Early Childhood Research Quarterly 30 (2015) 241–254

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

# Improving QRISs through the use of existing data: A virtual pilot of the California QRIS

Gail L. Zellman [a,*], Lynn A. Karoly [b,1]

[a] RAND Corporation, 1776 Main Street, Santa Monica, CA 90407, United States
[b] RAND Corporation, 1200 S. Hayes Street, Arlington, VA 22202, United States

ARTICLE INFO

*Article history:*
Available online 30 April 2014

*Keywords:*
Child care quality
Quality rating systems
Quality rating systems design
Quality rating system pilot studies

ABSTRACT

Available research underscores the value of using data to make and modify the many decisions required to design a child care quality rating and improvement system (QRIS). This paper argues for analyzing existing program data to address key questions and decisions in the early design stages of a QRIS, even in advance of pilot activities. We employed two datasets covering California ECE programs to provide cost-effective and timely input to policymakers for the proposed California QRIS, a block design system with five quality elements and five rating tiers. The first data source is the provider sample component of the 2007 RAND California Preschool Study (CPS), which represents all California providers. The second dataset derives from quality measurement of the ECE providers required to participate in San Francisco County's Gateway to Quality (GTQ) initiative. To address the study questions, we replicated as closely as possible the proposed QRIS rating structure for the available quality elements. Our "virtual pilot" analysis had limitations: we could examine only three of the five quality elements. Findings revealed that most programs in our statewide center-based sample would rate better on some quality elements than others. GTQ data revealed that center-based classrooms serving infants and toddlers did not score as well as those serving preschool-age children and home-based programs scored considerably lower on the applicable Environmental Rating Scale (ERS) than center-based programs.



## Introduction

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 multi-component assessments designed to make the quality of ECE 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 (Zellman & Perlman, 2008).

## Design decisions

The design of a QRIS involves a large number of decisions. Some design decisions are broad ones such as the ultimate goals of the QRIS and which providers may and must participate. For example, there is a growing sense that programs that receive public subsidies should be required to participate in QRISs, although no state had included such a mandate as of 2010 (Tout et al., 2010). Important decisions may also concern how to treat family child care (FCC) providers and infant-toddler programs: applying the same standards to programs of all types has considerable appeal as consistent standards will produce comparable ratings across program types. But FCC providers often find it difficult to attain the higher rating tiers, and teachers in infant-toddler programs often cannot meet education requirements at higher rating levels (Fuller & Kagan, 2000; Goerge et al., 2013; Helburn, 1995; Lahti et al., 2011; Scarr, Eisenberg, & Deater-Deckard, 1994; Whitebook et al., 2004).

### *Quality elements and measures*

Many decisions must be made about the definition and measurement of quality; decisions concerning what quality elements

* Corresponding author. Tel.: +1 310 393 0411; fax: +1 310 393 4818.
*E-mail addresses:* zellman@rand.org (G.L. Zellman), karoly@rand.org (L.A. Karoly).
[1] Tel.: +1 703 413 1100; fax: +1 703 413 8111.

http://dx.doi.org/10.1016/j.ecresq.2014.04.006
0885-2006/

to include in program ratings are particularly important as they essentially define quality standards for the system. These decisions also define what providers will attend to; for this reason, there is often pressure to include elements such as family engagement for which there is limited evidence concerning a link to important outcomes such as child learning (Zellman, Perlman, Le, & Setodji, 2008). Decisions also must be made about how to weight the different elements and combine them into a single summary program rating for each participating program. The research literature provides limited guidance concerning the most appropriate ways to do this (Lugo-Gil, Sattar, Ross, Boller, & Kirby, 2011; Tout, Zaslow, Halle, & Forry, 2009).

Once rating elements are selected, designers must decide which tools to use to measure each one. In general, Tout et al. (2010) find that states tend to employ similar tools, e.g., most use the applicable Environmental Rating Scale (ERS) (Harms, Clifford, & Cryer, 2005; Harms, Cryer, & Clifford, 2006; Harms, Cryer, & Clifford, 2007) to assess classroom environment. This reflects both limited options and a lack of empirical validation data (Zellman & Perlman, 2008). Costs drive some decisions about *how* to use these tools. For example, since observational tools are costly to administer, some states have chosen to conduct independent observations only at the upper rating tiers.

### *Rating structures*

The structures that drive how the selected elements are weighted to produce that summary program rating appear to have major implications for how programs focus their quality improvement efforts and how difficult it is for programs to improve their rating (Lahti, 2013; Tout et al., 2010). Rating structures typically rely on one of the three approaches: building block, point, or some combination of the two, typically labeled hybrid designs. In a building block structure, all of the standards at one rating level must be met before a program is eligible to receive a rating at the next higher level. This feature of block systems essentially forces programs to attend to all rating elements; a poor rating on a single quality element can prevent an otherwise well-rated program from receiving a higher summary rating. In a point system, points are earned for each element; these points are added together to determine the overall rating level. If a program achieves a score above an established threshold for a given rating tier, it receives the next higher rating. A feature of point systems is that programs may score relatively low on one or more elements but achieve a higher rating because of high scores on other elements. Some argue that block systems produce more comparable ratings for consumers because a given program rating assures that a program meets quality standards on each rating element. In contrast, point systems, which enable programs to excel on some rating elements but do far less well on others, may blur the meaning of a given rating; a highly rated program might actually score poorly on one or more quality elements. Supporters of point systems believe that the flexibility inherent in point systems allows programs to build on their strengths, which encourages programs to volunteer to participate (Zellman & Perlman, 2008). Hybrid systems may set quality floors for all elements but enable programs to focus on particular ones as they strive to improve their rating.

Designers must also decide how the various quality elements will be combined into summary program ratings. A particular issue here is the question of cut points. In block designs, designers must decide what score on each element qualifies a program for a given rating. In point structures, designers must decide the number of points a program must earn to qualify for a given rating.

## Empirical evidence in support of QRIS design

Few of these important design decisions are supported by robust empirical data. There are several reasons for this. First, QRISs are relatively new; the oldest are less than 15 years old. Given that full implementation takes years, there has been limited time to assess their functioning, