation standards for ECE programs and postsecondary ECE teacher preparation programs and included in some ECE program regulations (e.g., Head Start and California Title 5 programs). Data from California suggest that most center-based teachers rely on some form of child assessments to inform their work with children. Expected benefits include the enhanced ability of caregivers and teachers to provide individualized support to the children in their group, the early detection of developmental delays, better-informed parents who engage in developmentally supportive at-home activities, and data to inform staff development and program improvement. To be effective, caregivers and teachers must be well trained in the use of child assessments and

[1] As noted above, we focus on the use of child assessments for purposes of supporting and improving teaching and learning and for evaluating interventions. Thus, we do not focus on their use as a tool to screen for developmental delays or other handicapping conditions, although some rating systems consider whether assessments are used for screening purposes in measuring program quality.

Table S.1
Five Approaches to Incorporating Assessments of Child Functioning into State QI Efforts

Approach	Description and Purpose	Focus	Relationship to QRIS
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Expectation of use of child assessments by caregivers/ teachers to inform caregiving and instructional practice with individual children and to identify needs for staff professional development and other program quality enhancements	**Individual child** Assess developmental progress Apply differentiated instruction **Classroom/group or program** Identify areas for improved practice Determine guidance for technical assistance	Not explicitly incorporated into QRIS; can be focus of best practice in teacher preparation programs, ongoing professional development, and provider supervision; can be a requirement of licensing, program regulation, or accreditation
B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice	QRIS requires demonstrated use of child assessments by caregivers/teachers to inform caregiving and instructional practice with individual children and to identify needs for staff professional development and other program quality enhancements	Same as Approach A	QRIS rating element specifically assesses this component alone or in combination with other related practice elements
C: Independent Measurement of Child Outcomes to Assess Programs	Independent assessors measure changes in child functioning at the classroom/group or program level to assess program effects on child development or to assess the effectiveness of technical assistance or other interventions	**Classroom/group or program** Estimate value added Assess technical assistance effectiveness or other interventions	QRIS rating element is based on estimates of effects at the classroom/ group or program level
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	Independent assessors measure changes in child functioning to validate QRIS design (i.e., to determine if higher QRIS ratings are associated with better child developmental outcomes)	**Statewide QRIS** Assess validity of the rating portion of QRIS	Part of (one-time or periodic) QRIS evaluation
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Independent assessors measure child functioning to evaluate causal effects of specific ECE programs or groups of programs on child developmental outcomes at the state level	**Statewide ECE system or specific programs in system** Estimate causal effects of ECE programs	Part of QRIS (ongoing) quality assurance processes or, when no QRIS exists, part of evaluation of state ECE system

SOURCE: Authors' analysis.

in communicating results to parents. Program administrators also need to be able to use the assessment results to identify needs for staff development and program improvement.

- Approach B, labeled **QRIS-Required Caregiver/Teacher Assessments to Improve Practice**, has the same purpose as Approach A, but it has an explicit link to a QRIS. In this approach, a QRIS rating element *requires* the demonstrated use of assessments of child functioning to inform the approach a caregiver or teacher takes with an individual child, as well as efforts to improve program quality through professional development, technical assistance, or other strategies. Eleven of the 26 QRISs recently catalogued incorporate an indicator regarding the use of child assessments into the rating criteria for center-based

programs, whereas eight systems included such an indicator in its rating criteria for family child care homes. However, most systems do not include the use of assessments in their rating criteria for the lower tiers of their rating systems. The expected benefits are similar to those in Approach A, although the tie to the QRIS may increase compliance with the practice; caregivers and teachers may also be more effective in their use of assessments if the QRIS emphasizes the quality of implementation.

- For Approach C, labeled **Independent Measurement of Child Outcomes to Assess Programs**, the link between the QRIS and child developmental outcomes is even more explicit. In this case, the measurement of *changes* over time in child functioning at the classroom, group, or center level can be either an additional quality element incorporated into the rating system or a supplement to the information summarized in the QRIS rating. The appeal of this approach is that instead of relying solely on measured inputs to capture ECE program quality and calculate ratings, there is the potential to capture the outcome of interest—ECE program effects on child functioning—and to use the results when rating programs. At the same time, use of such data from three- and four-year-olds to hold individuals (here, caregivers or teachers) accountable has been deemed inappropriate because of reliability and validity concerns when assessing young children. Although this approach has not been used in QRISs to date, it is used in K–12 education, often as part of high-stakes accountability systems. In particular, value-added modeling (VAM) is a method that has quickly gained favor in the K–12 context for isolating the contributions of teachers or schools to student performance. Although VAM has many supporters, it remains controversial because of numerous methodological issues that have yet to be resolved, including the sensitivity of value-added measures to various controls for student characteristics and classroom peers and the reliability of value-added measures over time—issues that would likely be compounded with other issues unique to the ECE context. Since individual children in ECE programs would need to be assessed by independent assessors, it is also very resource-intensive.
- Approach D, labeled **Independent Measurement of Child Outcomes to Assess QRIS Validity**, collects child assessment data to address macro-level questions, in this case, the validity of the rating portion of the QRIS. For QRISs, the logic model asserts that higher-quality care will be associated with better child outcomes. Therefore, one important piece of validation evidence concerns whether higher program ratings, which are largely based on program inputs, are positively correlated with better child performance, the ultimate QRIS outcome. The required methods for this approach are complex and subject to various threats to validity, but there are strategies to minimize those concerns such as ensuring sufficient funding for the required sample sizes and the collection of relevant child and family background characteristics. The ability to base the QRIS validation design on a sample of programs and children means that it can be a cost-effective investment in the quality of the QRIS. To date, two states (Colorado and Missouri) have conducted such validation studies with mixed findings, and three other states (Indiana, Minnesota, and Virginia) have plans to implement this approach.
- Approach E, labeled **Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System**, also takes a macro perspective, but it differs from Approach D in using rigorous methods that enable an assessment of the causal effects of a statewide ECE program or group of programs on child developmental outcomes. To date, eight states have used a regression discontinuity design (a quasi-

experimental method that is appropriate when an ECE program has a strict age-of-entry requirement) to measure the effect of participating one year in their state preschool program on cognitive measures of school readiness. These evaluations have been conducted without reference to any statewide QRIS, but an evaluation using an experimental design or a quasi-experimental method could be a required QRIS component for determining at one point in time or on an ongoing basis if an ECE program or the ECE system as a whole is achieving its objectives of promoting strong child growth across a range of developmental domains. As in Approach D, this type of evaluation can be implemented with a sample of children and therefore is also a cost-effective way to bring accountability to ECE programs.

Policymakers Should Employ a Combination of Approaches

Our analysis of each of the five approaches, leads us to offer the guidance summarized in Table S.2 regarding the use of each of the strategies.

Table S.2
Guidance for Incorporating Child Assessments into State QI Efforts

Approach	Guidance	Rationale
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Implement either Approach A or Approach B, depending on whether a state-level QRIS has been implemented: If no QRIS exists, adopt Approach A; consider reinforcing through licensing, regulation, or accreditation if not already part of these mechanisms If a QRIS exists, adopt Approach B	Consistent with good ECE practice Important potential benefits in terms of practice and program improvement for relatively low incremental cost
B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice		Greater likelihood of use and appropriate use of assessments than with Approach A Important potential benefits in terms of practice and program improvement for relatively low incremental cost
C: Independent Measurement of Child Outcomes to Assess Programs	If considering adopting this approach as part of a QRIS, proceed with caution	Methodology is complex and not sufficiently developed for high-stakes use Costly to implement for uncertain gain Feasibility and value for cost could be tested on a pilot basis
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	Implement this approach when piloting a QRIS and periodically once the QRIS is implemented at scale (especially following major QRIS revisions)	Important to assess validity of the QRIS at the pilot stage and to reevaluate validity as the system matures Methodology Is complex but periodic implementation means high return on investment
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Implement this approach periodically (e.g., on a routine schedule or following major policy changes) regardless of whether a QRIS exists	Evidence of system effects can justify spending and guide quality improvement efforts Methodology is complex, but periodic implementation means high return on investment

SOURCE: Authors' analysis.

Promote the use of child assessments by ECE caregivers and teachers to improve practice either as part of a QRIS (Approach B) or through other mechanisms (Approach A). We suggest that all teachers and programs collect the child assessment data prescribed by Approaches A and B and that programs or states implement one or the other approach depending on whether the state has a QRIS. Key to effective use of both approaches is the provision of professional development that helps staff identify which measures are most appropriate for which purposes and teaches them how to use data from their assessments to improve practice. Our guidance stems from recognition that it is good practice for caregivers and teachers to use child assessments to shape their interactions with individual children in the classroom and to identify areas for program improvement; this approach is also endorsed by the NAEYC in its standards for accrediting ECE programs and postsecondary ECE teacher preparation programs. The use of child assessments in this manner has the potential to promote more effective individualized care and instruction on the part of caregivers and teachers and to provide program administrators with important information to guide professional development efforts and other quality improvement initiatives. The potential for widespread benefits from effective use of child assessments can be weighed against what we expect would be a relatively small incremental cost given the already widespread use of assessments, although costs would be higher if current practice does not include the needed professional development supports to ensure that assessments are used effectively to improve teaching and learning.

Undertake a QRIS validation study (Approach D) when piloting the implementation of a QRIS and repeat it periodically once the QRIS is implemented at scale. By validating the quality rating portion of a QRIS, Approach D can be a cost-effective investment in a state's QI efforts. We suggest that this approach be employed in the implementation pilot phase of a QRIS, assuming that there is such a phase, as that phase represents an opportune time in which to identify any weaknesses in the ability of a QRIS to measure meaningful differences in ECE program quality that matter for child outcomes. Incorporating a QRIS validation component into a pilot phase will ensure that needed refinements to the QRIS can be introduced before taking the system to scale. This will reduce the need to make changes in the QRIS structure once it is fully implemented. We further suggest that a QRIS validation study be repeated periodically (e.g., every five to ten years) or following major changes to a QRIS. This will ensure the continuing relevance of the QRIS given changes in the population of children served by ECE programs, the nature of ECE programs themselves, and other developments in the ECE field.

Implement a statewide, periodic evaluation of specific ECE programs or the broader ECE system (Approach E) regardless of whether a QRIS exists. Child assessments can be a critical addition to evaluation efforts that examine a range of program attributes. By using available cost-effective quasi-experimental methods, evaluators can determine if an ECE program (or the ECE system as a whole) is achieving its objectives of promoting strong child development across a range of domains. Approach E, especially when applied to ECE programs supported with public funding, fulfills a need for accountability, as part of either a QRIS or other state QI efforts. Favorable findings can be used to justify current spending or even to expand a successful program. Unfavorable results can be used to motivate policy changes such as modifications to an ineffective program. We suggest that such validation studies be conducted periodically, either to monitor the effect of a major policy change on an ECE program or to ensure that a program that performed well in the past continues to be effective.

Proceed with caution if considering a QRIS rating component that is based on estimates of a program's effect on child developmental outcomes (Approach C). Although the goal of measuring the effect of participating in a specific ECE classroom or program on child developmental outcomes and incorporating the results into a program's QRIS rating has merit, the available methods—short of an experimental design—are not sufficiently well developed to justify the cost of large-scale implementation or implementation in high-stakes contexts. Moreover, the reduced reliability and validity of measures of the performance of children under age five make this high-stakes use highly questionable. The K–12 sector has experienced a number of challenges in using methods such as VAM to make inferences about the contribution of a specific teacher, classroom, or school to a child's developmental trajectory. These challenges would be compounded in attempting to use such methods in the ECE context given the tender age of the children involved and the challenges in assessing their performance in a reliable and valid manner. If a state is considering incorporation of this approach into its QRIS, we suggest that the process begin with a pilot phase to assess feasibility, cost, and return on investment. Given experiences with VAM in the K–12 context, a number of challenges will need to be overcome before Approach C is likely to be a cost-effective tool for incorporating child outcomes into a QRIS.

In sum, although QRISs have gained currency as input-focused accountability systems, the focus on inputs does not preclude efforts to get to the outcome of interest: child cognitive, social, emotional, and physical functioning. This paper describes valuable and feasible approaches for incorporating assessments of child functioning into QRISs or QI efforts for ECE programs more generally as a means of improving instruction and assessing program and system validity and performance. Some approaches take a micro perspective, and others have a macro focus. Some are predicated on having a QRIS in place, and others can be implemented without one. Our guidance illustrates that multiple approaches can be used given their varied and complementary purposes. At the same time, some of these approaches raise methodological concerns that must be dealt with and that may override the potential benefits. Ultimately, policymakers at the state level need to determine the mix of strategies that will be most beneficial given the context of the ECE system in their state, the resources available, and the anticipated returns.

Acknowledgments

This work was sponsored by the David and Lucile Packard Foundation. We are particularly grateful for the guidance and feedback provided during the course of our work by Meera Mani of the foundation, who saw the importance of providing targeted research-based support to the Early Learning Quality Improvement System Advisory Committee, which oversaw the design of the QRIS, and the Early Learning Advisory Council, which was established to oversee the system's piloting, refinement, and implementation.

We received valuable comments on an earlier draft from Kelly Maxwell at the Frank Porter Graham Child Development Institute, University of North Carolina, Chapel Hill. We also appreciate the careful research assistance provided by RAND Pardee Graduate School fellow Ashley Pierson. Our work also greatly benefited from administrative support provided by Christopher Dirks.

The RAND Labor and Population review process employs anonymous peer reviewers, including at least one reviewer who is external to RAND. Three anonymous reviewers provided thorough and constructive feedback on the draft paper, for which we are grateful.

Abbreviations

ASQ	Ages and Stages Questionnaire
CAELQIS	California Early Learning Quality Improvement System
CDD	Child Development Division
CDE	California Department of Education
DRDP	Desired Results Developmental Profile
ECCRN	Early Child Care Research Network
ECE	early care and education
ECERS-R	Early Childhood Environment Rating Scale–Revised
FDCRS	Family Day Care Rating Scale
ITERS	Infant/Toddler Environment Rating Scale
NACCRRA	National Association of Child Care Resource and Referral Agencies
NAEYC	National Association for the Education of Young Children
NCLB	No Child Left Behind
NICHD	National Institute of Child Health and Human Development
P.L.	Public Law
PPVT	Peabody Picture Vocabulary Test
QI	quality improvement
QRIS	quality rating and improvement system
RD	regression discontinuity
VAM	value-added modeling
WJ	Woodcock-Johnson (achievement test)

CHAPTER ONE

Introduction

The ultimate goal of the development and implementation of a state early care and education (ECE) quality rating and improvement system (QRIS) is to raise the quality of child care and early learning settings; these higher-quality settings are expected to improve child functioning, including school readiness, in relevant domains. QRIS logic models focus on one set of inputs to child development—various dimensions of ECE quality—with the goal of improving system outcomes, namely, child cognitive, social, emotional, and physical development. Yet although improved child outcomes are the ultimate goal, QRISs rarely directly assess children's developmental outcomes to determine if the system itself improves child functioning, nor do they require child assessments for purposes of evaluating specific programs. This is because it is costly and difficult to accurately measure these outcomes and difficult to link the contribution of any given child care or early learning setting to a particular child's developmental trajectory.

Despite these challenges, it is important that QRISs use child assessments to at least some extent because they do represent the ultimate goal of these systems. Such assessments also can be used to examine the viability of the logic models underlying these systems. The purpose of this paper, then, is to identify options for states to consider for incorporating child assessments into the design, implementation, and evaluation of their QRISs or related quality improvement (QI) efforts. Our analysis draws on decades of research regarding the measurement of child development and the methods available for measuring the contribution of child care and early learning settings to children's developmental trajectories. We also reference new research documenting the approaches taken in other states to include measures of child development in their QRISs and lessons learned from those experiences.

In focusing on the options for incorporating child assessments into QRISs, we consider approaches that are relevant for the child age ranges and setting types covered by state QRISs. In terms of child ages, the strategies we discuss can apply throughout the early years, from birth to kindergarten entry. In many cases, although the appropriate assessment tools may vary with the age of the child, the general approaches we cover apply to children across that age span. We also consider strategies that are relevant for the various ECE settings that serve children, from home-based care to center-based care, in both subsidized and unsubsidized settings. Again, the application of a given approach may vary with the type of setting, but the general approach is typically the same regardless. Where important differences arise with respect to child age or setting type, they are noted in our discussion.

In the remainder of this chapter, we set out definitions for key terms used throughout the paper, as our usage may differ from how terms have been employed in other literature. We conclude this introduction with a road map for the rest of the paper.

Defining Key Terms

Assessing young children and using these data to achieve assessment aims are complicated undertakings. Children can be assessed in many different ways and for many different purposes. Those conducting the assessments may be teachers, specially trained assessors, or researchers. Assessments may occur at different points in the life of a child, in an intervention, or in a QRIS. In addition, terminology for all of these components may differ as well. Here, we briefly describe the terms we will use in this paper and their meanings, which may differ in some cases from how these terms are commonly used. In later chapters, we will articulate some of these terms in more detail as the discussion focuses on different aspects of assessments.

Assessment. This is the most generic term we use, and it refers to any effort made to determine a child's level of functioning through use of one or more specified approaches or measures. Instruments may be structured or not and also may provide data on the performance of other children (normative data), which are used to determine whether an individual child's performance is within or outside normal developmental trajectories. Some approaches may be quite informal, e.g., observations and notes that teachers make on children's progress during the course of a day. The goal of any assessment is to identify how well a child is functioning. As discussed below, assessments may be used to screen an individual child for learning problems, to identify areas where a child needs extra help in the learning process, to improve a program, or to assess how well an intervention is working. Specific goals for assessments should always be set to ensure that instruments that best address those goals are selected, that appropriate technical requirements are met, and that the appropriate type of assessor is used (Shepard, Kagan and Wurtz, 1998). In addition, when assessments are focused on the performance of teachers, programs, or interventions, steps may be taken to reduce the assessment burden on any individual child, as discussed in a later chapter.

Functioning. This term describes how well a child is doing. Unlike performance, listed next, it does not refer to a specific assessment. Therefore, it is a more generic term and is used frequently in this paper.

Performance. Performance represents the findings of an assessment and describes how well a child has done on a given assessment at a particular time in a particular circumstance.

Outcomes. This term is used to describe the results of child assessments that may be used for a particular purpose: to determine how well an activity, program, or intervention (here, high-quality ECE or a QRIS) is contributing to child functioning. As discussed in more detail in a later chapter, use of the word *outcomes,* therefore, implies that the causal effect of some activity or intervention is being evaluated and that child assessments are being used as an input into that evaluation.

Screening. This term is used to describe assessments conducted by teachers and caregivers to map an individual child's developmental progress to better understand the factors that may be contributing to demonstrated learning delays. Since the purpose of such assessments is to determine the learning needs of an individual child and how to meet them, the child must carry the full assessment burden. If the screening suggests disabilities or behavioral or other issues outside the range of normal development, the child is referred to a specialist for further assessment. The conduct of such screenings is very useful in identifying children who have problems early on, which is why they have been incorporated as a quality element in a number of QRISs. However, because such assessments do not have a programmatic focus, we do not discuss them at length in this paper.

Road Map for the Paper

We continue in the next chapter by providing relevant background for this paper, with a discussion of the motivation for QRISs and a brief description of their history. We conclude that chapter with a discussion of some of the key challenges encountered in assessing young children and using assessment data. In Chapter Three, we present a framework for classifying approaches to using child assessments in state efforts to improve ECE quality, including QRISs. We describe five approaches that vary in terms of purpose, who conducts the assessment, and the sort of design needed to ensure that the resulting child assessment data can be used in a meaningful way. We then review each approach in terms of its current use, lessons learned from prior experience, the resources required for implementing it, the benefits of the approach, and possible barriers to success and ways to mitigate those barriers. We conclude in Chapter Four by providing guidance regarding the use of the five strategies based on our assessment of their overall strengths and weaknesses and the potential benefit relative to cost of each approach.

CHAPTER TWO

The Ultimate Goal of State QRISs: Improving Child Developmental Outcomes

Research findings point to the importance of the period from birth to school entry for children's development and focus attention on the quality of care and early learning experiences that young children receive (Lamb, 1998; Scarr, 1998; Vandell and Wolfe, 2000; Shonkoff and Phillips, 2000; Bowman, Donovan, and Burns, 2001; Center on the Developing Child, National Forum on Early Childhood Program Evaluation, and National Scientific Council on the Developing Child, 2007). Numerous studies have demonstrated that higher-quality care, defined in various ways, predicts positive developmental outcomes for children, including improved language development, cognitive functioning, social competence, and emotional adjustment (e.g., Howes, 1988; Burchinal et al., 1996; National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network [ECCRN], 2000; Peisner-Feinberg et al., 2001; Clarke-Stewart et al., 2002). High-quality programs for the most vulnerable infants and toddlers can lead to higher cognitive test scores through young adulthood and are associated with higher achievement in school and a greater likelihood of attending college (Campbell and Ramey, 1995; Ramey et al., 2000). Disadvantaged children who attend effective preschool programs have stronger language and math skills in the first years of elementary school and are less likely to repeat a grade, require special education classes, or drop out of school (Karoly, Kilburn, and Cannon, 2005).

In an attempt to better understand the consistent but usually modest associations between quality and children's developmental outcomes found in the literature (e.g., Nelson, Westhues and MacLeod, 2003; Karoly, Kilburn, and Cannon, 2005; Burchinal et al., 2009; Burchinal, Kainz, and Cai, 2011), recent work has examined in more nuanced ways how program quality operates to improve child outcomes. These efforts, summarized in Zaslow et al. (2010), have analyzed why studies of the relationships of quality to child outcomes so often produce small and often inconsistent results. One important approach to addressing this issue has been to question the assumption of linearity between quality and child outcomes: that is, that higher quality will produce better child outcomes regardless of where programs begin on the quality spectrum. These studies 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), thresholds (whether a particular quality level must be achieved to demonstrate effects on children), and quality features (which aspects of care matter most in improving child outcomes). Findings suggest that dosage is a key factor, and its effect is magnified when the program provides high-quality care (Zaslow et al., 2010). A small group of studies that consider both dosage and thresholds find stronger effects on child outcomes when quality is high (e.g., Votruba-Drzal, Coley, and Chase-Lansdale, 2004; Dearing, McCartney, and Taylor, 2009). Together, these studies suggest that a more differentiated, multidimensional understanding of

quality that considers dosage and thresholds may help to explain and improve it (Zaslow et al., 2010).

These new analyses underscore even more clearly the importance of high-quality care in improving children's outcomes. At the same time, we continue to find that many children participate in programs whose quality is not sufficient to improve their outcomes and may, in fact, undermine their development. Armed with knowledge about its importance and its absence from many programs, policymakers in states and some localities have turned to QRISs as a strategy for improving the quality of ECE programs in the public and private sectors so that more children can benefit from high-quality ECE. In this chapter, we review the factors motivating the use of QRISs and provide a brief history of the evolution of these systems. We then offer background on the use of child assessments in QRISs, as a preview for a more in-depth treatment in Chapter Three. We conclude with a discussion of the challenges associated with using child assessments appropriately and effectively.

Motivation for State QRISs

We articulate three primary factors that have propelled the use of QRISs in the last decade: gaps in quality in existing ECE programs, the inability of the current ECE system to promote uniformly high quality, and features of the market for child care and early learning programs that limit the consumption of high-quality services. With this perspective on motivation, it is possible to see the logic of QRISs.

Quality Shortfalls in Existing ECE Programs

Despite evidence showing the benefits of high-quality care, the ECE experiences received by many children in home and center settings do not meet quality benchmarks, often falling far short of even "good" care (Peisner-Feinberg and Burchinal, 1997; Fuller and Kagan, 2000; NICHD ECCRN, 2003; Duncan, 2003; Whitebook et al., 2004; Barnett and Ackerman, 2006; Karoly et al., 2008). In one of the earliest large-scale studies of child care quality covering four states (including California), researchers used the Infant/Toddler Environmental Rating Scale (ITERS) to conduct on-site assessments of 401 centers. The study found that 75 percent of infant classrooms and over 67 percent of toddler classrooms did not meet the "good" benchmark (Helburn, 1995). Nearly 50 percent of the infant and toddler rooms provided poor-quality care, including many rooms that scored low on basic measures of health and safety. A recent California study (Karoly et al., 2008) found that, depending on the quality measure, between 30 and 80 percent of preschool-age children in center-based programs with the largest gaps in school readiness and subsequent achievement do not participate in center-based programs that meet quality benchmarks in terms of common input indicators such as staff-child ratios and teacher qualifications. When the researchers assessed programs using the process measures that are most closely linked to school readiness, e.g., instruction in thinking and language skills, they found that 80 to 90 percent of the disadvantaged children in the California study enrolled in center-based programs were receiving care that would not meet quality benchmarks.

Concerns about poor-quality care have been exacerbated by a policy focus in recent years on children's academic achievement and the degree to which ECE promotes school readiness and improves children's academic performance in kindergarten and beyond. The K–12

accountability provisions in the No Child Left Behind (NCLB) Act of 2001 (Public Law [P. L.] 107-110) have drawn attention to the social and cognitive skills children need to build successful careers at school and have led K–12 leaders to scrutinize the skills of children entering kindergarten. Often, these skills are limited at best. K–12 leaders argue that if children are entering kindergarten unprepared to learn, then K–12 actors should not be expected to meet rigorous standards for their progress in elementary school.

Existing ECE Systems Do Not Ensure High Quality

The generally low quality of child care has led to calls for improvement, amid recognition that the current ECE system in the United States, if it can be called a system at all, does little to promote quality (National Early Childhood Accountability Task Force, 2007). Indeed, the U.S. "system" of child care and early learning has been described as "a nonsystem of microenterprises" (Kagan, 2008). Most providers are underfunded and only loosely regulated.

ECE programs are delivered by a variety of providers, including publicly subsidized center-based programs (such as Early Head Start, Head Start, and state-funded child development and prekindergarten programs), private child care centers (which may also serve children whose care is subsidized), home-based family child care programs, and friend-and-neighbor care. Centers and family child care homes are the most likely to be licensed; they are also the types of care settings that are typically the focus of QRISs.

At the lowest level, quality standards in most states are largely defined by licensing requirements, which are set by states and vary widely in their scope and rigor. For example, whereas states (including the District of Columbia) generally require that centers be licensed, and most states (33) require that child care homes serving four or more children be licensed, 16 states do not impose licensing requirements unless a program serves five or more children, and two states do not require any license for family child care homes. Centers under religious aegis are license-exempt in nine states (Smith and Sarkar, 2008).

Although much care is licensed, licensing represents a fairly low quality bar, focused as it is on the adequacy and safety of the physical environment. Licensing requirements address such things as fencing, square footage, and protecting children's health and well-being by covering plugs and locking up cleaning supplies. They essentially ignore other aspects of program quality, although many states also impose ceilings on group sizes and staff-child ratios or require minimal caregiver training (National Association of Child Care Resource and Referral Agencies [NACCRRA], 2011). In recent years, in response to fiscal constraints, even these minimal requirements are less likely to be monitored. In some states, licensing visits occur as infrequently as every five years, despite NACCRRA's recommendation for quarterly licensing inspections, with at least some of these inspections unannounced (NACCRRA, 2011).[1] Moreover, in its focus on easily assessed environmental features, the licensing process emphasizes compliance with a checklist of items and does not encourage continuous quality improvement (Zellman and Perlman, 2008).

For some publicly funded programs, a higher level of quality standards applies. The federally funded Early Head Start and Head Start programs, for example, are subject to a set of program standards that set minimal requirements on such program structural features as group sizes, ratios, and staff qualifications that are typically more stringent than the comparable state

[1] This standard is followed for Department of Defense–sponsored care.

licensing standards. Likewise, as more states have introduced publicly funded prekindergarten programs, the associated program standards typically represent a higher bar than the licensing requirements. For example, although just one state currently requires a bachelor's degree in the ECE field for the lead teacher in a licensed center-based classroom, 22 states require that degree level for lead teachers in the state preschool program (Barnett et al., 2010; NACCRRA, 2011). Starting in 2013, Head Start will require that at least half of teachers nationwide have a bachelor's degree in ECE. While setting a higher bar, program regulations for publicly funded programs can vary considerably by funding source and may still fall short of the benchmarks that are defined for high-quality programs. For example, California's Title 5 State Preschool Program is rated as meeting just four of ten benchmarks for high-quality preschool programs established by the National Institute for Early Education Research (Barnett et al., 2010).

Features of ECE Markets Limit Use of High-Quality Services

As ECE consumers, parents must be able to distinguish between low- and high-quality settings when making choices about their desired level of quality, subject to the price of care they can afford.[2] However, parents are not very good at evaluating the quality of care settings. Although some believe that quality is obvious and parents will "know it when they see it," research suggests that this is not the case (e.g., Helburn, Morris, and Modigliani, 2002). Parents tend to rate child care providers very positively (e.g., Helburn, 1995; Barraclough and Smith, 1996; Cryer and Burchinal, 1997; Wolfe and Scrivner, 2004), and their ratings are uncorrelated with observer quality ratings (e.g., Barraclough and Smith, 1996; Cryer and Burchinal, 1997; Cryer, Tietze, and Wessels, 2002). Helburn (1995), for example, found that parents rated centers nearly twice as highly as did trained assessors on such key elements as health, safety, and interactions between staff and children.

Some parents rely on information that comes from licensing or program standards to assess program quality. Yet, most parents do not fully understand the licensing process: In a recent survey of parents, 62 percent believed that all child care programs must be licensed, and 58 percent believed that the government inspects all child care programs. Many believe that licensing includes scrutiny of program quality and that licensure indicates that a program is of high quality (NACCRRA, 2011). As the California Early Learning Quality Improvement System (CAELQIS) Advisory Committee (2010a) report notes, these findings highlight the need for reliable and valid information about the health and safety of ECE settings and about the quality of the care and early learning environments provided. Parents need this information to make wise choices on behalf of the real "consumers," their children. In the same way, government agencies that subsidize ECE program through direct provision (i.e., programs funded by grants or contracts) or through subsidies (i.e., vouchers) also need reliable and valid information about program quality, either to set a bar for the programs that will be subsidized or to tie reimbursement rates to the quality of the program.

But even if parents better understood licensing and quality more generally, the limited availability of care in many locations and for key age groups (particularly infants) provides ready clients for most providers, even those who do not offer high-quality services. The high cost of high-quality care and limited public funding to subsidize the cost of ECE programs for low-income families further constrain the demand for high-quality care. Providing high-

[2] See Iruka and Carver (2006) for ECE use patterns by key child and family characteristics based on the 2005 National Household Education Survey.

quality care is costly, and most parents cannot afford to pay the full cost of such care.[3] With the exception of universal programs, most subsidized ECE programs are not fully funded so that many eligible families are not served.

This strong demand for care at existing quality levels limits incentives for providers to take often-costly steps to improve. In some cases, providers may not know how to improve, even if they are motivated to do so. In addition, there are few empirical data available that providers can use to help them select the best ways to invest limited funds to maximize improvements in quality. Another constraint on quality improvement is parents' inability to recognize high-quality care and distinguish it from care of moderate or mediocre quality. At the same time, parents' inflated quality judgments may represent only a minor handicap in choosing care, since for many, care decisions are strongly influenced by supply and such convenience factors as cost, hours of operation, availability of a space, and location (Barraclough and Smith, 1996; Blau, 1991; Leslie, Ettenson, and Cumsille, 2000; Van Horn et al., 2001; Seo, 2003). Perceived center quality was not one of the reasons most often considered by parents when selecting a child care center (Seo, 2003).

The Logic of QRISs

The growing scrutiny of ECE settings, the lack of market incentives to improve, and the lack of QI skills and knowledge among some well-meaning providers have fueled concerns about the level of ECE quality and have focused attention on ways to improve it. In recent years, QRISs have become an increasingly popular policy tool to improve quality and have been adopted in many localities and states. QRISs incorporate rating systems based on multicomponent assessments designed to make child care quality transparent and easily understood. Nearly all systems explicitly include feedback and technical assistance and offer incentives to both motivate and support quality improvement. QRISs are essentially accountability systems centered around quality ratings that are designed to improve ECE quality by defining quality standards, making program quality transparent through ratings, and providing supports for quality improvement. A comprehensive QRIS provides workforce development, financial incentives, and other supports to improve quality and strengthen the core components of an early learning system (CAELQIS, 2010b; Kauerz and Thorman, 2011). QRISs recognize that providers need help to improve and that ratings alone may not be adequate for formulating improvement plans. Moreover, quality improvements cost money. In particular, lower staff-child ratios and better-educated and trained staff, two components that are generally viewed as critical to improving quality, are major cost drivers.

QRISs are sometimes contrasted with accreditation, another tool to improve ECE quality. Accreditation, generally associated with the National Association for the Education of Young Children (NAEYC) (although other organizations also accredit ECE programs), is designed to help ECE providers improve the services they provide by engaging staff in a self-study process followed by a validation visit. However, because of the rigor and cost of the process and the high standards attached to attaining it, accreditation has been taken up by very few providers. As of 2008, NACCRRA reports that only 10 percent of all ECE programs are accredited,

[3] A significant exception to the association between cost and quality may be found at Head Start centers and at Child Development Centers sponsored by the Department of Defense for military dependents. In both of these settings, substantial subsidies enable the children of low-income families to receive high-quality care at very low cost (Zellman and Gates, 2002; U.S. Department of Health and Human Services, 2004).

with the majority of them accredited by NAEYC (see Smith and Sarkar, 2008; NAEYC, 2011). Further, there is some evidence that fewer providers have been seeking NAEYC accreditation in recent years (Smith and Sarkar, 2008). Some attribute this decline to the growing prevalence of QRISs, which allow programs to participate in these systems and set quality goals at different levels.[4]

The focus of QRISs on improving ECE inputs and children's outcomes through increased accountability is consistent with policy efforts in K–12 education, where student test scores are used to assess school performance. Advocates for high-quality ECE are generally quite enthusiastic about the potential of these systems, largely because of their scope, the infusion of public funds into them, and their focus on improving quality in providers that enter the system at very different levels of quality. At the same time, QRISs differ in a key way from K–12 accountability efforts because they focus almost exclusively on inputs into caregiving and caregiving processes rather than on outcomes of the process, which for K–12 accountability systems are measures of student performance on standardized assessments. There are, of course, good reasons for this. The key one is that the paper and pencil assessments on which K–12 accountability systems rely as outcome measures cannot be used with young children, as discussed in more detail below.

Instead, most QRISs assert a link between improved program quality and enhanced child functioning. Such effects are more likely when the measure of quality in question is more closely related to a child's direct experiences in care (e.g., Mashburn et al., 2008). This link is supported by findings of correlations between the two. In addition, the results of rigorous assessments of carefully controlled interventions using randomized control designs such as Abecedarian (e.g., Ramey and Ramey, 2006) find significant, long-term effects of a high-quality program on children's functioning. Research also finds that attending a quality preschool program is associated with higher achievement in elementary school for children in all income groups, although the benefits tend to be largest for children from disadvantaged backgrounds (e.g., Gormley and Gayer, 2005; Gormley et al., 2005; Karoly, 2009; Pianta et al., 2009). However, more recent research has examined whether the relationship between improved program quality and enhanced child functioning is linear. This work suggests that relationships between quality and child functioning are stronger when quality is high (e.g., Zaslow et al., 2010) and that threshold effects, in which effects on child functioning depend on a certain level of quality being in place, may affect these relationships.

The general theory underlying QRISs is that improved program quality will contribute to better child functioning. However, as Zellman et al. (2008) note, this theory has not yet been tested, although evaluations have examined selected parts of these systems.[5] QRIS logic models articulate each system's underlying theory. These models may differ slightly, but they generally include similar processes and outcomes (see Zellman et al., 2011, for a discussion of several QRIS logic models). Basically, a logic model is a systematic and visual way to present the relationships that are expected to exist among the resources available to the effort or program, the activities or polices that are to be put in place, and the changes or results that are expected to follow (Kellogg Foundation, 2004). Logic models provide stakeholders with a road

[4] See Zellman et al. (2008) for further discussion of accreditation in the rating portion of QRISs.

[5] For example, several evaluations have examined the relationship between QRIS ratings and environment rating scales such as the Early Childhood Environment Rating Scale–Revised (ECERS-R) or child assessments (e.g., Elicker et al., 2007; Elicker and Thornburg, 2011; Langill et al., 2009; Tout et al., 2010b).

map describing the sequence of inputs and activities and their connections with the program's desired results. Presented graphically, as in Figure 2.1, logic models display designers' theory of how proposed activities and policies will lead to desired goals through a logical chain of "if-then" relationships. A well-articulated logic model describes key steps in the process and key outputs of each step. Those outputs, when well-defined, identify measurable behaviors or indicators at each stage of the implementation process, e.g., "meetings are held at least quarterly between specified actors" rather than "more collaboration occurs" (e.g., Rossi, Lipsey, and Freeman, 2004; Chen, 2005). These indicators constitute the measures of the initiative's progress toward meeting its stated goals (Zellman et al., 2011).

The model presented in Figure 2.1, based on Colorado's QRIS (Zellman et al., 2008), focuses on parents and providers and articulates the process assumed to be involved in implementing a QRIS in some detail. In particular, it indicates the multiplicity of changes in behavior that are required to achieve the longer-term and ultimate outcomes of a QRIS, which are better emotional and cognitive outcomes, including school readiness. This model is read "bottom to top," emphasizing how the currently available resources and activities lead to measured outputs and outcomes. However, for strategic planning, it is often most useful to read a logic model from "top to bottom," starting with the effects and outcomes desired, then considering what activities are likely to yield those outcomes and what resources are required to implement those activities (Breitner, Brandon and Lalic, 2010). (See Zellman et al., 2011 for examples of other QRIS logic models.)

A Brief History of State QRISs

The compelling logic of QRISs has led to their rapid adoption, starting with the first system in 1998 in Oklahoma and now reaching more than two-thirds of the states. A brief overview of the evolution of the QRIS landscape and the shape of QRIS designs, including the use of child assessments, provides additional perspective on this phenomenon.

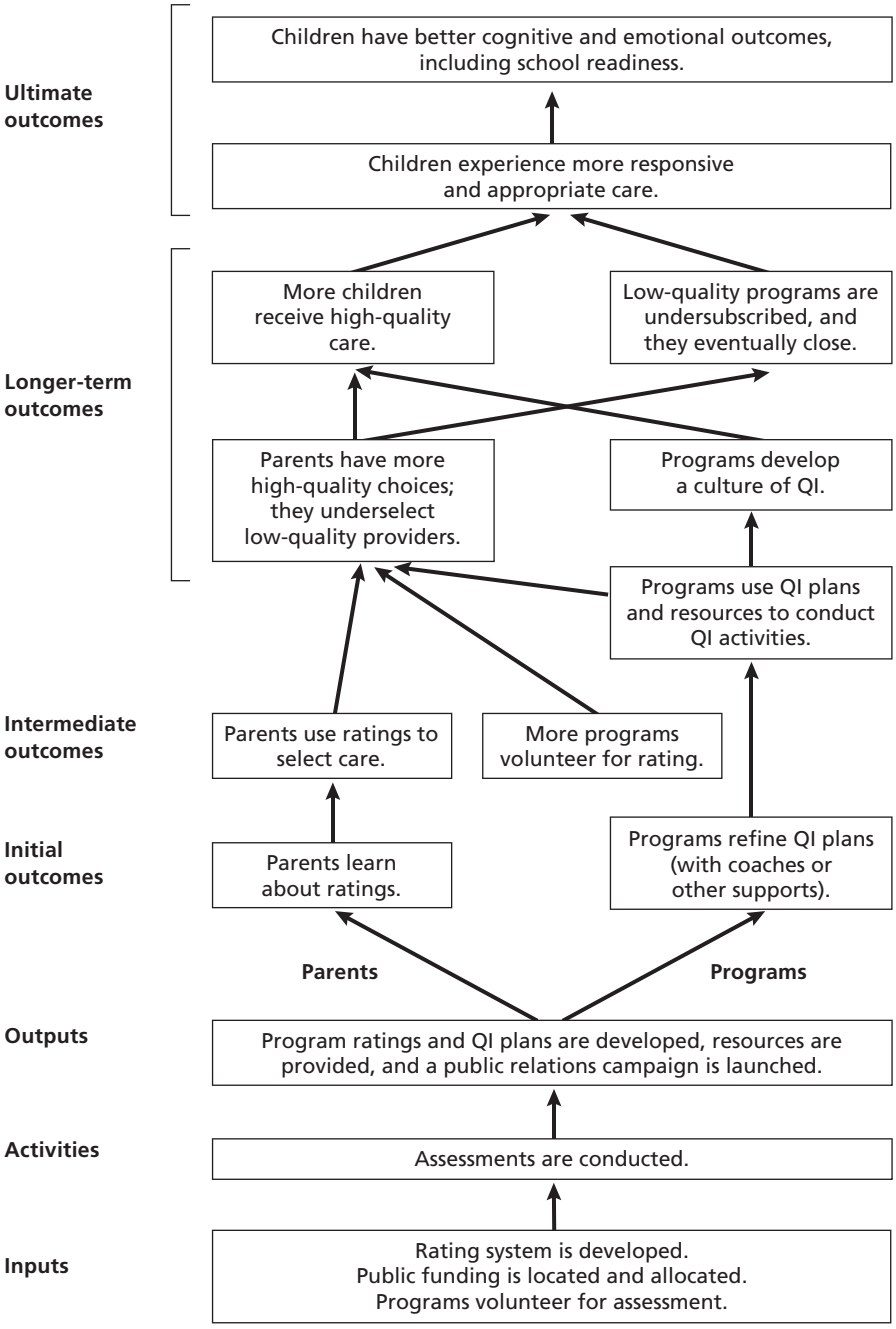
The QRIS Landscape

QRISs have proved popular with state legislatures in recent years because they represent a conceptually straightforward way to improve quality that appeals both to child advocates—because of the promise of support for improved quality—and to those who support market-based solutions—because QRISs incentivize improvement. They are also consistent with a general trend toward demanding accountability in government-funded programs; the NCLB legislation noted above is a good example of this trend. The number of states that are implementing some form of rating system, including system pilots, has increased from 14 in early 2006 to 35 at the time of this writing in 2011.[6]

The earliest QRIS was designed and implemented in Oklahoma. It came about because system designers had learned that the state legislature was not willing to increase public funding for the care of low-income children given the poor quality of many programs that served

[6] Information on state rating systems and pilots was compiled by the authors in January 2011 from existing sources, including publications from the National Child Care Information and Technical Assistance Center, the U.S. Department of Health and Human Services Administration for Children and Families, Office of Planning, Research and Evaluation (Tout et al., 2010b), and online documents including state websites.

Figure 2.1
A Logic Model for QRISs



SOURCE: Zellman and Perlman (2008), Figure 1.1.

RAND *OP364-2.1*

them. It was hoped that a QRIS would make legislators more willing to allocate additional reimbursement funds. Designers' goals were to improve the quality of child care by increasing the training and education of providers and to provide parents with a simple tool to evaluate care quality. An additional goal was to increase provider reimbursements, which would help

to increase the number of spaces available to low-income families by making caring for those children more remunerative.[7]

Oklahoma began its QRIS design process in 1997 and launched *Reaching for the Stars* in 1998. Oklahoma did not pilot its rating system but has changed its system regularly since its initial rollout. Oklahoma began with a two-level system. One star was awarded automatically with licensing. A second star required that a program meet internal quality criteria or achieve NAEYC accreditation. In 1999, a third star was added to its two-level system. The following year, the "1-star plus" level was added because so few programs could reach the 2-star level (see Zellman and Perlman, 2008, for further discussion of Oklahoma's QRIS and that of four other QRIS pioneer states).

Oklahoma QRIS designers bemoan the fact that as pioneers, they had no other states to look to in designing their system (Zellman and Perlman, 2008). Other states quickly began to look to Oklahoma, however. Key lessons learned included the importance of piloting and the need to carefully balance the realities of current quality against long-term quality goals. Oklahoma designed its system knowing that few programs could meet rigorous standards and created a system with few levels so that providers starting at or near the bottom would not be discouraged. But they quickly learned that multiple steps were important both to encourage growth and to ensure that small improvements mattered (Zellman and Perlman, 2008).

QRIS Design

The first states that adopted QRISs tended to follow similar processes in developing and implementing them (Zellman and Perlman, 2008). Each state set goals; assessed feasibility; and designed, implemented, and assessed its system. The process for most is continuous; the outputs from the "final" stage (evaluation) are, in turn, used to reassess feasibility and make further design and implementation changes.[8]

One of the most important decisions in QRIS design concerns the selection of the quality components for the rating system. As one QRIS designer interviewed by Zellman and Perlman (2008) astutely noted, "What matters is to include what matters." In all states, interviewees who discussed component choice noted that what gets included is attended to, whereas what is excluded is likely to be ignored. In selecting the components for their systems, all of the pioneering states that Zellman and Perlman (2008) included in their study turned to the research literature to determine which care components were associated with better child outcomes.[9] Interviewees in Colorado and Ohio specifically mentioned relying on the results of the Cost, Quality, and Child Outcomes Study (Helburn, 1995; Peisner-Feinberg and Burchinal, 1997; Peisner-Feinberg et al., 1999). Oklahoma also included components that developers believed needed attention because they were either absent or set at low levels in licensing requirements. And several states, heeding the idea that ignored components tend to be ignored, made com-

---

[7] See Mitchell (2005) for further discussion of the many goals that underlie these systems.

[8] QRIS developers have discovered over time that there is a tension between continuous improvement and acceptance of the system. When Oklahoma, for example, decided to change its system after discovering that the standards were too low, providers who signed on believing that they could meet the standards were upset to discover that the standards had changed and their ability to qualify for the rating to which they had aspired was now in question (Zellman and Perlman, 2008).

[9] Zellman and Perlman (2008) selected five states to study from among the 14 states that had a statewide QRIS in place as of January 2007. They designated Colorado, North Carolina, Ohio, Oklahoma, and Pennsylvania as pioneer states as all had statewide systems in place by 2004, and three of the five had implemented their system before 2000.

ponent decisions that did not rely completely on research findings. For example, several interviewees mentioned a lack of data to support the inclusion of parent involvement and the lack of a clear way to operationalize the construct. Nonetheless, a decision was made in three of the five pioneer states included in Zellman and Perlman's (2008) study to include a parent involvement component in the rating system because designers wanted to encourage programs to promote it.

These reviews were then subjected to discussions about which components could be well measured and which were feasible to measure given wide variation in the cost of measurement. For example, Ohio interviewees noted that parent involvement was originally included in their rating system but was later dropped when the measures they were using to assess this component, such as number of parents who attend meetings, began to be seen as not credible.

In general, the rating system component choices made by the five states covered by Zellman and Perlman (2008) were fairly similar. Succeeding states have tended to make similar ones as well. The QRIS Compendium found that six quality categories were included in the majority of the 26 center-focused systems assessed (Tout et al., 2010b). These categories include licensing compliance (26 systems), environment (24 systems), staff qualifications (26 systems), family partnership (24 systems), administration and management (23 systems), and accreditation (21 systems). Three categories—curriculum (14 systems), ratio and group size (13 systems), and child assessment (11 systems)—are included in half or just under half of the QRISs assessed. The quality components assessed by rating systems focused on family child care were quite similar to those for centers.[10]

The Role of Child Assessments in QRISs

As noted above, QRISs were conceptualized as input-focused accountability systems, different from K–12 accountability schemes or most other accountability systems, which focus on the products of the system, whether these products are academic skills in second- to twelfth-graders, higher rates of mammograms in high-risk patients, or the quality of donuts produced by an industrial bakery (Stecher et al., 2010). The focus on measures of inputs to quality rather than outputs—such as children's level of school readiness, literacy, or numeracy or noncognitive skills such as self-regulation or the ability to follow instructions or get along with peers—was considered a necessary concession to the reality that the performance and longer-term outcomes of young children are difficult and costly to measure and that measures of these attributes are less reliable and accurate than those for older children, as discussed below. The absence of a link between child outcomes and ratings, reimbursements, or technical assistance in QRISs is consistent with the recommendations of the National Research Council (Heubert and Hauser, 1999), which urges extreme caution in basing high-stakes decisions on assessment

[10] The QRIS Compendium's authors note that licensing requirements frequently represent a minimal set of provisions to ensure that care and education environments are safe and healthy and provide for children's basic needs (Tout et al., 2010b). Consequently, it is necessary to understand licensing requirements to make sense of the rating components of a QRIS. For example, QRIS developers in a particular state may conclude that the existing requirements for ratio and group size that are part of licensing requirements are sufficient to ensure children's safety in the environment. If they reach this conclusion, they might decide that there is no need to include a separate measure of ratios and group size in the rating portion of the QRIS. However, the lack of such an indicator among the rating components clearly does not mean that maintenance of appropriate ratios and group sizes is not a priority. Similarly, health and safety requirements are typically included in licensing, so further indicators in this category may not be included.

outcomes. Snow and Van Hemel (2008) likewise urge ". . . even more extreme caution" when dealing with assessments of children under age five.

Moreover, although not frequently discussed, it was understood that requiring assessments for purposes of evaluating program effectiveness would divert limited funds from supporting the sorts of improvements to quality inputs such as teacher education, reduced ratios and group sizes, and more professional development that QRIS designers believed were key to making a difference for children's development. This focus on inputs did not reflect any abandonment by QRIS designers of the ultimate goal of these systems, which remained the improvement of child functioning through exposure to higher-quality care. Indeed, QRISs were a policy tool created to improve quality and to capture more public funding in support of the goals of improved quality and greater accessibility of high-quality care, especially for low-income children. Despite the absence of discussion about the measurement of children's functioning in most pioneering systems, advocates understood this to be the ultimate goal. Their focus on improving inputs was supported by a literature that frequently asserted and sometimes demonstrated that higher-quality care was associated with improved child functioning, as discussed above. Nevertheless, only a few states have included child outcomes in their QRIS evaluations to date, as discussed below.

As QRISs have developed and been refined over time, child assessments have increasingly found their way into QRISs, although they generally serve a different purpose from the one discussed above. Indeed, the purpose of many of these assessments, which are typically used to clarify the learning process and growth in skills of the children assessed, more closely aligns with the input focus of QRISs. By determining how well children are doing, what they have learned, and in what areas they may need more support, it is hoped that these assessments can help teachers and administrators to improve the quality of the care they are providing both to individual children and to all children in a given classroom. By refining curriculum to better meet children's needs, and by individualizing their efforts to meet the specific needs of individual children, children's learning and emotional needs can be better served.

These sorts of child assessments are not common in QRISs; the Compendium lists only 11 of 26 systems that include them (Tout et al., 2010b). And in most of these systems, expectations for assessments do not kick in until the higher levels of the rating system. In three of these 11 systems, the assessment tool is specified; in five of the others, program staff may select from an approved list of measures. No information about the nature of the tools is provided for the remaining three systems.

In some cases, the child assessments may serve a different purpose: to screen children and identify those who may need referrals to specialized developmental or mental health services (more commonly, however, developmental screening is a separate activity carried out by specially trained assessors). As the Compendium notes, "Developmental screening of children is a related but different process from child assessment. While child assessments are used to individualize curriculum and instruction, screening is used to identify children who may need a referral to determine if they have a developmental disability" (Tout et al., 2010b, p. 110).

Efforts to use child assessments as outcomes that contribute to a determination about how well a QRIS is working are relatively rare. As of early 2011, just five state-level systems have

included child assessments in such efforts.[11] Key factors limiting the use of child assessments as outcomes are the many challenges associated with assessing young children. Some of the key challenges are discussed next.

Challenges in Assessing Young Children and Using Assessments

The focus of QRISs on the inputs side of the logic model in Figure 2.1 stems from the significant challenges associated with assessing children from birth to school entry and then using those assessments appropriately. A review of assessment issues and objectives helps to set the stage for the discussion in Chapter Three.[12]

Assessment Issues

Assessing young children presents a number of challenges and obligations. We briefly summarize some of these as a prelude to our discussion below of a number of strategies for using child assessments in QRISs. Detailed information on assessing young children may be found in Snow and Van Hemel (2008).

Characteristics of Young Children. The key constraint on the assessment of young children is the nature of the skills and forms of expression that characterize them. Unlike their older siblings, young children lack the skills and motivation to complete paper and pencil assessments with little to no adult monitoring (Bredekamp and Rosegrant, 1992, 1995; Guddemi and Case, 2004). Indeed, their tender age requires a great deal of adult involvement, as well as a set of assessment tools that recognize the reality that young children cannot read, often cannot hold a pencil, and learn and express themselves most articulately in ways that cannot be captured with paper and pencil assessments. Moreover, their attention spans are limited, and the expression of a given skill may vary substantially from one day to the next (Guddemi and Case, 2004). Beyond such developmental considerations, there is also a need to account for the cultural and linguistic contexts experienced by young children that can affect performance on assessments and interpretation of skills and limits (Bowman, Donovan, and Burns, 2001). Children's cultural experiences can have a substantial effect on how they view the assessment situation and how able and willing they are to respond to requests from an examiner (e.g., McCauley, undated). Assessment practices also must be sensitive to children's language development—both in English and in children's native language (Shepard, 1994).

Young children are most comfortable in familiar settings and with familiar people. Ideally, if they are to perform at their best, they will be assessed by someone they know in a familiar setting (Guddemi and Case, 2004). And, many argue, the very best assessments are those that do not appear to be assessments at all; assessments by teachers and caregivers, often through observations of children engaging in daily activities, are likely to produce the most valid information about children's abilities. But such in situ assessments may not meet the

[11] Our examination of the QRIS Compendium and of state websites indicates that Colorado, Missouri, Minnesota, Indiana and Virginia are using child assessments as part of QRIS assessments (Tout et al., 2010b). Elicker and Thornburg (2011) also list Ohio in this group.

[12] For more in-depth discussion of and guidance regarding approaches to assessment of young children, see Shepard, Kagan, and Wurtz (1998), Epstein et al. (2004), and Snow and Van Hemel (2008). See Halle et al. (2011) for a compendium of commonly used assessment and screening tools and a review of their reliability and validity.

needs of adults or children. Much depends on the purpose of the assessments and the way in which the resulting data are to be used. Even more than is the case for older children, it is critical to clarify the purpose of data collection, anticipate how data will be used, and find assessment tools and processes that deliver needed data without overburdening children. A number of researchers have noted that high-stakes assessments designed to determine a child's eligibility or placement (for example, in kindergarten) may not be appropriate, given the difficulties of assessing young children, and may exclude children who might benefit most from participation in high-quality programs (e.g., Shepard, 1994; Shepard, Kagan and Wurtz, 1998). In contrast, although ordinary classroom assessments designed to improve instruction also affect individual children, the consequences of these decisions for any one child are not nearly as great. And when assessments are used to evaluate programs or other interventions, the interest is in the group's average performance with few to no implications for an individual child. Of course, group assessments may need to meet very high standards for other reasons, e.g., schools or prekindergarten programs may lose autonomy or funds if children do not perform well (Shepard, 1994).

Ensuring Reliability and Validity. In selecting instruments to use with young children, it is critical to select tools that meet the guidelines for reliability and validity established by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999). These two criteria are not always easy to meet.

Reliability refers to the consistency of a measure over time and across different assessors. A test or other assessment is considered reliable if it produces the same results repeatedly. Given this definition, it is clear why reliability is a challenge in assessing young children: Their levels of knowledge and understanding can change in a very short period of time, and they may not express a skill in a consistent way (e.g., Shepard, Kagan, and Wurtz, 1998). Reliability also may be threatened by the complexity of the assessment and the limited amount of training that assessors are given. It may also be threatened by assessor drift—a tendency for assessors to diverge from a training standard over time. This latter threat can be addressed by following a protocol in which assessors are retested at specified intervals and retrained to standard as needed. However, such a practice is rare, suggesting that reliability may be a real but obscured problem in many assessments of young children and their environments (see, for example, Le et al., 2006).

Validity is also an important issue in assessing young children. Validation is a process that assesses the degree to which evidence and theory support the conclusions and interpretations derived from multicomponent assessments conducted in a specified context. The purpose of validation activities is to allow conclusions to be drawn about whether assessments measure what they purport to measure (Cizek, 2007). A number of ECE professionals have noted that children's culture and English or native language skills are important considerations when determining the validity of an assessment or assessment protocol and making interpretations about a child's learning capacity from his or her performance (see, for example, Shepard, 1994; Lynch and Hanson, 1996; Bowman, Donovan, and Burns, 2001; National Association of School Psychologists, 2009a; 2009b). Another key validity issue that is only rarely considered in the context of QRISs is that validity refers to the use of a specific instrument in a specific context. Often, assessments used in QRISs were designed for use in low-stakes settings such as research studies and program self-assessments. But QRISs increasingly represent high-stakes settings, where the outcomes of assessments affect public ratings, reimbursement rates, and

the availability of technical assistance, as discussed below. Assessment tools developed for low-stakes uses should be validated for use in high-stakes settings (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999).

Assessment Objectives

Child assessments may take a number of forms, including observations, standardized assessments, home inventories, portfolios, and running records. These assessments are generally understood to have three basic purposes: screening individual children for handicapping conditions, supporting and improving teaching and learning, and evaluating interventions. As Snow and Van Hemel (2008) note, it is critical that the goals of assessment be clearly understood in advance and that measures be selected that are valid for those purposes. Each of these assessment goals is discussed in turn below.

Screening Individual Children for Handicapping Conditions. Assessments of handicapping conditions typically have included two phases: The first is a screening assessment conducted by a teacher or caregiver on children who appear not to be making expected progress. The second is an in-depth developmental assessment conducted by a specialist on those children whose screening reveals some evidence of significant learning problems (Shepard et al., 1998). A third type of screening involves simpler assessments of large groups of children with no demonstrated learning delays to identify those who may need further assessment (Council of Chief State School Officers, 2003). This latter type of screening, which has increasingly been advocated as a way to ensure that learning barriers are addressed as early as possible, has raised concerns; they are mainly centered around the unnecessary singling out of children (Snow and Van Hemel, 2008). Universal screenings have been included in a number of QRISs as one indicator of program quality.

Screening by program staff of children with demonstrated learning or other developmental delays requires staff training and a clear sense of the purpose: to assess developmental progress to better understand the factors that may be contributing to delays (Shepard, 1994). Such assessments should be limited to those children whose learning profiles and performance suggest significant learning issues outside the wide variation in development found in groups of young children (Shepard, 1994). When such screenings document significant delay, referrals are typically made to specialists who then conduct in-depth assessments. Because these assessments are focused on individual children with identified behavioral, emotional, or learning issues, the assessment tools generally are developmental screening measures, focused on determining the nature and extent of any problem and the degree to which the problem represents a significant deviation from normal development. This latter point is key: Children, and particularly young children, develop at very different rates, so it is critical before labeling a child as delayed and referring him or her to special services to be able to examine the child's performance against performance norms for children in similar circumstances. This need generally leads to the selection of assessment tools that are widely used and standardized and make normative information available to assessors. However, it is critical in selecting instruments and particularly in interpreting results that norms be clearly understood before inferences about the child in question are made (e.g., Snow and Van Hemel, 2008). Most assessments were originally normed on samples of white middle class children (e.g., Reynolds, 1982), although efforts are increasingly being made to use more diverse samples. The focus on an individual child also

demands that that child undergo the assessments without any possibility of other children sharing the assessment burden, which may be possible in the case of assessments designed to serve other purposes, as described below. Ideally, such assessments can be conducted iteratively, so a child need not be assessed for too long during a single session.

Supporting and Improving Teaching and Learning. These low-stakes assessments are designed to determine how well children are learning so that curricula and other interventions can be modified to better meet children's learning needs, at the level of the individual child, the classroom, and the program. Ideally, the content of assessments should reflect and model progress toward important learning goals, which may include physical and socioemotional development, as well as cognitive skills (Shepard, 1994).

Key to these assessments is the establishment of a plan for using the data that are collected to actually improve programs and interventions. Hebbeler and Taylor (2011) note that using child outcome data for program improvement requires that programs develop a continuing process of program improvement that involves formulating key questions, analyzing data, interpreting data, developing hypotheses, and identifying and implementing a course of action based on findings, hypotheses, and inferences. Once changes have been implemented, new data will help to clarify the extent to which the actions taken were successful in helping children improve.

A key advantage of assessment for program improvement is that the focus can be largely on the program. With this goal in mind, planners can use the flexibility that the goal implies to reduce the assessment burden imposed on individual children. For example, it may be possible to test a lot of children on different pieces of an assessment protocol (Shepard, 1994).

Evaluation of Interventions. These child assessments represent a way to determine if interventions have been successful in achieving their goals. Unlike assessments designed to improve programs, which may be formal or informal, these assessments should be embedded in a rigorous research design that increases the likelihood of finding effects, if they exist, to the greatest extent possible. And as with assessments for program improvement, it may not be necessary to impose the full assessment burden on each child; it is often possible to assess children on different parts of an assessment protocol.

Since the logic models underlying QRISs typically assert that improved child outcomes represent the ultimate goal of the system, use of child assessments to examine QRIS functioning makes some sense. At the same time, designing and implementing such evaluations is very challenging for a number of reasons, described in more detail in Chapter Three. Most important, it is not easy to clearly attribute child outcomes or changes in child outcomes to a given teacher or program or to a QRIS. Indeed, some have argued that QRISs do not really represent interventions in the traditional sense: They do not impose a standardized curriculum, specify minimum exposures, or specify how content is to be delivered. On the other hand, QRISs do represent a *policy intervention* with the intended goal of improving the quality of care in ECE settings. Viewed in this way, assessing whether a QRIS achieves its objective is a legitimate question for an evaluation.

If a decision has been made to assess changes in child functioning over time, high attrition rates in many child care programs can undermine statistical power and reduce the likelihood of finding effects. Further, assessments of QRISs that involve comparisons with programs that are not participating in a QRIS may not be entirely fair, since voluntary QRISs are generally assumed to attract programs of higher quality. Although efforts are generally made to

control for differences across QRIS participants and nonparticipants, unmeasured differences may remain.[13]

[13] See Zellman et al. (2011) for more information about designing QRIS outcome studies.

CHAPTER THREE

Approaches to Using Assessments of Child Functioning in State ECE QI Efforts

The challenges of conducting assessments of child functioning and the logic of QRISs discussed in Chapter Two do not preclude the use of such assessments as part of state QI efforts. In this chapter, we begin by first establishing a framework that defines five strategies for using assessments of child functioning in state efforts to improve ECE quality, three of which are predicated on the existence of a QRIS. We then review each strategy in terms of its current use in state systems, lessons learned from prior experience, the resources required for implementing it, its benefits, and possible barriers to success and strategies for mitigation of these barriers.

A Framework for Classifying Approaches to Using Assessments of Child Functioning

After reviewing the literature and current state practices, we distilled current and potential approaches for incorporating child developmental outcomes into state QI efforts into five distinct strategies.[1] Table 3.1 labels each one and summarizes its purpose and approach, focus, and relationship to a QRIS. The strategies are arrayed in Table 3.1 from those that focus on assessments of child functioning at the micro level—the developmental progress of an individual child or group of children in a classroom—to those that have a macro focus—the performance of the QRIS at the state level or the effect of a specific ECE program or the ECE system on children's growth trajectories at the state level. Given the differences in purpose and approach, any strategy may be implemented on its own or in combination with one or more of the other four strategies.

The following provides a brief summary of the five strategies, before we delve into a more in-depth discussion of each approach:

- With Approach A, labeled **Caregiver/Teacher- or Program-Driven Assessments to Improve Practice**, individual caregivers or teachers are trained as part of their formal education or ongoing professional development to use developmentally appropriate

[1] Elicker and Thornburg (2011) discuss three ways in which measures of child development may be used in QRIS evaluations; these uses differ to some extent from our framework. First, they note that such assessments may be used to provide descriptive information about the children enrolled in the QRIS. Such information can help planners determine whether the QRIS is reaching targeted populations. A second purpose, similar to our Approach D, is to use child assessments to validate a QRIS. Third, child outcomes may be used to determine how well a QRIS is working; this use assumes that the QRIS is an intervention and child outcomes a measure of its success. This use is somewhat like our Approach E, described below, although Approach E addresses the effect of a given ECE program or the ECE system as a whole rather than the effect of the QRIS.

Table 3.1
Five Approaches to Incorporating Assessments of Child Functioning into State QI Efforts

Approach	Description and Purpose	Focus	Relationship to QRIS
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Expectation of use of child assessments by caregivers/teachers to inform caregiving and instructional practice with individual children and to identify needs for staff professional development and other program quality enhancements	**Individual child** Assess developmental progress Apply differentiated instruction **Classroom/group or program** Identify areas for improved practice Determine guidance for technical assistance	Not explicitly incorporated into QRIS; can be focus of best practice in teacher preparation programs, ongoing professional development, and provider supervision; can be a requirement of licensing, program regulation, or accreditation
B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice	QRIS requires demonstrated use of child assessments by caregivers/teachers to inform caregiving and instructional practice with individual children and to identify needs for staff professional development and other program quality enhancements	Same as Approach A	QRIS rating element specifically assesses this component alone or in combination with other related practice elements
C: Independent Measurement of Child Outcomes to Assess Programs	Independent assessors measure changes in child functioning at the classroom/group or program level to assess program effects on child development or to assess the effectiveness of technical assistance or other interventions	**Classroom/group or program** Estimate value added Assess technical assistance effectiveness or other interventions	QRIS rating element is based on estimates of effects at the classroom/group or program level
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	Independent assessors measure changes in child functioning to validate QRIS design (i.e., to determine if higher QRIS ratings are associated with better child developmental outcomes)	**Statewide QRIS** Assess validity of the rating portion of QRIS	Part of (one-time or periodic) QRIS evaluation
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Independent assessors measure child functioning to evaluate causal effects of specific ECE programs or groups of programs on child developmental outcomes at the state level	**Statewide ECE system or specific programs in system** Estimate causal effects of ECE programs	Part of QRIS (ongoing) quality assurance processes or, when no QRIS exists, part of evaluation of state ECE system

SOURCE: Authors' analysis.

assessments to evaluate each child in his or her care.[2] Caregivers and teachers then use the results of the assessments to set future developmental goals for each child and to tailor their activities and interactions with each child to promote development in those areas identified as needing additional attention. In addition, program leadership may aggregate

[2] As noted in Chapter Two, child development assessments may include a combination of structured observation, portfolios, or other tools to track children's developmental progress and individualize curriculum and instruction. Such assessments are different from developmental screening tools for purposes of identifying children who may have developmental disabilities (Tout et al., 2010b).

the assessment results to the classroom or program level to improve practice and identify needs for professional development or other quality enhancements. This approach does not assume a formal link to a QRIS but rather that the use of child assessments is part of standard practice as taught in teacher preparation programs or other, ongoing professional development programs (e.g., coaching or mentoring) and as reinforced through provider supervision (e.g., by a center director or through a network of family child care providers). The practice of using child assessments may also be stipulated in licensing requirements, program regulations, or accreditation standards.

- Approach B, labeled **QRIS-Required Caregiver/Teacher Assessments to Improve Practice**, has the same purpose as Approach A, but it has an explicit link to a QRIS. In this approach, demonstrated use of assessments of child functioning to inform the approach a caregiver or teacher takes with an individual child, as well as efforts to improve program quality through professional development, technical assistance, or other strategies, is required to reach specified rating levels, usually the higher ones.
- For Approach C, labeled **Independent Measurement of Child Outcomes to Assess Programs**, the link between the QRIS and child developmental outcomes is even more explicit. In this case, the measurement of *changes* over time in child functioning at the classroom, group, or center level can be either an additional quality element incorporated into the QRIS rating or a supplement to the information summarized in the QRIS rating. Typically, these analyses attempt to control for differences in children's backgrounds as well as classroom and teacher characteristics in predicting changes over time (see, for example, Zellman et al., 2008). Value-added modeling (VAM), a method quickly gaining popularity in studies of K–12 education, is a more sophisticated approach to doing this. Complex VAMs attempt to isolate the contributions of teachers or schools to student performance (McCaffrey et al., 2004). VAM measures the effect of a specific caregiver/teacher or a specific program on the developmental progress of the students in that setting. The analyses used to assess program or teacher effects can also be used to assess the effectiveness of technical assistance or other professional development interventions.
- Approach D, labeled **Independent Measurement of Child Outcomes to Assess QRIS Validity**, collects child assessment data to address macro-level questions, in this case, the validity of the rating portion of the QRIS. As discussed above, validation is a process that assesses the degree to which evidence and theory support the conclusions and interpretations derived from multicomponent assessments conducted in a specified context. For QRISs, the logic model asserts that higher-quality care will be associated with better child outcomes. Therefore, one important piece of validation evidence concerns whether programs that score higher on the QRIS rating scale are associated with better child outcomes. In other words, through a one-time or periodic evaluation, this strategy aims to determine if higher program ratings, which are largely based on program inputs, are positively correlated with better child outcomes, the ultimate QRIS goal.
- Approach E, labeled **Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System**, also takes a macro perspective, but it differs from Approach D in using rigorous methods that enable an assessment of the causal effects of a statewide ECE program or group of programs on child developmental outcomes. Such an evaluation of effects may be a required one-time evaluation or periodic quality assurance component for the QRIS. For states without a QRIS, such an evalua-

tion may be conducted to measure the effect of specific publicly funded programs or the state's early care and learning system more generally.

We now discuss each of these approaches in more depth, drawing on the additional detail about measurement and methods for each approach summarized in Table 3.2. As part of the discussion, we identify (1) where the approach is currently in use, if it has been implemented, as well as associated best practices or lessons learned; (2) the resources that would be required for the approach to yield useful and valid information; (3) the benefit that would accrue from making the investment in the associated data gathering and analysis; and (4) threats to the validity of the approach and possible strategies for mitigating those risks. These points are summarized in Table 3.3.

Table 3.2
Measurement Details and Analysis Methods for Each Approach to Incorporating Child Assessments

Approach	Who Is Assessed	Who Conducts the Assessment	What Is Assessed	Frequency of Assessment	Analysis Methods
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Each child in the group/classroom	Caregiver or teacher	Multiple specific domains of development using recognized assessment methods	As needed to support the child's developmental progress	Caregiver/teacher tracks progress over time in manual or automated system; results are reported to parents in parent-teacher conferences and to program leadership to target professional development/program improvement
B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice	Same as Approach A	Same as Approach A	Same as Approach A	Same as Approach A	Same as Approach A
C: Independent Measurement of Child Outcomes to Assess Programs	For group/classroom Impact: all children in group/classroom For program impact: random sample of children in each group/classroom	Trained, reliable independent assessor	Same as Approach A and child/family background characteristics (and possibly caregiver/teacher characteristics) to serve as controls	At least start of year (baseline) and end of year (follow-up)	Evaluator conducts analysis (e.g., VAM) that produces estimates of effects, which may be input into program rating
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	All children or a random sample of children in a random sample of programs across the state participating in the QRIS	Same as Approach C	Same as Approach C	Same as Approach C	Evaluator examines the statistical relationship between QRIS rating and child outcomes
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Random sample of children in relevant programs across the state	Same as Approach C	Same as Approach C	Depends on design	Evaluator estimates causal effect of ECE system or specific programs on child outcomes

Approach A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice

As noted above, Approach A takes a micro perspective—the level of an individual child in a group or classroom or the level of a classroom or program —and is not predicated on the existence of a QRIS. Instead, the use of assessments of child functioning by caregivers and teachers to track child progress and support program improvement is motivated simply by the expectation that this is good or standard practice in the field. As shown in Table 3.2, the approach assumes that each child in the group or classroom is assessed by a well-trained caregiver or teacher, regardless of the child's current or prior circumstances (e.g., children with developmental delays are assessed, as well as children showing advanced development). Children are assessed across multiple specific domains of development (e.g., social, emotional, cognitive, physical) using recognized assessment methods (some of which may be tailored to the curriculum being used), where the frequency of assessment is appropriate for supporting the child's developmental progress. The resulting information at a point in time and over time (which may be recorded manually or in an automated system) is used by caregivers or teachers to inform their work with the child. The results are reported to the parents as well, typically through parent-caregiver/teacher conferences. In addition, program administrators can aggregate the results within or across classrooms or groups to identify needs for professional development or program improvement.

Current Practice

In many respects, what we have labeled as Approach A is recognized as good practice in the provision of ECE (although implementation may not always occur or be done well). Child assessments have long been recognized by the child development field as a critical tool for monitoring children's development, both in support of further development and in the application of a particular curriculum (Bowman, Donovan, and Burns, 2001). Assessments can also support program improvement. No single approach is viewed as superior, and multiple approaches are usually recommended by experts.

The value placed on conducting child assessments as part of effective practice is reinforced through the accreditation standards defined by NAEYC, the premier accreditation body for early childhood programs. Among the ten accreditation standards, the fourth pertains to "Assessment of Child Progress" and states that "the program is informed by ongoing, systematic, formal, and informal assessment approaches to provide information on children's learning and development" (NAEYC, 2008, p. 2). The standard notes that assessments are used to inform decisions about children, teaching, and program improvement. With regard to children, the assessments may inform the need for more intensive instruction or for further developmental evaluation.

The use of child assessments is also recognized as part of standard practice through accreditation standards for teacher preparation programs. Since 2006, NAEYC has been accrediting associate degree teacher preparation programs focused on early childhood. Among the seven NAEYC core (or "initial") accreditation standards set forth by the NAEYC Commission on Early Childhood Associate Degree Accreditation, the third standard pertains to "observing, documenting, and assessing to support young children and families" (NAEYC, 2010). The standard is predicated on the understanding "that child observation, documentation, and other

Table 3.3
Additional Features of Each Approach to Incorporating Child Assessments

Approach	Current Use	Potential Resources Required	Potential Benefits	Potential Barriers to Success	Strategies for Mitigation
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Incorporated into NAEYC ECE program accreditation criteria Incorporated into NAEYC teacher training program accreditation criteria	Identification of appropriate assessment tools Training for caregivers/teachers and supervisor in value of and use of assessment tool(s) Time for caregivers/teachers to administer assessments and record results Recording forms or database for inputting assessment results and analyzing data at child, classroom, and program levels Parent conference policy and time to communicate results to parents Time for program administrators to analyze aggregate results and craft a program improvement plan specifying areas for staff development and other quality improvements	Interactions and instructional activities can be tailored to meet the individual child's developmental needs Handicapping conditions are identified early Parents are informed about their child's developmental progress and can integrate activities at home that target areas where development is lagging Assessments aggregated to classroom or program level can identify areas for staff professional development, technical assistance, and other program improvement	Inadequate training in the use of assessment tools Caregivers/teachers do not have the needed time or classroom environment to conduct proper assessments	Ensure that those who complete training courses or programs achieve competency and periodically reassess competency Address other aspects of work environment through regulations, the QRIS, or classroom supervision
B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice	**Center QRIS**: CA[a], DE, FL[b], FL[c], LA, ME, MN, MS, NM, OH, PA **FCCH QRIS:** CA[a], DE, FL[b], ME, MN, NM, OH, PA	Same as Approach A, as well as: Possible development of new tool Review, and approve recommended assessment tool(s) or database software Training for caregivers/teachers and supervisors in use of any required assessment tool(s) and database	Same as Approach A, as well as: Increased use of assessments	Same as Approach A	Same as Approach A

Table 3.3—Continued

Approach	Current Use	Potential Resources Required	Potential Benefits	Potential Barriers to Success	Strategies for Mitigation
C: Independent Measurement of Child Outcomes to Assess Programs	None	Identification of appropriate assessment tools Training of independent assessors to reach reliability and regular retraining to maintain reliability Time for assessors to administer assessments, collect other data, and record results Time for evaluator to analyze classroom or program effects	QRIS rating includes estimates of classroom or program effects on child outcomes in program ratings	Difficulties associated with assessing young children reliably Data collection requirements are costly Analytic methods do not control for confounding factors so estimates of effects are biased	Ensure reliability of assessors through rigorous training and regular testing Assess programs less frequently Plan to use rigorous methods to address methodological issues and subject design and findings to peer review
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	**Completed:** CO, MO **Planned:** IN, MN, VA	Identification of appropriate assessment tools Training of independent assessors to reach reliability and regular retraining during study to maintain reliability Time for assessors to administer assessments, collect other data, and record results Time for evaluator to design study, manage data collection, and conduct analyses	Evaluation determines if QRIS is measuring dimensions of ECE program quality that are important for different domains of child development	Difficulties associated with assessing young children reliably Assessment instruments may not capture those aspects of child development affected by high-quality ECE Other methodological issues may bias statistical inference Study may not have adequate resources to implement rigorous design	Ensure reliability of assessors through rigorous training and ongoing testing Plan to assess multiple domains of child functioning and use well-validated instruments Plan to use rigorous methods to address methodological issues and subject design and findings to peer review Ensure adequate resources for all phases of study design, including sufficient sample sizes to account for attrition and subgroup analyses and measurement of key family characteristics
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Completed or ongoing: AR, CA, MI, NJ, NM, OK, SC, WV	Same as Approach D	Evaluation determines if participation in ECE program or programs affects child development in a range of domains	Same as Approach D	Same as Approach D

SOURCE: Authors' analysis of information in Tout et al. (2010b).

NOTE: FCCH is family child care home.

[a] QRIS is for Los Angeles County.

[b] QRIS is for Miami-Dade County.

[c] QRIS is for Palm Beach County.

forms of assessment are central to the practice of all early childhood professionals" (NAEYC, 2010, p. 32). Key elements of the standard include

- understanding the goals, benefits, and uses of assessments to set goals, incorporate a curriculum, and employ teaching strategies
- knowing how to use assessments to form partnerships with families and professional colleagues
- knowing how to use assessment tools that involve observation, documentation, and other techniques, including the use of technology for documentation
- understanding how to practice responsible assessment with each child, including the use of assistive technology for children with disabilities.

The NAEYC standards are also incorporated into the accreditation of early childhood baccalaureate and graduate degree programs through the partnership between NAEYC and the National Council for Accreditation of Teacher Education.

Employing child assessments may also be dictated as part of the regulatory requirements for specific early childhood programs. For example, the federal Head Start Program Performance Standards require ongoing child developmental assessments to chart progress and plan program activities (U.S. Department of Health and Human Services, undated). In California, although Title 22 licensing standards for centers and family child care homes are silent with respect to child assessments, the regulations governing Title 5 child development programs, including the California State Preschool Program, specify that programs use the Desired Results Developmental Profile (DRDP) (California Department of Education [CDE], 2008b).[3] The DRDP is a research-based observational assessment tool administered by the caregiver or teacher that measures a child's progress toward the set of milestones outlined in the DRDP system, which itself is aligned with the state's early learning standards and the Head Start Child Outcome Framework (CDE and Center for Child and Family Studies at WestEd, undated; CDE, 2010a). The regulations require the use of the DRDP within 60 days of enrollment and every six months thereafter to "plan and conduct age and developmentally appropriate activities." The Title 5 regulations also specify the use of accommodations for children with exceptional needs. Although not required to do so, many programs throughout California not covered by Title 5 or Head Start regulations are implementing the DRDP as part of their routine practice (CDE, 2009a).

The CDE Child Development Division (CDD) provides support for the use of the DRDP through the Desired Results Training and Technical Assistance Project, which is funded with federal Child Care and Development Fund quality improvement dollars. The training program targets administrators and teachers in CDE-administered Title 5 contract centers and family child care homes. As of state fiscal year 2009–10, about 500 contractors were expected to participate in the three-day training program each year.

Recent survey data for California indicate that the use of child assessments is near universal, at least for preschool-age children in center-based settings. In the teacher survey component of the RAND California Preschool Study, Karoly et al. (2008) estimated that most three- and four-year-old children in center-based settings were routinely assessed, using either ratings

[3] There are several different DRDP instruments and they have been revised over time. As of 2011, infant/toddler, preschool, and school-age instruments were available. A school readiness version is under development.

based on observation or work sampling (47 percent), standardized tests or assessment instruments (6 percent), or a combination of observation and direct assessment (42 percent). Just 5 percent of children in center-based programs were in a classroom where the teacher reported only informal assessment or no assessment method. One limitation of these data is that they do not indicate if the assessments are implemented well or used appropriately.

Resources Required

As indicated in Table 3.3, the use of Approach A requires identification by program administrators of appropriate assessment tools and the resources to train the caregivers, teachers, and supervisors to use the identified tools effectively. Those employing the assessments need to understand the benefits of tracking children's developmental progress for their own practice and for communicating with parents. They also need the general training associated with observing and measuring children's development across the array of relevant domains—social, emotional, cognitive, and physical—as well as any specialized training associated with the use of particular assessment tools. Program administrators will also need additional training in the use of the assessment results to identify needs for staff development and program improvement, possibly through the use of database software.

Beyond the required training, another key resource required is the time it takes caregivers and teachers to conduct the child assessments and record and analyze the results (whether on paper forms or in an automated database). A parent conference is a critical piece of the effort as well; in these conferences, caregivers and teachers take the time to communicate assessment findings to parents and engage them in their children's developmental progress. This latter effort may require that programs develop activities or resources that parents can use at home to support their children's progress. The time required for conducting the assessments and communicating with parents is nontrivial and increases linearly with the number of children in the group, the frequency with which assessments are conducted, and the number of parent conferences held. Time is also required for program administrators to analyze aggregate results and develop a program improvement plan that specifies needs for staff development and other program enhancements.[4]

Expected Benefits

The use of child development assessment as part of standard practice and its incorporation into various accreditation standards reflect the expected benefits from evaluating the progress of young children in multiple domains of development (Bowman, Donovan, and Burns, 2001). By identifying where a child's development may be trailing given his or her stage of development, caregivers or teachers can provide more individualized support in their interactions with the child and in the use of the curriculum and other activities. Further, through routine assessments, possible developmental delays may be identified earlier, with the possibility that targeted interventions will be more successful when started sooner rather than later. When the results of assessments are routinely communicated to parents through conferences and other mechanisms, parents are informed about their child's progress and also can implement developmentally appropriate activities at home to reinforce the targeted supports provided by the

[4] Snow and Van Hemel (2008) note that since assessment activities typically deflect time and resources from instruction, it is important to ensure that the value of the information collected outweighs the costs—in terms of both possible unpleasantness for children and time and money spent.

caregiver or teacher in the early care setting. Finally, aggregated results on child developmental progress at the classroom or program level can be used to identify areas where individual caregivers or teachers need improvement or broader needs for program-wide technical assistance or quality upgrading.

Potential Barriers to Success and Strategies for Mitigation

To be successful, Approach A requires that caregivers and teachers be effective in their use of the assessment tool(s), both in administering the tool and in interpreting the results. The same is true for program administrators who use the tool to identify needs for program improvement. If training is inadequate, either as part of postsecondary degree programs or as part of specific trainings, the use of assessments by caregivers/teachers will be less reliable and their communication to parents will be less valuable. The aggregate results at the classroom or program level may also be misleading if the assessments are not conducted in a consistent way by caregivers and teachers. Thus, effective training is a critical element. To ensure that training programs are effective, practitioners should be assessed at the end of the training to determine if they have achieved competency with the tool. It is important as well to plan for periodic redetermination of competency, including reliability, perhaps administered through a program director or other supervisor.

Given the time to conduct effective assessments and the nature of the assessment tool, it is also important for caregivers and teachers to be in a supportive environment in the classroom or home-based setting. Thus, if the group size is too large relative to the number of available adults or if management of the group of children is poor, it may be challenging for the assessor to concentrate on evaluation of a particular child. Ensuring that the ECE environment is supportive can potentially be addressed through licensing or regulatory standards regarding group size and ratios, through attention to these structural features in the rating portion of the QRIS or through supervision and supports.

Approach B: QRIS-Required Caregiver/Teacher Assessments to Improve Practice

What we have labeled Approach B is an extension of Approach A, where now the assessments of child functioning and the demonstrated use of those assessments to inform classroom practices, staff development and other quality enhancements is *a required element to achieve specified QRIS rating levels*. As shown in Table 3.2, Approach B is defined to be essentially identical to Approach A in terms of who is measured, how measurement is conducted, what is measured, the frequency of measurement, and how assessments are used.

Current Practice

Some form of Approach B has been incorporated into the QRISs of several states and areas.[5] According to Tout et al. (2010b), 11 of the 26 QRISs they studied as of 2009 incorporated

[5] It is worth noting that the use of child assessments is not a program feature that is evaluated as part of the most commonly used early childhood environment rating scales, such as the ECERS-R, the Family Day Care Rating Scale (FDCRS), or the ITERS. For this reason, although most state QRISs incorporate these scales into their rating, it is necessary to add a separate element for child assessments if that is a criterion that states want to incorporate into their quality standards.

an indicator regarding the use of child assessments among the rating components of their QRIS for center-based programs, and eight systems had such an indicator in the rating system for family child care homes (see Table 3.3). Of the 11 center-based systems, four (Colorado, Minnesota, Ohio, and Pennsylvania) required that assessment results be shared with parents. Three of the 11 center-based systems (California–Los Angeles, Minnesota, and Pennsylvania) reported that they have a process for reviewing the child assessment tools, and seven systems (California–Los Angeles, Florida-Palm Beach, Louisiana, Minnesota, Mississippi, Ohio, and Pennsylvania) required that programs use an approved tool. The comparable figures for the eight family child care systems were four (assessments are shared with parents), two (review process exists), and four (approved assessment tools required).[6] In some state QRISs, required child assessments also have a developmental screening purpose, in addition to using assessments to inform classroom practice and program improvement.

Existing systems with an assessment requirement differ in terms of the rating tier in which child assessments are first required. Some systems mandate the use of assessments starting at level 2 of 5 or level 3 of 5, but others do not introduce the requirement until a program reaches one of the top tiers. Most states do not vary the nature of the assessment requirement across applicable tiers. One exception is Pennsylvania where, to reach the second of four tiers, a center-based program must complete a developmentally appropriate screening of the child that is shared with parents within 45 days of program entry (Pennsylvania Office of Child Development and Early Learning, 2010). At the third tier, in addition to the second tier requirement, a "developmentally appropriate authentic assessment" must be completed on a designated time line, shared with the family three times per year, recorded in a statewide online system, and used "for curriculum, individual child planning, and referral to community resources."[7]

The QRIS proposed by the CAELQIS Advisory Committee (2010a, 2010b) would incorporate child assessments as part of the Teaching and Learning domain, specifically the component that requires alignment with the state's Early Learning Foundations and Frameworks (CDE, 2008a, 2009b, 2010b). As proposed, the rating system adds a requirement in the third through fifth tiers (out of five total) to use "developmentally, culturally, linguistically appropriate" child assessments tied to lesson plans in the social, emotional, cognitive, and physical domains (CAELQIS Advisory Committee, 2010b). The proposed QRIS design does not reference the use of such assessments at the classroom or program level for targeting professional development or other program improvements.

Resources Required

As shown in Table 3.3, Approach B potentially requires some additional resources beyond those enumerated for Approach A. First, as noted above, a number of states have chosen to specify one or more recommended child assessment tools to be used in mandated assessments. Selecting these tools requires a process for clearly specifying assessment goals, reviewing avail-

[6] Examples of approved tools include the Ages and Stages Questionnaire (ASQ), the Brief Infant-Toddler Social Emotional Assessment, the Early Childhood Screening Assessment, the Ounce Scale, the Preschool Kindergarten Behavior Scale, and the assessment tools associated with specific ECE curricula such as Creative Curriculum and High/Scope (Tout et al., 2010b). Some of these tools (e.g., ASQ) are classified as developmental screening tools by Halle et al. (2011).

[7] The state's QRIS requirements with respect to the child assessment rating element has varied over time. The discussion here reflects the standards as of 2010–2011, which differ from the standards that applied at the time information was collected by Tout et al. (2010b).

able measures, and selecting the tools that will be recommended or required. States may also wish to develop their own tools, consistent with their early learning standards, as was done in California in the case of the DRDP. This is obviously an even more resource-intensive activity given the need to develop, test, and validate a new tool, a process that can take several years.[8]

Second, by formalizing the use of child assessments through the rating portion of the QRIS, it may become apparent that there is some benefit to developing software for a database that allows caregivers and teachers to input results for individual children and then readily analyze patterns for a given child or across children, ideally with easy-to-use graphical interfaces. The same software tool can be used by administrators to analyze results for groups of children. An online tool, such as the one used in Pennsylvania, also allows for analysis at even more aggregated levels. Development of such a tool can require substantial time and resources.

Third, in addition to more general training of caregivers, teachers, and supervisors on the value of and use of child assessments required in Approach A, if the QRIS specifies the acceptable assessment tools, providers will need additional training to achieve proficiency with the required tool(s).

Expected Benefits

All of the benefits discussed in the context of Approach A also apply to this approach, namely, differentiated instruction, early detection of developmental delays, well-informed parents who engage in developmentally supportive at-home activities, and data to inform staff development and program improvement. By incorporating the use of child assessments into a QRIS requirement, it is possible that compliance with the practice would increase compared with the voluntary approach in A. Caregivers and teachers may also be more effective in their use of assessments if the QRIS places an emphasis on the quality of implementation.

Potential Barriers to Success and Strategies for Mitigation

As with Approach A, this approach is less effective if caregivers and teachers are not well trained in the use of child developmental assessments more generally and any required tools more specifically. Having the required time and a supportive ECE environment is important as well. Thus, as with Approach A, it is essential to ensure that staff, including program administrators, are well versed in the use of assessment tools and the interpretation of results. Post-training evaluations are key, along with periodic refresher trainings and redeterminations of competency. Further, when appropriate given the nature of the assessment tool, it is necessary to ensure that ECE classrooms or home settings have adequate ratios and are well managed so that child assessments can proceed in a setting that accommodates the needs of the assessor.

Approach C: Independent Measurement of Child Outcomes to Assess Programs

In moving from Approaches A and B to the remaining three approaches, we transition from strategies that center on child *assessments* to those that focus on child *outcomes*. As noted in

[8] The process of identifying or developing ECE assessment tools under Approach B may be coordinated and aligned with efforts to develop a kindergarten readiness assessment tool. Assessing children's developmental status at kindergarten entry is increasingly part of state kindergarten school readiness practices (Daily, Burkhauser, and Halle, 2010).

Chapter One, the use of the term *outcome* implies that we are interested in measuring the effect of some program or intervention on child functioning, as measured by a given assessment tool. Given this assessment objective, we also now shift from assessments conducted by caregivers or teachers in the child's group or classroom to assessments conducted by well-trained, independent assessors. This increases the reliability of the assessments within and across settings, which is critical for making valid inferences about causal effects based on the assessment results.

In the case of Approach C, we still take a more micro perspective, as we do in Approaches A and B, in that our interest is in the effect of a given ECE setting—a classroom or a program—on the functioning of the children in that setting. As with Approach B, Approach C is incorporated into the rating structure of the QRIS. However, given that the assessment goal is to identify the contribution of a classroom or program environment to a child's developmental trajectory, the analytic requirements of Approach C are considerably more demanding than those employed in Approaches A and B. That is because Approach C, in seeking to isolate the specific contribution of the ECE program or another intervention, must identify and control for all other possible factors that might affect child development. As noted in Table 3.2, Approach C differs from Approaches A and B on each of the measurement and method requirements. First, although Approaches A and B assume that all children will be assessed, depending on the nature of the evaluation question being asked, Approach C may require assessments of only a sample of children. This would be the case, for example, if the goal were to measure the average effect of the ECE program on child outcomes across all children in that program. Depending on the number of children in the program, it might be possible to obtain sufficient statistical power with assessments on a random sample of the children. This is less likely to be the case if the goal is to measure effects on child outcomes at the level of an individual classroom or group.[9]

Second, as already noted, to be rigorous, Approach C requires that the child assessments be conducted by well-trained independent assessors who have met a standard for reliability and who undergo periodic retraining to ensure that they remain reliable in the use of the assessment tool. Typically, child functioning would be assessed at the start and end of a program year (e.g., the fall and spring for academic-year programs), so that it is possible to take the child's initial level of functioning into account and assess changes over time.

Third, to be most valid, other factors that can influence child developmental outcomes must be statistically controlled. This means measuring child and family background characteristics (e.g., parental education and family income) to account for potential selectivity of children into programs. If the goal is to measure the effect of a particular intervention, the measurement of caregiver/teacher characteristics may also be important to account for how those characteristics (e.g., teacher education) mediate the effect of the intervention.

Finally, in terms of analytic methods, Approach C would be expected to use an independent evaluator to analyze the child assessment data and report estimates of classroom or program effects. The resulting estimates of effects would then contribute to how a program is rated according to the structure of the rating portion of the QRIS.

[9] Shepard, Kagan, and Wurtz (1998) note that direct measures of learning outcomes for three- and four-year-olds can be developed and used in large-scale program evaluations, such as Head Start, Even Start, and Title I in the preschool years but must be administered under controlled conditions and use matrix sampling. Results should not be reported for individual children.

Current Practice

To our knowledge, Approach C has not been used to date in state or local QRISs, on either a pilot or large-scale basis. As noted above, Approach C is comparable to approaches currently being used in K–12 education to assess the contribution of teachers in a given classroom or the effects of a particular school on student achievement. Thus, it is instructive to review how these methods, particularly VAM, are being used in the K–12 context and the methodological issues that arise from their use.

VAM is a collection of statistical techniques employed in the K–12 context that uses multiple years of student achievement test score data to estimate the unique contributions of the school or teacher over the course of a year rather than the cumulative effects of education or student background factors. The teacher's effect may be defined as *the average causal effect on student achievement across all students of interest.* Estimation of the teacher's effect typically entails subtracting the achievement test scores of a teacher's students at the beginning of the year from their score at the end of the same year, adjusting statistically to account for the effects of student background or school-level factors outside the teacher's control (McCaffrey, Koretz, et al., 2004). Unlike traditional methods, VAM analyses implicitly hold student background factors constant (Buddin, 2011). The focus on improvement over the course of a year has meant that some schools that have consistently been rated as excellent using traditional methods have ranked far lower using VAM analyses because they are not contributing substantially to the improvement of their students, who come to the school from supportive family backgrounds and with a history of high achievement. VAM has recently attracted a great deal of attention among both researchers and policymakers (McCaffrey, Koretz, et al., 2004). Two aspects of VAM are particularly appealing. First, VAM is theoretically able to assess the separate effects of teachers and schools and of family background on student performance. This enables schools and teachers to be rated in terms of what they have accomplished. It neither rewards schools that attract students with strong cognitive skills and supportive family backgrounds nor penalizes schools that attract students with fewer skills and less support at home. Second, some recent VAM studies have found very large differences among teachers in their effectiveness (Glazerman et al., 2010). If these differences can be substantiated and then linked to specific teacher characteristics, important improvements in education could be made by basing hiring decisions and professional development policies on them (McCaffrey, Koretz, et al., 2004; Buddin, 2011).

To conduct VAM analyses, student achievement test scores must be available for at least two points in time, typically one year apart. Although NCLB requires such testing and the cataloguing and linking of such scores, at least some school districts are not able to do this well. VAM analyses are complex; a high level of statistical sophistication is necessary to carry them out. Many schools districts do not possess such capacity. Political will is necessary to conduct the analyses and apply them in considering policy. Ensuring such will is challenging. Not surprisingly, teacher unions and teachers have reacted with concern to these approaches and have noted the many statistical issues associated with this approach, some of which are noted below. Many question the emphasis on student test scores as a way to assess individual teachers given the many challenges associated with producing reliable and valid measures of student performance (e.g., Hannaway and Hamilton, 2008). And, the fact that test scores are available in only a small number of subjects increase these concerns. Certainly, use of VAM will require additional training for principals in the use of the results and in how to integrate them with other measures of teacher performance. Parent education is also required; parents must

be helped to understand that VAM analyses may represent only one of several assessments of teacher and school performance.

Despite the obvious intuitive appeal of VAM, its use raises a number of important statistical and psychometric issues. McCaffrey, Koretz, et al. (2004) identify and discuss four categories, including basic issues of statistical modeling, issues involving confounding and omitted variables and missing data, issues arising from the use of achievement test scores as dependent measures, and uncertainty about estimated effects. Many are arcane, but if they are not addressed, VAM is likely to produce incorrect estimates of school and teacher effectiveness, which could hamper efforts to improve education and give VAM an undeserved bad name (McCaffrey, Koretz, et al., 2004). Yet the consensus among researchers is that the technique has value and should not be discarded because of statistical and political concerns. Indeed, Glazerman et al. (2010) note that all teacher evaluation approaches have methodological and practical flaws that must be considered when VAM's limitations are highlighted.

The precise specification and estimation of value added models is not straightforward (Reardon and Raudenbush, 2008). VAM, like most statistical models, will produce unbiased or consistent estimates of a particular effect only when certain untestable assumptions hold. The particular assumptions being employed for estimating a particular effect should be identified and evaluated for plausibility and formally tested where possible; the effect of violations of these assumptions will depend on the desired effect and the particular model used (see McCaffrey, Koretz, et al., 2004, for a discussion of these assumptions). To refine the use of VAM, several important statistical and psychometric issues need to be addressed including the sensitivity of value-added measures to various controls for student characteristics and classroom peers and the reliability of value-added measures over time. For example, if assessments of teacher and school effectiveness vary substantially from year to year, then value-added estimates will not be helpful in identifying the factors that appear to improve student learning. Alternative specifications of the statistical model used in the analyses may also produce substantially different estimates of teacher and school effects (Buddin, 2011). Currently, the research base does not justify the use of VAM in high-stakes decisions (McCaffrey, Koretz, et al., 2004). Other barriers in the K–12 context include the fact that student achievement tests are not administered until second grade, so the measures provide no indication of the effectiveness of kindergarten or first grade teachers. Second, most districts test annually only in the subjects required in NCLB: English language arts and math. Although these subjects are clearly foundational, tests in just two subjects do not provide a comprehensive indication of what students have learned. In addition, standardized tests are imperfect measures of learning because students may misunderstand what is expected or because individual students may have test anxiety or other issues on the day of the test.

Use of VAM in ECE settings involves all of the challenges encountered in grades 2–12 as well as additional ones unique to ECE settings. As noted in Chapter Two, assessments of young children are difficult, costly, and less reliable than those of older children; this is one reason why NCLB does not include kindergarten or first grade students. Relying on these assessments to evaluate ECE caregivers or teachers or even the contribution of ECE programs to child development is problematic at best, especially given that child functioning is a multidimensional concept and cannot be readily summarized in a single assessment tool. Certainly, it would be critical to validate measures that might be used in this high-stakes way before implementing such analyses. If the method were used to assess the performance of individual caregivers or teachers, the typical practice in ECE settings of having multiple staff in a classroom

or group and the high staff turnover rates in many settings mean that it would be challenging to identify the contribution of specific individuals to a child's developmental progress. At best, it may be possible to estimate the effect of a child being in a particular classroom or group (without reference to the staff in that setting) or of being in a particular program. However, with the movement of some center-based children from one classroom to another in the course of a year (if they "age out" into the next level of care, for example) or the mobility of some children across different programs, even this type of assessment is problematic. Finally, average ECE class or group size also presents serious problems. Some of the statistical imprecision in assessing students in grades 2–12 is assumed to "average out" across students in a classroom or school (Buddin, 2011). Such averaging is far more difficult to assume and achieve in ECE settings, where class and group sizes are far smaller.

Resources Required

Table 3.3 summarizes the potential resources required to implement Approach C. First, it is necessary to identify the appropriate assessment tools given the purpose for which they will be used. For example, the assessments used in Approaches A and B, which are usually those designed to inform practice and program improvement, may not be appropriate for an analysis that seeks to identify the causal effect on child functioning and incorporate that information into ECE program ratings. The use of assessments in high-stakes circumstances, such as the use of achievement tests to evaluate teachers or schools in K–12 education, places a much higher burden on the assessment tool and on the assessor (Bowman, Donovan, and Burns, 2001). Standards of evidence for the tool's psychometric properties and its appropriateness for use with children of the different ages and backgrounds being assessed must be very high (Snow and Van Hemel, 2008). Thus, it is important that the assessments used are well validated and that the assessors are thoroughly trained to meet a high standard of reliability.

The most resource-intensive part of Approach C is the time required for assessors to administer the assessment tool or tools. The relevant tools, without exception, will require that the assessor work one-on-one with the child being assessed, and when multiple tools are used to capture different domains of functioning, the time required may be measured in hours rather than minutes. The need to repeat assessments so that gains over time are measured further adds to the time burden. And since child or family background information is required, resources may also need to be allocated to the design and administration of a parent questionnaire.

Once the assessments have been conducted and other data collected, resources are required for an evaluator, presumably independent from the program (which is not always the case in the K–12 context), to conduct the appropriate analyses to measure classroom-, group-, or program-level effects. The analytic resources are likely to be most intensive at the outset when the method is developed for estimating classroom or program effects on child developmental outcomes. Ideally, that process would involve extensive sensitivity testing to determine how robust the estimated effects are to methodological choices such as model specification, the use of specific combinations of control variables, and so on.

Expected Benefits

The primary motivation for Approach C is to more directly incorporate child outcomes into QRIS ratings. Instead of relying solely on measured inputs to capture program quality and calculate ratings, Approach C has the potential to capture the outcome of interest—ECE program effects on child functioning—and to use the results when rating programs. In principle,

if Approach C could be implemented with confidence, program ratings could be based exclusively on the results of the estimates of effects of a given ECE program on child outcomes. However, given the current state of the art for generating such estimates of effects as discussed above, this would be unwise. Instead, estimates of the effect of a particular ECE setting on child development using VAM or other methods could, at best, be one component of the QRIS rating structure, and the weight it receives in calculating ratings could reflect the confidence attached to the estimates.

Potential Barriers to Success and Strategies for Mitigation

Several factors could limit the ability to implement Approach C and validate the resulting estimates (see Table 3.3). First, if the independent assessors are not well trained or do not achieve or maintain reliability, the validity of any resulting estimates of causal impact would be compromised. The obvious solution is to ensure that assessors are well trained and reach and sustain reliability in their administration of the assessments. This may be particularly challenging if a large number of assessors is required given the number of children that need to be assessed and the frequency of assessments.

A second related point is that again, depending on the number of children who need to be assessed, the collection of assessment data to implement Approach C may be very costly. One strategy might be to conduct a classroom or program analysis of effects less frequently, such as every third or fourth year, depending on the schedule for conducting program ratings as part of the QRIS. Even so, mounting such a data collection effort in a highly reliable way would be a substantial undertaking.

Finally, the biggest obstacle to overcome in using Approach C is resolving the methodological challenges discussed above. In the K–12 context where VAM and other methods have at least some history of use, the debate is far from settled over the appropriate statistical models required to obtain valid estimates. Any given method ultimately rests on untestable assumptions, so all stakeholders must have some level of confidence in the robustness of the methodology used. At a minimum, it is necessary to perform extensive sensitivity analyses before identifying a preferred methodology; the resulting approach should be subject to rigorous peer review to provide further validation that relevant methodological issues have been addressed.

Approach D: Independent Measurement of Child Outcomes to Assess QRIS Validity

In moving to Approaches D and E, we shift again, this time from a micro focus to a macro perspective. In the case of Approach D, the goal is to use assessments of child functioning to assess the validity of a QRIS. Thus, instead of a focus on a particular classroom or program as in Approach C, we are now interested in a statewide or systemwide perspective. Validity can be assessed in many ways (see, for example, Zellman et al., 2011), but for QRISs, the central question that can be addressed using child outcomes is whether programs rated in the highest-quality tiers in the QRIS are associated with or produce larger developmental gains than lower-rated programs. This question aligns with the logic model underlying QRISs, which posits that improvements in program quality will be associated with improved child outcomes, as discussed above. Such an investigation could be conducted once in the course of the development

of the QRIS, or it could be instituted on a periodic basis once the QRIS is fully operational, tied, for example, to major revisions of the QRIS.

As shown in Table 3.2, there are some differences in measurement and methods between Approaches C and D. Although Approach D could be implemented by using assessment data for all children in programs participating in the QRIS, it is sufficient and more cost-effective to base the evaluation on assessments for children (or a sample of children) in a sample of ECE programs participating in the QRIS. In theory, any valid assessment of child functioning could be used. In practice, however, researchers conducting such evaluations tend to draw on child assessments used in other outcomes-based research studies, as discussed below, assessments that typically would not be used in Approaches A or B. Otherwise, as with Approach C, Approach D requires independent assessors, the collection of additional data on child and family characteristics, and measurement at least twice (typically in the fall and spring for academic-year programs or the fall and the following fall for year-round programs). The statistical methods involve estimating the relationship between quality ratings and child developmental outcomes, typically using the child as the unit of analysis.

Current Practice

Although a number of states have conducted or plan to conduct an evaluation of their QRIS, it is relatively rare that the evaluation incorporates child outcomes (Tout et al., 2010b). To our knowledge, just two states have conducted and released results from an Approach C–type validation study of their QRIS: Colorado and Missouri. Three other states report that they have such a validation study in process: Indiana, Minnesota, and Virginia (Tout et al., 2010b; Langill et al., 2009).

The Zellman et al. (2008) evaluation of Qualistar Colorado, a statewide QRIS first implemented in 1999, was the first to examine the relationship between the rating portion of a state QRIS and child developmental outcomes. Their sample covered 65 child care centers and 38 family child care homes. A target of 20 preschool-age children in one or more classrooms in each center and at least four preschool-age children in each home were assessed in terms of their social development, emotional development, and cognitive functioning as determined by teacher surveys and direct observation by trained assessors. Socioemotional developed was measured using the Child Behavior Inventory, which is administered to teachers, and the Strength and Difficulties Questionnaire, which is administered to parents. Cognitive functioning was assessed by direct observation using the Peabody Picture Vocabulary Test–Third Edition (PPVT-III) and three subsets of the Woodcock-Johnson–Third Edition (WJ-III) achievement test (Letter-Word Identification, Passage Comprehension, and Applied Problems). Data were collected over three waves, approximately one year apart. Extensive family background information was also collected from parents. Over 1,300 children participated in the first wave of data collection, but substantial attrition at the program and child levels meant that only 7 percent of the original sample remained by the third wave, a methodological limitation that constrained the study findings.

Although the Colorado evaluation found that quality ratings for individual programs did improve over time (consistent with the QRIS logic model), an analysis of the relationship between individual components in the rating system and child outcomes, both in cross-sectional and longitudinal models that also controlled for family and provider characteristics, showed few significant relationships (and when significant, effect sizes were small), and there was no significant relationship between the overall star ratings (ranging from one to four stars)

and child outcomes. These patterns were found for children in both center-based programs and family child care homes and also for subgroups of at-risk children defined by low family income or high doses of exposure to nonparental care.

The evaluation of Missouri's QRIS followed a similar design but produced somewhat more positive findings (Thornburg et al., 2009). The study analyzed outcomes for 350 children participating full-time (25 or more hours per week) in 66 classrooms or groups from 32 centers and six family child care homes located in three communities in the state where concentrated efforts had been made to raise ECE program quality. Preschool-age children were assessed in the fall of 2008 and spring of 2009 on a range of assessments that covered vocabulary (PPVT-IV); early literacy skills (Test of Early Reading Ability–Third Edition); math skills (WJ-III Applied Problems); basic knowledge of shapes, colors, and upper case letters; fine and gross motor skills; and socioemotional development (Devereux Early Childhood Assessment, measuring initiative, self-control, attachment, and behavior problems). Child and family background characteristics were collected through a parent survey.

The analysis showed that for all children in the sample, those in one- or two-star rated programs had significantly smaller gains (or even losses) relative to children in four- and five-star programs on most of the measures of socioemotional development, with effect sizes in the medium to high range. The gains in overall socioemotional skills were also significantly larger for those in three-star programs relative to the one- and two-star programs. There were no significant differences in the gains children experienced in low- versus high-quality and low- versus medium-quality programs on any of the other developmental domains assessed. And the contrast between medium-quality (three-star programs) and high-quality (four- and five-star programs) did not show statistically significant differences for any of the developmental domains measured. The study also examined the subsample of children in poverty and found significant differences in developmental gains across quality tiers for both socioemotional skills and vocabulary, with the largest contrasts again between the lowest-rated programs (one and two stars) and the highest-rated programs (four and five stars).

As noted above, three other states plan to conduct similar evaluations of the relationship between QRIS ratings and child developmental outcomes, with some differences in methods from the Colorado and Missouri studies. For example, the planned methodology for Indiana's validation study indicates that child assessments will cover both infants and toddlers and preschool-age children (Langill et al., 2009). Infants and toddlers will be assessed for their cognitive and language development (Mullen Scales of Early Learning) and socioemotional development (Brief Infant Toddler Social and Emotional Assessment). The domains for preschoolers are the same although the assessments differ, with measures of cognitive development (WJ-III Applied Problems and Letter Word Identification subtests), language development (PPVT-IV), and social emotional development (Social Competence and Behavior Evaluation).

The Colorado and Missouri validation studies demonstrate the importance of conducting such analyses but also some of the challenges. In the case of the Colorado study, the lack of a relationship between child developmental outcomes and either the overall quality ratings or the individual quality components suggests that the rating system may not be capturing the dimensions of ECE quality that matter most for child development or that the measures for any given quality dimensions are not sufficiently refined to accurately assess meaningful differences in program quality. Either explanation would require revising the rating portion of the QRIS to incorporate better measures of the quality constructs or to modify the quality components included in the rating system. Alternatively, the child assessments may not have

measured the relevant aspects of child development that are affected by high-quality ECE programs. It is interesting that the Missouri evaluation showed the strongest and most robust relationship between quality ratings and child development for socioemotional functioning rather than for the cognitive domain. One reason may be that the Missouri study used a different assessment tool for socioemotional functioning than the Colorado study. Thus, the decision about which domains of child development to use in a validation study and the specific assessments used to measure each domain may affect the study findings.

Other methodological issues may also affect the ability to detect statistically significant relationships. In both studies, the sample sizes were relatively small for some analyses. In the Qualistar Colorado evaluation, attrition played a role. Notably, eight out of 65 centers dropped out of the study between wave 1 and wave 2, six of them because they went out of business. Likewise, six of the 38 family child care homes also dropped out between the first two waves and another three home-based providers left the study between the second and third waves, all of which closed. In the Missouri evaluation, the sample had relatively few centers with the lowest and highest quality ratings (specifically, just one center with a one-star rating serving nine children and just one center and one home with a five-star rating serving a total of nine children). The small samples at the extremes of the rating scale meant that the analyses could contrast programs in only the two lowest tiers with programs in the middle or two highest tiers.

Another methodological concern is the ability to make inferences about the causal relationship between program quality, as measured by the QRIS rating components or the summary program ratings, and children's developmental outcomes, given that parents select providers. The ideal study design would randomly assign children to programs with different quality ratings so that children would be equivalent on both observed and unobserved child and family characteristics that might also affect their developmental trajectory. Any differences in child functioning after program participation could then be attributed to the differences in program quality. However, since an experimental design is unlikely to be feasible, the design of a QRIS validation study must contend with the potential bias from the selectivity inherent in the fact that the level of quality that a given child's provider offers is not independent of a family's background and values. For example, if children living in more supportive home environments are more likely to participate in high-quality programs, a study finding that child developmental gains are larger in high-quality programs may reflect not the causal effect of the program but rather the contribution of positive family factors, which are positively correlated with quality. The study design can try to account for possible selectivity bias by controlling for as many relevant observable family background factors as possible, but there may still be unobservable factors that cannot be controlled for and would bias the estimated effects of ECE program quality. For this reason, caution must be used in making causal statements about the role of program quality in explaining associations found between rated quality and child outcomes (Elicker and Thornburg, 2011).

Resources Required

Although many of the resource requirements for Approach D mirror those for Approach C, the overall cost would likely be substantially less because Approach D requires a one-time or periodic validation study in contrast to the ongoing data collection required for Approach C. In addition, Approach D can be implemented by assessing children in a sample of programs participating in the QRIS, as opposed to children in all programs in the rating system. As seen

in Table 3.3, as in Approach C, Approach D involves identifying the appropriate assessment tools, training the independent assessors, and collecting the child assessment and family and provider background data. Typically, the validation study design would involve at least two rounds of data collection. Resources are also required for the evaluator to develop the study design, manage the data collection process, and conduct the associated analyses.

Expected Benefits

The motivation for conducting a QRIS validation study is to determine if the theory behind the logic model holds in the context of a given state's design and implementation of its QRIS ratings. As QRISs incorporate tiered reimbursement systems and other high-stakes features, it is especially important to determine empirically if the quality ratings are associated with meaningful differences in quality that in turn affect child development. If a validation study confirms that programs with higher quality ratings produce larger gains in child functioning, the QRIS will have more credibility with various stakeholders in the system, including parents and agencies in the public and private sectors that subsidize the cost of program participation. If, on the other hand, the validation study shows no relationship or only a weak or limited relationship between program ratings and children's developmental outcomes, then QRIS designers may want to make appropriate revisions to the rating system (e.g., change the quality components in the rating system or the way components are measured or weighted), guided by the evaluation results.

In all likelihood, an Approach D validation study will show at least some areas of weakness in the QRIS design, such as particular domains of child development that are not affected by the quality ratings or particular components in the quality rating system that are not related to any child outcomes. When a validation study is conducted in the pilot stage of the QRIS development, the findings can be used to make refinements before going to scale, and a subsequent validation study can be used to see if further modifications are needed. If the validation study does not occur until after the QRIS goes to scale, the findings may necessitate making adjustments to a fully operational program, which Zellman and Perlman (2008) point out is not ideal given the experience of implementing QRISs in their pioneering states. For this reason, incorporating Approach D in a pilot phase can be particularly valuable.

Potential Barriers to Success and Strategies for Mitigation

The primary factors that may limit the value of a QRIS validation study are methodological. As with the other approaches, valid inferences depend on having well-trained reliable assessors who measure child functioning at each wave of data collection. In addition, the domains of child functioning that are assessed need to be ones that would be expected to be affected by participation in high-quality ECE and the assessment tools must be valid measures of the desired developmental constructs. As noted above, other methodological issues, such as small sample sizes resulting from the study design or attrition or the selectivity associated with children from different family backgrounds participating in low- and high-quality programs, also threaten the validity of statistical inferences based on the data. In some cases, these and other methodological flaws arise because a study does not have sufficient resources to implement a rigorous design with adequate sample sizes and the ability to collect essential family background data.

In most cases, these potential barriers to success can be remedied by a strong study design with access to the required resources. As with the other approaches, having well-trained asses-

sors who are regularly retested for reliability can mitigate unreliable measurement of children's developmental outcomes, a problem that is largely invisible if reliability testing is not conducted during the assessment process. Likewise, the study design can incorporate multiple measures of child functioning across the relevant domains and rely upon well-validated instruments. To guard against small sample sizes, the study design should include statistical power calculations that anticipate initial response rates, the likely rate of attrition from the sample over time, and planned subgroup analyses. Measuring and controlling for extensive child and family background characteristics can minimize the potential for selectivity bias. And sensitivity analyses can be used to examine the robustness of the findings to the use of alternative methods. Subjecting the study design and findings to peer review can also ensure that robust state-of-the-art methods are employed and that inferences from the data are valid given any methodological concerns. Finally, all of these strategies rely upon having adequate resources for each phase of the study design and implementation.

Approach E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System

Approach E shares a macro perspective with Approach D, but the focus is on the performance of specific ECE programs or the ECE system as a whole, rather than on the performance of the QRIS. In fact, Approach E can be adopted even if there is no QRIS, although it might be incorporated into an existing QRIS as a one-time or ongoing mechanism to assess the effect on child outcomes of specific ECE programs or the system as a whole. The objective of Approach E is to use child assessments to estimate the causal effect of a specific ECE system component (e.g., Head Start, a state-funded prekindergarten program, all publicly subsidized ECE programs, or all center-based programs) or the effect of ECE regardless of program type on child developmental outcomes.

As indicated in Table 3.2, Approach E shares a number of measurement and analysis features with Approach D (and Approach C). As in Approach D, the evaluation can be based on a sample of children in the programs of interest across the state. Assessments are conducted by well-trained, reliable assessors on multiple domains of child functioning. Studies to date have used many of the same assessment instruments employed in the Approach D studies and other ECE research. The frequency of the assessments will depend on the study design. The regression discontinuity (RD) design, a quasi-experimental approach for making causal inferences about policy or program effects discussed next, requires annual assessments of a sample of children at the same time each year, typically in the fall when children enter an ECE program or kindergarten. Information on child and family background is incorporated as well, in part to examine subgroup differences. The resulting data are used by the evaluator to estimate the causal relationship between participating in a specific ECE program, a set of programs, or the ECE system as a whole on child functioning.

Current Practice

To our knowledge, no states have employed Approach E as a required component of their QRIS, either for a specific program or the system as a whole. However, several states have used this strategy to evaluate the effect of their state prekindergarten or preschool programs on measures of child functioning that are characterized as capturing dimensions of school readiness.

A well-implemented experimental design with random assignment to the program to be evaluated or to a no-program or other comparison program would be considered the gold standard for evaluating the causal effect of a state preschool program or any other component of the ECE system on school readiness or other measures of child functioning. However, random assignment often cannot be carried out, either because it is impractical or because it is not possible to find a "no ECE program" control group given the high rates of ECE participation among three- and four-year-olds in most states.[10] Thus, researchers have turned to rigorous quasi-experimental approaches that allow causal inferences to be made with appropriate caveats. The most commonly used method in recent studies of state ECE programs is an RD design, a quasi-experimental approach that uses the "accident" of birth as a random event that determines which children will enter an ECE program (because they must meet an age cutoff) and which will not. The birth date cutoff creates a break or "discontinuity" in the continuous age spectrum between the cohort of children who participate in the program in a given year versus those who must wait another year to enroll.[11] The method then estimates program effects conditional on which children are enrolled in the program. In other words, the method cannot be used to estimate the effect for children who never enroll, as that group may be different in observable and unobservable ways from the children who do participate. Despite this limitation, the RD approach has gained currency because it is relatively straightforward to implement. In principle, other quasi-experimental methods could be used as well, such as matched treatment-control group designs (using propensity score matching or other methods).

Table 3.4 summarizes the results from nine states, including California, where the RD approach has been employed to date to evaluate voluntary state programs. In all cases, the targeted or universal state preschool program was the subject of the evaluation, and the evaluation measured the effect of participation for one year before kindergarten entry (see Cannon and Karoly, 2007, and Lipsey et al., 2011, for additional detail on the preschool program features). Two evaluations have been conducted for Oklahoma, one specific to the Tulsa school district and the other using a statewide sample. The New Mexico evaluation has been repeated annually for three years and results reported for each separate year and pooled over three years. The evaluation in Tennessee is a first report of findings from one region of the state that will cover additional regions in the course of the five-year study.[12] The other state program evaluations have been one-time efforts thus far. The sample sizes across the studies are in the range of 600 to 2,500 children for a single-year evaluation.

---

[10] The national Head Start evaluation is one example of the use of a large-scale random assignment experiment (U.S. Department of Health and Human Services, 2005). However, in that study, 18 percent of children randomized to the control group actually participated in some other Head Start program (Ludwig and Phillips, 2007).

[11] In the RD studies discussed in this chapter, the child assessments are conducted at the same time for both the treatment group (children who attended preschool in the prior year because they made the age cutoff and are now entering kindergarten) and the control group (children who are just entering preschool because they did not meet the age cutoff the prior year). The models estimate the relationship between child assessments and age (controlling for other factors), separately for the treatment and control groups, typically using a nonlinear relationship. The treatment effect is measured as the difference between the expected assessment score of a treated child who just made the age cutoff and the expected assessment score of a control child who just missed the age cutoff. See Gormley and Gayer (2005) and Gormley et al. (2005) for more detail on this approach.

[12] The Tennessee evaluation also includes a randomized control trial component for 23 participating schools in 14 districts across the state that had more preschool applicants than could be accommodated by the number of available spaces (see Lipsey et al., 2011). Results showed significantly better achievement in children who had participated than in those who had not; kindergarten teachers also rated them as significantly more school-ready.

Table 3.4
Estimated Effects of State Preschool Programs on School Readiness Using Quasi-Experimental Designs

		Effect Size			
			WJ Subtest		
Program	Sample Size	Vocabulary (PPVT)	Letter-Word Identification	Spelling	Applied Problems
Arkansas	1,408	0.36*	—	—	0.24*
California[a]	2,304	0.30*–0.47*	—	—	0.31*–0.38*
Michigan	871	0.03	—	—	0.51*
New Jersey	2,072	0.34*	—	—	0.19*
New Mexico (year 1)	886	0.36*	—	—	0.39*
New Mexico (year 2)	924	0.25*	—	—	0.50*
New Mexico (year 3)	1,333	0.17*	—	—	0.43*
New Mexico (pooled)	3,153	0.25*	—	—	0.37*
Oklahoma (Tulsa)	2,484	—	0.79*	0.64*	0.38*
Oklahoma	838	0.32*	—	—	0.49*
South Carolina	777	0.05	—	—	—
Tennessee (Central West)	608	—	0.82*	0.99*	0.48*
West Virginia	720	0.18	—	—	0.52*

SOURCES: Cannon and Karoly (2007), Tables 4.2, 4.3, 4.4, and 4.5; Hustedt, Barnett, and Jung (2007), Figure 4; Hustedt et al. (2008), Figure 1; Hustedt et al. (2009), Figure 1; Lipsey et al. (2011), Table 7; and Barnett, Howes, and Jung (forthcoming), Table 12.

NOTES: The effect sizes are for the treatment-on-treated program effects. Estimates for Oklahoma, Michigan, New Jersey, South Carolina, and West Virginia are based on the pooled sample regression discontinuity model. For Tennessee, data are from children whose ages range within 12 months around the cutoff date. Results are also available for small groups within six months and within three months of the cutoff date; nearly all are also significant.

[a] The range of estimates is based on alternative model specifications using the regression discontinuity design methodology. See Barnett, Howes, and Jung (forthcoming) for details.

* Denotes statistically significant effects at the 5 percent level or better.

— Indicates not available.

Table 3.4 reports results for the PPVT and WJ, the two assessment tools used consistently across these studies. These assessments center on cognitive skills; the studies to date have not aimed to measure effects on noncognitive skills, such as social, behavioral, or emotional functioning. This is not a limitation of the method; these studies could include such domains. With the exception of South Carolina, each evaluation found at least one statistically significant effect of program participation on a cognitive measure, indicating that children who participate in the program have a higher level of functioning in that domain than children who did not participate. The effect sizes are typically in the 0.3 to 0.4 range, although some fall above or below that range. Effects sizes of that magnitude are considered large in the context of education interventions (Cannon and Karoly, 2007).

As noted above, the results in Table 3.4 report the effects of preschool participation for one year before kindergarten entry. In principle, the RD design could be used to assess the effect of participation at any age provided there is a strict age cutoff that determines when children are eligible to enter the program. In practice, this is more likely to be the case for education-based preschool programs that begin at ages three or four using a fixed age cutoff (e.g., turning three or four by September 1), rather than for child care centers that accept children at any age starting as young as infancy. In such cases, evaluators would need to rely on other quasi-experimental methods or an experimental design. The RD method also does not allow longer-term effects of program participation to be assessed, as can be done with experimental designs or with quasi-experimental designs that have a "no program" control group.

Resources Required

The resources required for Approach E are similar to those for Approach D in terms of the general resource categories: study design, including identifying assessment tools and training in their use; data collection; and analysis. The primary drivers of any cost differentials between the two approaches would be overall sample size and the required number of assessment waves. Examples of prior studies using Approach D have tended to have smaller samples than with the RD method under Approach E. However, as noted above, the Colorado and Missouri evaluations would have benefited from larger sample sizes. Whereas Approach D would typically require at least two waves of assessments, the use of the RD design under Approach E could be accomplished with just one wave of data collection. Other quasi-experimental methods or an experimental design under Approach E may also require two or more waves of data collection.

Expected Benefits

As states have increased their investments in ECE programs, there has been an interest in measuring their effects on the ultimate outcome of interest: child functioning, often conceptualized in these studies as school readiness. The examples provided under Approach E relied on RD designs and aimed to document whether publicly supported one-year preschool programs, whether universal or targeted, were having their intended effect on child development. Although the studies have focused primarily on cognitive outcomes to date, a broader range of developmental domains could be examined in future studies. The confirmation of positive effects of ECE program participation on child development for most of the states listed in Table 3.4 and the substantial magnitude of the measured effects relative to other education interventions have boosted support for such programs. In cases where no favorable effects are found or effects are limited to a subset of developmental domains, the results can be used to target resources into areas of professional development and program improvement that would support stronger effects on child development in the future.

Potential Barriers to Success and Strategies for Mitigation

As noted in Table 3.3, the factors that may compromise the validity of Approach E are comparable to those discussed already for Approach D. Although some of the detailed concerns differ between the two approaches, the general issues are the same: the reliability and relevance of the child assessments, other methodological issues that may bias statistical inferences, having insufficient resources to ensure adequate sample sizes, and so on. The associated strategies for mitigation are likewise similar: ensuring that assessors are well trained, using well-validated assessments covering multiple domains, applying rigorous methods to account for possible

biases, subjecting research designs and findings to peer review, and ensuring that there are adequate resources for a well-designed, rigorous evaluation.

CHAPTER FOUR

Conclusions and Policy Guidance

Our goal in this paper has been to identify strategies for incorporating assessments of child functioning into state QRISs or other QI efforts. We are motivated by the reality that QRISs and their rating systems typically focus on the input side of the equation, by measuring the components that are assumed to define quality in ECE settings, whereas the ultimate outcome of interest—whether ECE programs promote children's cognitive, social, emotional, and physical development—is rarely addressed.

Our approach has been to define five strategies that vary in how they incorporate child assessments into state QI efforts and in several cases into a QRIS. The five strategies approach child assessments with different objectives. Two use child assessments to inform and shape classroom practices and to support program improvements. The remaining three approaches use child assessments to measure the effects of participating in a given classroom, program, or ECE system on child functioning. As noted at the outset of Chapter Three, each approach may be implemented on its own or in combination with one or more other approaches.

In this concluding chapter, we offer guidance concerning which approaches to employ and in what circumstances, using our analysis of the experiences to date with each method, the payoff relative to the costs, and the ability to mitigate potential impediments to success. Our guidance for each approach is summarized in Table 4.1 and further discussed below.

Suggestion: Implement Either Approach A or Approach B, Depending on Whether a QRIS Exists

Approaches A and B are the same except that the former is not predicated on the existence of a QRIS, whereas the latter is explicitly part of a QRIS. Our suggestion is that all teachers and programs collect the child assessment data prescribed by these approaches and that programs or states implement one or the other approach, depending upon the existence of a QRIS. Our suggestion stems from recognition that it is good practice for caregivers and teachers to use child assessments to shape their interactions with individual children in the classroom and to identify areas for program improvement. This practice is endorsed by its inclusion in NAEYC program accreditation standards and the accreditation standards for postsecondary ECE teacher preparation programs. The use of child assessments is this manner has the potential to promote more effective individualized care and instruction on the part of caregivers and teachers and to provide program administrators with important information to guide professional development efforts and other quality improvement initiatives. Since the practice appears to be widespread, at least according to ECE teachers in California in center-based classrooms

Table 4.1
Guidance for Incorporating Child Assessments into State QI Efforts

Approach	Guidance	Rationale
A: Caregiver/Teacher- or Program-Driven Assessments to Improve Practice	Implement either Approach A or Approach B depending on whether a state-level QRIS has been implemented: If no QRIS exists, adopt Approach A; consider reinforcing through licensing, regulation, or accreditation if not already part of these mechanisms If a QRIS exists, adopt Approach B	Consistent with good ECE practice Important potential benefits in terms of practice and program improvement for relatively low incremental cost
B: QRIS-Required Caregiver/ Teacher Assessments to Improve Practice		Greater likelihood of use and appropriate use of assessments than with Approach A Important potential benefits in terms of practice and program improvement for relatively low incremental cost
C: Independent Measurement of Child Outcomes to Assess Programs	If considering adopting this approach as part of a QRIS, proceed with caution	Methodology is complex and not sufficiently developed for high-stakes use Costly to implement for uncertain gain Feasibility and value for cost could be tested on a pilot basis
D: Independent Measurement of Child Outcomes to Assess QRIS Validity	Implement this approach when piloting a QRIS and periodically once the QRIS is implemented at scale (especially following major QRIS revisions)	Important to assess validity of the QRIS at the pilot stage and to reevaluate validity as the system matures Methodology Is complex but periodic implementation means high return on investment
E: Independent Measurement of Child Outcomes to Evaluate Specific ECE Programs or the Broader ECE System	Implement this approach periodically (e.g., on a routine schedule or following major policy changes) regardless of whether a QRIS exists	Evidence of system effects can justify spending and guide quality improvement efforts Methodology is complex, but periodic implementation means high return on investment

SOURCE: Authors' analysis.

serving preschool-age children, calling attention to the importance of the practice and providing needed supports (e.g., through teacher preparation programs or professional development opportunities) using either Approach A or Approach B may serve to expand its use (especially in home-based programs and for infants and toddlers) and contribute to the improved use of these assessments through more focus on them in teacher preparation coursework and professional development offerings. The potential for widespread benefit can be weighed against what we expect would be a relatively small incremental cost given the already widespread use of assessments, although costs would be higher if current practice does not include the needed professional development supports to ensure that assessments are used effectively to improve teaching and learning.

Suggestion: Undertake Approach D When Piloting a QRIS and Periodically Once the QRIS Is Implemented at Scale

Approach D is an important component for validating the rating portion of a QRIS. Since this strategy may identify weaknesses in the ability of a QRIS to measure meaningful differ-

ences in ECE program quality that matter for child outcomes, we suggest that this approach be employed in the pilot phase of a QRIS when implementation is at a smaller scale, assuming that there is such a phase.[1] Incorporating a QRIS validation component into a pilot phase will ensure that needed refinements to the QRIS can be introduced before taking the system to scale. This will reduce the need to make changes in the QRIS rating structure once it is fully implemented. In addition to its implementation during the pilot phase, we suggest that a QRIS validation study be repeated periodically (e.g., every five to ten years) or following major changes to a QRIS. This will ensure the continuing relevance of the QRIS rating system given changes in the population of children served by ECE programs, the nature of ECE programs themselves, and other developments in the ECE field. Although the required methodology to implement Approach D is complex and subject to various threats to validity, there are strategies to minimize those concerns such as ensuring sufficient funding for the required sample sizes and the collection of relevant child and family background characteristics. The ability to base the validation design on a sample of programs and children means that it can be a cost-effective investment in the quality of the QRIS.

Suggestion: Implement Approach E Periodically Regardless of Whether a QRIS Exists

Approach E is also a valuable tool, complementary to Approach D, for evaluating the effect of a given ECE program or the entire ECE system, regardless of the existence of a QRIS. Assessing the effect of participation in a given ECE program or group of programs, particularly those that are supported with public dollars, fulfills a need for accountability borne by publicly funded programs. Favorable findings can be used to justify current spending or even to expand a successful program. Unfavorable results can be used to motivate policy changes such as modifications to an ineffective program. Ideally, such a study would be repeated periodically, either to monitor the effect of a major policy change on an ECE program's effect or to ensure that a program that performed well in the past continues to be effective. Although the RD approach used to evaluate preschool programs in several states, including California, has some limitations (e.g., the requirement of a strict age-of-entry requirement and the inability to measure effects for more than one year), when the method can be used, it is relatively straightforward to implement and the findings readily understood. Because it can be implemented using a sample in the range of 1,000 to 2,000 children, it, too, is a very cost-effective approach for determining if an ECE program is achieving its objectives of promoting strong child growth across a range of developmental domains.

Suggestion: If Approach C Is Under Consideration for Inclusion in a QRIS, Proceed with Caution

Although our guidance endorses the other four approaches discussed in this paper, we are considerably less sanguine about Approach C. Although Approach C's aim of measuring the

[1] In some states, the QRIS pilot phase focuses only on refining the QRIS design, rather than on testing the implementation of the QRIS.

causal effect of participating in a specific ECE classroom or program has merit, in reality, the available methods—short of an experimental design—are not sufficiently well developed to justify the cost of large-scale implementation or implementation in high-stakes contexts. As noted in Chapter Three, the K–12 sector has experienced a number of challenges in using methods such as VAM to make inferences about the contribution of a specific teacher, classroom, or school to a child's developmental trajectory. These challenges would be compounded in attempting to use the method in the ECE context. If a state is considering incorporating Approach C into its QRIS, we suggest starting with a pilot phase to assess feasibility, cost, and return on investment. A pilot phase could determine the relative advantage of the approach to alternative methods. Given experiences with VAM in the K–12 context, there will be a number of challenges to overcome before Approach C is likely to be a cost-effective tool for incorporating child outcomes into a QRIS.

In sum, although QRISs have gained currency as input-focused accountability systems, the focus on inputs does not preclude efforts to get to the outcome of interest: child cognitive, social, emotional, and physical functioning. This paper has demonstrated that there are valuable and feasible approaches for incorporating assessments of child functioning into QRISs or QI efforts for ECE programs more generally. Some approaches take a micro perspective, and others have a macro focus. Some are predicated on having a QRIS in place, and others can be implemented without one. Our guidance illustrates that multiple approaches can be used, given their varied and complementary purposes. At the same time, some of these approaches raise methodological concerns that must be dealt with and that may override the potential benefits. Ultimately, policymakers at the state level need to determine the mix of strategies that will be most beneficial given the context of the ECE system in their state, the resources available, and the anticipated returns.

Bibliography

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, *Standards for Educational and Psychological Testing,* Washington, D.C.: American Educational Research Association, 1999.

Barnett, W. Steven, and Debra J. Ackerman, "Costs, Benefits, and Long-Term Effects of Early Care and Education Programs: Recommendations and Cautions for Community Developers," *Journal of the Community Development Society,* Vol. 37, No. 2, Summer 2006. As of January 18, 2012: http://government.cce.cornell.edu/doc/pdf/86-100%20barnett%20ackerman.pdf

Barnett, W. Steven, Dale J. Epstein, Megan E. Carolan, Jen Fitzgerald, Debra J. Ackerman, and Allison H. Friedman, *The State of Preschool 2010,* New Brunswick, N.J.: National Institute for Early Education Research, 2010.

Barnett, W. Steven, Carollee Howes, and Kwanghee Jung, "California's State Preschool Program: Quality and Effects on Children's Cognitive Abilities at Kindergarten Entry," *Final Report to the First 5 California Children and Families Commission,* forthcoming.

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

Blau, David M., "The Quality of Child Care: An Economic Perspective," in David M. Blau, ed., *The Economics of Child Care,* New York: Russell Sage Foundation, 1991, pp. 145–173.

Bowman, Barbara T., M. Suzanne Donovan, and M. Susan Burns, eds., *Eager to Learn: Educating Our Preschoolers,* Washington, D.C.: National Academy Press, 2001.

Bredekamp, Sue, and Teresa Rosegrant, eds., *Reaching Potentials: Transforming Early Childhood Curriculum and Assessment,* Vol. 1, Washington, D.C.: National Association for the Education of Young Children, 1992.

———, *Reaching Potentials: Transforming Early Childhood Curriculum and Assessment,* Vol. 2, Washington, D.C.: National Association for the Education of Young Children, 1995.

Breitner, Leslie K., Richard Brandon, and Nevina Lalic, Budgeting as a Tool for Policy Development, training materials prepared for the UNICEF Social Protection and Inclusion Project in Bosnia and Herzegovina, 2010. As of January 18, 2012: http://evans.washington.edu/files/Module1_Background_Materials.pdf

Buddin, Richard, "Measuring Teacher Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools," Santa Monica, Calif.: RAND Corporation, unpublished working paper, March 2011.

Burchinal, Margaret R., Kirsten Kainz, and Yaping Cai, "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 Martha Zaslow, Ivelisse Martinez-Beck, Kathryn Tout, and Tamara Halle, eds., *Quality Measurement in Early Childhood Settings,* Baltimore, Md.: Brookes Publishing, 2011.

Burchinal, Margaret, Nathan Vandergrift, Robert Pianta, and Andrew Mashburn, "Threshold Analysis of Association between Child Care Quality and Child Outcomes for Low-Income Children in Pre-Kindergarten Programs," *Early Childhood Research Quarterly,* Vol. 25, 2009, pp. 166–176.

Burchinal, Margaret R., Joanne E. Roberts, Laura A. Nabors, and Donna M. Bryant, "Quality of Center Child Care and Infant Cognitive and Language Development," *Child Development,* Vol. 67, 1996, pp. 606–620.

CAELQIS—*See* California Early Learning Quality Improvement System.

California Department of Education (CDE), *California Preschool Learning Foundations,* Vol. 1, Sacramento, Calif, 2008a. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/documents/preschoollf.pdf

———, "Changes to the Developmental Profile of the Desired Results Regulations," Management Bulletin 08-12, Sacramento, Calif, 2008b. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/ci/mb0812.asp

———, *Child Care and Development Fund Plan for California: FFY 2010–2011,* Sacramento, Calif., October 2009a. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/documents/stateplan1011final.doc

———, *California Infant/Toddler Learning and Development Foundations,* Sacramento, Calif., 2009b. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/itf09aavcontents.asp

———, *Introduction to Desired Results,* Sacramento, Calif., 2010a. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/ci/desiredresults.asp

———, *California Preschool Curriculum Framework,* Vol. 1, Sacramento, Calif., 2010b. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/documents/psframeworkkvol1.pdf

California Department of Education (CDE) and Center for Child and Family Studies at WestEd, *Desired Results for Children and Families,* website, Camarillo, Calif., undated. As of January 18, 2012:
http://www.wested.org/desiredresults/training/index.htm

California Early Learning Quality Improvement System (CAELQIS) Advisory Committee, *Dream Big for Our Youngest Children: Executive Summary,* Sacramento, Calif.: 2010a. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/documents/fnlrptexecsummary.pdf

———, *Dream Big for Our Youngest Children: Final Report,* Sacramento, Calif.: 2010b. As of January 18, 2012:
http://www.cde.ca.gov/sp/cd/re/documents/fnlrpt2010.pdf

Campbell, Francis A., and Craig T. Ramey, "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention," *American Educational Research Journal,* Vol. 32, No. 4, 1995, pp. 743–772.

Cannon, Jill S., and Lynn A. Karoly, *Who Is Ahead and Who Is Behind? Gaps in School Readiness and Student Achievement in the Early Grades for California's Children,* Santa Monica, Calif.: RAND Corporation, TR-537-PF/WKKF/PEW/NIEER/WCJVSF/LAUP, 2007. As of January 18, 2012:
http://www.rand.org/pubs/technical_reports/TR537.html

CDE—*See* California Department of Education.

Center 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,* Cambridge, Mass., Harvard University, 2007. As of January 18, 2012:
http://developingchild.harvard.edu/resources/reports_and_working_papers/policy_framework/

Chen, Huey-Tsyh, *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness,* Thousand Oaks, Calif.: Sage Publications, 2005.

Cizek, Gregory J., *Introduction to Validity,* Presentation to the National Assessment Governing Board of NAEP, August 2007.

Clarke-Stewart, K. Alison, Deborah Lowe Vandell, Margaret Burchinal, Marion O'Brien, and Kathleen McCartney, "Do Regulable Features of Child-Care Homes Affect Children's Development?" *Early Childhood Research Quarterly,* Vol. 17, No. 1, 2002, pp. 52–86.

Council of Chief State School Officers, *Building an Assessment System to Support Successful Early Learners: The Role of Child Assessment in Program Evaluation and Improvement,* Washington, D.C., 2003.

Cryer, Debby, and Margaret Burchinal, "Parents as Child Care Consumers," *Early Childhood Research Quarterly,* Vol. 12, 1997, pp. 35–38.

Cryer, Debby, Wolfgang Tietze, and Holger Wessels, "Parents' Perceptions of Their Children's Child Care: A Cross-National Comparison," *Early Childhood Research Quarterly,* Vol. 17, 2002, pp. 259–277.

Daily, Sarah, Mary Burkhauser, and Tamara Halle, "A Review of School Readiness Practices in the States: Early Learning Guidelines and Assessments," *Early Childhood Highlights,* Vol. 1, Issue 3, Washington, D.C.: Child Trends, June 17, 2010. As of January 18, 2012: http://www.childtrends.org/Files/Child_Trends-2010_06_18_ECH_SchoolReadiness.pdf

Dearing, Eric, Kathleen McCartney, and Beck A. Taylor, "Does Higher Quality Early Child Care Promote Low-Income Children's Math and Reading Achievement in Middle Childhood?" *Child Development,* Vol. 80, No. 5, 2009, pp. 1329–1349.

Duncan, Greg J., "Modeling the Impacts of Child Care Quality on Children's Preschool Cognitive Development," *Child Development,* Vol. 74, No. 5, 2003, pp. 1454–1475.

Elicker, James and Kathy Thornburg, *Evaluation of Quality Rating and Improvement Systems for Early Childhood Programs and School-Age Care: Measuring Children's Development,* Research-to-Practice Brief, OPRE 2011-11a, Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2011.

Elicker, James, Carolyn Clawson Langill, Karen Ruprecht, and Kyong-Ah Kwon, *Paths to Quality: A Child Care Quality Rating System for Indiana. What Is Its Scientific Basis?* West Lafayette, Ind.: Purdue University, Center for Families and Department of Child Development and Family Studies, 2007.

Elicker, James, Carolyn Clawson Langill, Karen Ruprecht, Joellen Lewsader, and Treshawn Anderson, *The Relationship Among Paths to QUALITY Level, Child Care Quality and Child Outcomes: Findings from Indiana's Voluntary QRIS,* West Lafayette, Ind.: Purdue University, 2011, As of January 18, 2012: http://www.cfs.purdue.edu/cdfs/documents/SRCD-2011/Elicker-Langill-Ruprecht-Lewsader-Anderson_2011.ppt

Epstein, Ann S., Lawrence J. Schweinhart, Andrea DeBruin-Parecki, and Kenneth B. Robin, *Preschool Assessment: A Guide to Developing a Balanced Approach,* New Brunswick, N.J.: National Institute for Early Education Research, July 2004. As of January 18, 2012: http://nieer.org/resources/policybriefs/7.pdf

Fuller, Bruce, and Sharon Lynn 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, Calif.: University of California, Graduate School of Education—Policy Analysis for California Education, February 2000.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst, *Evaluating Teachers: The Important Role of Value-Added,* Washington, D.C.: Brookings Institution, November 17, 2010. As of January 18, 2012: http://www.brookings.edu/~/media/Files/rc/reports/2010/1117_evaluating_teachers/1117_evaluating_teachers.pdf

Gormley, William T., and Ted Gayer, "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program," *Journal of Human Resources,* Vol. 40, No. 3, Summer 2005, pp. 533–558.

Gormley, William T., Ted Gayer, Deborah Phillips, and Brittany Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology,* Vol. 41, No. 6, 2005, pp. 872–884.

Guddemi, Marcy, and Betsy J. Case, *Assessing Young Children,* Santa Monica, Calif.: Pearson Education, Inc., 2004.

Halle, Tamara, Martha Zaslow, Julia Wessel, Shannon Moodie, and Kristen Darling-Churchill, *Understanding and Choosing Assessments and Developmental Screeners for Young Children Ages 3–5: Profiles of Selected Measures,* OPRE 2011-23, Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, June 2011. As of January 18, 2012: http://www.acf.hhs.gov/programs/opre/hs/dev_screeners/reports/screeners_final.pdf

Hannaway, Jane, and Laura Hamilton, *Performance-Based Accountability Policies: Implications for School and Classroom Practices,* Washington, D.C.: The Urban Institute and RAND Corporation, 2008.

Hebbeler, Kathy, and Cornelia Taylor, *Using Outcomes Data for Program Improvement,* Early Childhood Outcome Center, SRI International, Webinar, April 12, 2011. As of January 18, 2012: http://www.fpg.unc.edu/~eco/pages/archive.cfm

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

Helburn, Suzanne W., John R. Morris, and Kathy Modigliani, "Family Child Care Finances and Their Effect on Quality and Incentives," *Early Childhood Research Quarterly,* Vol. 17, 2002, pp. 512–538.

Heubert, Jay P., and Robert M. Hauser, eds., *High Stakes: Testing for Tracking, Promotion and Graduation,* Washington, D.C.: National Academies Press, 1999.

Howes, Carollee, "Relations Between Early Child Care and Schooling," *Developmental Psychology,* Vol. 24, 1988, pp. 53–57.

Hustedt, Jason T., W. Steven Barnett, and Kwanghee Jung, *The Effects of the New Mexico PreK Initiative on Young Children's School Readiness,* New Brunswick, N.J.: National Institute for Early Education Research, 2007.

Hustedt, Jason T., W. Steven Barnett, Kwanghee Jung, and Alexandra Figueras, *Impacts of New Mexico PreK on Children's School Readiness at Kindergarten Entry: Results from the Second Year of a Growing Initiative,* New Brunswick, N.J.: National Institute for Early Education Research, June 2008.

Hustedt, Jason T., W. Steven Barnett, Kwanghee Jung, and Alexandra Figueras-Daniel, *Continued Impacts of New Mexico PreK on Children's Readiness for Kindergarten: Results from the Third Year of Implementation,* New Brunswick, N.J.: National Institute for Early Education Research, September 2009.

Iruka, IIheoma U., and Priscilla R. Carver, *Initial Results from the 2005 NHES Early Childhood Program Participation Survey,* NCES 2006-075, Washington, D.C.: National Center for Education Statistics, U.S. Department of Education, 2006.

Kagan, Sharon Lynn, "On Buckets, Banks, and Hearts: Aligning Early Childhood Standards and Systems," presentation at the Build Conference on Quality Rating Improvement Systems, Minneapolis, Minn., June 4, 2008. As of January 18, 2012: http://www.buildinitiative.org/files/MSP-Build%205.19.08.ppt

Karoly, Lynn A., *Preschool Adequacy and Efficiency in California: Issues, Policy Options, and Recommendations,* Santa Monica, Calif.: RAND Corporation, MG-889-PF/WKKF/PEW/NIEER/WCJVSF/LAUP, 2009. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG889.html

Karoly, Lynn A., Bonnie Ghosh-Dastidar, Gail L. Zellman, Michal Perlman, and Lynda Fernyhough, *Prepared to Learn: The Nature and Quality of Early Care and Education for Preschool-Age Children in California,* Santa Monica, Calif: RAND Corporation, TR-539-PF/WKKF/PEW/NIEER/WCJVSF/LAUP, 2008. As of January 18, 2012: http://www.rand.org/pubs/technical_reports/TR539.html

Karoly, Lynn A., M. Rebecca Kilburn, and Jill S. Cannon, *Early Childhood Interventions: Proven Results, Future Promise,* Santa Monica, Calif.: RAND Corporation, MG-341-PNC, 2005. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG341.html

Kauerz, Kristie, and Abby Thorman, *QRIS and P-3: Creating Synergy Across Systems to Close Achievement Gaps and Improve Opportunities for Young Children,* Boston, Mass.: BUILD Initiative, March 2011. As of January 18, 2012: http://www.buildinitiative.org/files/QRIS_P-3brief.pdf

Kellogg Foundation, *Logic Model Development Guide: Using Logic Models to Bring Together Planning, Evaluation, and Action,* Battle Creek, Mich., January 2004. As of January 18, 2012: http://www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx

Lamb, Michael E., "Nonparental Child Care: Context, Quality, Correlates, and Consequences," in W. Damon, I. E. Sigel, and K. A. Renninger, eds., *Handbook of Child Psychology,* Vol. 4: Child Psychology in Practice, New York: Wiley, 1998, pp. 73–133.

Langill, Carolyn, James Elicker, Karen Ruprecht, Kyong-Ah Kwon, and Joellen Guenin, *Paths to Quality—A Child Care Quality Rating and Improvement System for Indiana: Evaluation Methods and Measures,* Technical Report No. 2. West Lafayette, Ind.: Purdue University, Center for Families, 2009.

Le, Vi-Nhuan, Michal Perlman, Gail L. Zellman, and Laura S. Hamilton, "Measuring Child-Staff Ratios in Child Care Centers: Balancing Effort and Representativeness," *Early Childhood Research Quarterly,* Vol. 21, 2006, pp. 267–279.

Leslie, Leigh A., Richard Ettenson, and Patricio Cumsille, "Selecting a Child Care Center: What Really Matters to Parents?" *Child and Youth Care Forum,* Vol. 29, 2000, pp. 299–322.

Lipsey, Mark W., Dale C. Farran, Carol Bilbrey, Kerry G. Hofer, and Nianbo Dong, *Initial Results of the Evaluation of the Tennessee Voluntary Pre-K Program,* Nashville, Tenn.: Vanderbilt University, Peabody Research Institute, April, 2011. As of January 18, 2012: http://peabody.vanderbilt.edu/Documents/pdf/PRI/Initial%20Results%20of%20the%20Evaluation%20of%20TN_VPK.pdf

Ludwig, Jens, and Deborah A. Phillips, "The Benefits and Costs of Head Start," *Social Policy Report,* Vol. 21, No. 3, 2007. As of January 18, 2012: http://www.srcd.org/index/php?option=com_docman&task=doc_download&gid=83 [restricted access]

Lynch, Eleanor W., and Marci J. Hanson, "Ensuring Cultural Competence in Assessment," in M. McLean, D. B. Bailey, and M. Wolery, eds., *Assessing Infants and Preschoolers with Special Needs,* second edition, Columbus, Ohio: Merrill, 1996, pp. 69–95.

Mashburn, Andrew J., Robert C. Pianta, Bridget K. Hamre, Jason T. Downer, Oscar Barbarin, Donna Bryant, Margaret Burchinal, Diana M. Early, and Carollee Howes, "Measures of Classroom Quality in Prekindergarten an Children's Development of Academic, Language, and Social Skills," *Child Development,* Vol. 79, No. 3, May/June 2008, pp. 732–749.

McCaffrey, Daniel F., Daniel M. Koretz, J. R. Lockwood, and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability,* Santa Monica, Calif.: RAND Corporation, MG-158-EDU, 2004. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG158.html

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, Thomas A. Louis, and Laura S. Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics,* Vol. 29, No. 1, Spring 2004, pp. 67–101.

McCauley, Louise, "The Developmental Assessment of Young Children," *Child and Adolescent Psychiatry On-Line,* undated. As of January 18, 2012: http://priory.com/psych/assessyoung.htm

Mitchell, Anne W., *Stair Steps to Quality: A Guide for States and Communities Developing Quality Rating Systems for Early Care and Education,* United Way, Success by 6, Fairfax, Va.: National Child Care Information Center, July 2005.

NACCRRA—*See* National Association of Child Care Resource and Referral Agencies.

NAEYC—*See* National Association for the Education of Young Children.

National Association for the Education of Young Children (NAEYC), *Overview of the NAEYC Early Childhood Program Standards,* Washington, D.C., 2008. As of January 18, 2012: http://www.naeyc.org/files/academy/file/OverviewStandards.pdf

———, *2010 NAEYC Standards for Initial and Advanced Early Childhood Professional Preparation Programs,* Washington, D.C., January 2010. As of January 18, 2012: http://www.naeyc.org/files/ecada/file/2010%20NAEYC%20Initial%20&%20Advanced%20Standards.pdf

———, *Summary of Accredited Programs,* Washington, D.C., 2011. As of January 18, 2012: http://oldweb.naeyc.org/academy/summary/center_summary.asp

National Association of Child Care Resource and Referral Agencies, *We Can Do Better, 2011 Update: NACCRRA's Ranking of State Child Care Center Regulations and Oversight,* Washington, D.C., 2011. As of January 18, 2012:
http://www.naccrra.org/publications/naccrra-publications/we-can-do-better-2011.php

National Association of School Psychologists, *School Psychologists' Involvement in Assessment: Position Statement,* Bethesda, Md., 2009a.

———, *Early Childhood Assessment: Position Statement,* Bethesda, Md., 2009b.

National Early Childhood Accountability Task Force, *Taking Stock: Assessing and Improving Early Childhood Learning and Program Quality,* Washington, D.C.: Pew Charitable Trusts, October 2007.

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,* Vol. 74, No. 4, 2000.

———, "Does Quality of Child Care Affect Child Outcomes at Age 4 ½?" *Developmental Psychology,* Vol. 39, No. 3, 2003, pp. 451–469.

Nelson, Geoffrey, Ann Westhues, and Jennifer MacLeod, "A Meta-Analysis of Longitudinal Research on Preschool Prevention Programs for Children [electronic version]," *Prevention and Treatment,* Vol. 6, No. 1, December 2003.

NICHD—*See* National Institute of Child Health and Human Development.

Peisner-Feinberg, Ellen S., and Margaret R. Burchinal, "Relations Between Preschool Children's Child-Care Experiences and Concurrent Development: The Cost, Quality, and Outcomes Study," *Merrill-Palmer Quarterly,* Vol. 43, No. 3, 1997, pp. 451–477.

Peisner-Feinberg, Ellen S., Margaret R. Burchinal, Richard M. Clifford, Noreen Yazejian, Mary L. Culkin, J. Zelazo, Carollee Howes, P. Byler, S. L. Kagan, and J. Rustici, *The Children of the Cost, Quality, and Outcomes Study Go to School: Technical Report,* Chapel Hill, N.C.: University of North Carolina, Frank Porter Graham Child Development Center, 1999.

Peisner-Feinberg, Ellen S., Margaret R. Burchinal, Richard M. Clifford, Mary L. Culkin, Carollee Howes, Sharon Lynn Kagan, and Noreen Yazejian, "The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Developmental Trajectories through Second Grade," *Child Development,* Vol. 72, No. 5, 2001, pp. 1534–1553.

Pennsylvania Office of Child Development and Early Learning, *Keystone Stars: Center Performance Standards for FY 2011–2012,* Harrisburg, Pa.: Pennsylvania Departments of Public Welfare and Education, Office of Child Development and Early Learning, 2010. As of January 18, 2012:
http://www.pakeys.org/docs/2011-2012%20Keystone%20STARS%20Performance%20Standards%20for%20Centers.pdf

Pianta, Robert C., W. Steven Barnett, Margaret Burchinal, and Kathy R. Thornburg, "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,* Vol. 10, No. 2, 2009, pp. 49–88.

Ramey, Craig T., Frances A. Campbell, Margaret Burchinal, Martie L. Skinner, David M. Gardner, and Sharon L. Ramey, "Persistent Effects of Early Intervention on High-Risk Children and Their Mothers," *Applied Developmental Science,* Vol. 4, 2000, pp. 2–14.

Ramey, Craig T., and Sharon L. Ramey, "Early Learning and School Readiness: Can Early Intervention Make a Difference?" in N. F. Watt, C. Ayoub, R. H. Bradley, E. Puma, and W. A. LeBouef, eds., *The Crisis in Youth Mental Health: Critical Issues and Effective Programs. Early Intervention Programs and Policies,* Vol. 4. Westport, Conn.: Praeger, 2006, pp. 291–318.

Reardon, Sean F., and Stephen W. Raudenbush, "Assumptions of Value-Added Models for Estimating School Effects," paper presented at the National Conference on Value-Added Modeling, April 22–24, 2008.

Reynolds, Cecil, "The Problem of Bias in Psychological Assessment," in C. R. Reynolds and T. B. Gutkin, eds., *The Handbook of School Psychology,* New York: Wiley, 1982.

Rice, Jennifer King, *Teacher Quality: Understanding the Effectiveness of Teacher Attributes,* Washington, D.C.: Economic Policy Institute, 2003.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman, *Evaluation: A Systematic Approach,* Thousand Oaks, Calif.: Sage Publications, 2004.

Scarr, Sandra, "American Child Care Today," *American Psychologist,* Vol. 53, 1998, pp. 95–108.

Seo, Sojung, "Early Child Care Choices: A Theoretical Model and Research Implications," *Early Child Development and Care,* Vol. 173, 2003, pp. 637–650.

Shepard, Lorrie A., "The Challenges of Assessing Young Children Appropriately," *Phi Delta Kappa,* November 1994.

Shepard, Lorrie A., Sharon Lynn Kagan, and Emily Wurtz, eds., *Principles and Recommendations for Early Childhood Assessments,* Washington, D.C.: National Education Goals Panel, 1998. As of January 18, 2012: http://govinfo.library.unt.edu/negp/reports/prinrec.pdf

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

Smith, Linda K., and Mousumi Sarkar, *Making Quality Child Care Possible: Lessons Learned from NACCRRA's Military Partnerships,* Washington, D.C.: National Association of Child Care Resource and Referral Agencies (NACCRRA), 2008. As of January 18, 2012: http://www.naccrra.org/publications/naccrra-publications/publications/LesnsLrnd%20Rprt-m2.pdf

Snow, Catherine E., and Susan B. Van Hemel, eds., *Early Childhood Assessment: Why, What and How,* Washington, D.C., National Research Council, 2008

Stecher, Brian M., Frank A. Camm, Cheryl L. Damberg, Laura S. Hamilton, Kathleen J. Mullen, Christopher D. Nelson, Paul Sorensen, Martin Wachs, Allison Yoh, Gail L. Zellman, and Kristin J. Leuschner, *Toward a Culture of Consequences: Performance-Based Accountability Systems for Public Services,* Santa Monica, Calif.: RAND Corporation, MG-1019, 2010. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG1019.html

Thornburg, Kathy R., Wayne A. Mayfield, Jacqueline S. Hawks, and Kathryn L. Fuger, *The Missouri Quality Rating System School Readiness Study,* Columbia, Mo.: University of Missouri, Center for Family Policy and Research, 2009. As of January 18, 2012: http://mucenter.missouri.edu/MOQRSreport.pdf

Tout, Kathryn, Rebecca Starr, Tabitha Isner, Jennifer Cleveland, Margaret Soli, and Katie Quinn, *Evaluation of Parent Aware: Minnesota's Quality Rating and Improvement System Pilot Year 3 Evaluation Report,* November, 2010a.

Tout, Kathryn, Rebecca Starr, Margaret Soli, Shannon Moodie, Gretchen Kirby, and Kimberly Boller, *Compendium of Quality Rating Systems and Evaluations,* Washington, D.C.: Administration for Children and Families, U.S. Department of Health and Human Services; Mathematica Policy Research, Inc.; and Child Trends, 2010b. As of January 18, 2012: http://www.acf.hhs.gov/programs/opre/cc/childcare_quality/compendium_qrs/qrs_compendium_final.pdf

U.S. Department of Health and Human Services, *Head Start Program Performance Standards and Other Regulations,* Washington, D.C.: U.S. Department of Health and Human Services, undated. As of January 18, 2012: http://eclkc.ohs.acf.hhs.gov/hslc/Head%20Start%20Program/Program%20Design%20and%20Management/Head%20Start%20Requirements/Head%20Start%20Requirements

———, *The Role of Early Head Start Programs in Addressing the Child Care Needs of Low-Income Families with Infants and Toddlers: Influences on Child Care Use and Quality,* Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, Early Head Start Research and Evaluation Project, 2004. As of January 18, 2012: http://www.mathematica-mpr.com/publications/PDFs/roleofearly.pdf

———, *Head Start Impact Study: First Year Findings,* Washington, D.C.: Administration for Children and Families, June 2005. As of January 18, 2012: http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf

Van Horn, M. Lee, Sharon Landsman Ramey, Beverly A. Mulvihill, and Wanda Y. Newell, "Reasons for Child Care Choice and Appraisal Among Low-Income Mothers," *Child and Youth Care Forum,* Vol. 30, 2001, pp. 231–249.

Vandell, Deborah L., and Barbara 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.

Votruba-Drzal, Elizabeth, Rebekah L. Coley, and P. Lindsay Chase-Lansdale, "Child Care and Low-Income Children's Development: Direct and Moderated Effects," *Child Development,* Vol. 75, 2004, pp. 296–312.

Whitebook, Marcy, Deborah Phillips, Dan Bellm, Nancy Crowell, Mirella Almaraz, and Joon Yong Jo, *Two Years in Early Care and Education: A Community Portrait of Quality and Workforce Stability,* Berkeley, Calif.: Center for the Study of Child Care Employment, April 2004. As of January 18, 2012: http://www.irle.berkeley.edu/cscce/wp-content/uploads/2010/07/twoyears_final.pdf

Wolfe, Barbara, and Scott Scrivner, "Child Care Use and Parental Desire to Switch Care Type among a Low-Income Population," *Journal of Family and Economic Issues,* Vol. 25, 2004, pp. 139–162.

Zaslow, Martha, Rachel Anderson, Zakia Redd, Julia Wessel, Louisa Tarullo, and Margaret Burchinal, *Quality Dosage, Thresholds, and Features in Early Childhood Settings: A Review of the Literature,* OPRE 2011-5, Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2010.

Zellman, Gail L., Richard Brandon, Kimberly Boller, and J. Lee Kreader, *Effective Evaluation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care,* Research-to-Practice Brief, OPRE 2011-11a, Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, June 2011. As of January 18, 2012: http://www.acf.hhs.gov/programs/opre/cc/childcare_technical/reports/quality_rating.pdf

Zellman, Gail L., and Susan M. Gates, *Examining the Cost of Military Child Care,* Santa Monica, Calif.: RAND Corporation, MR-1415-OSD, 2002. As of January 18, 2012: http://www.rand.org/pubs/monograph_reports/MR1415.html

Zellman, Gail L., and Michal Perlman, *Child-Care Quality Rating and Improvement Systems in Five Pioneer States: Implementation Issues and Lessons Learned,* Santa Monica, Calif.: RAND Corporation, MG-795-AECF/SPF/UWA, 2008. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG795.html

Zellman, Gail L., Michal Perlman, Vi-Nhuan Le, and Claude Messan Setodji, *Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child-Care Quality,* Santa Monica, Calif.: RAND Corporation, MG-650-QEL, 2008. As of January 18, 2012: http://www.rand.org/pubs/monographs/MG650.html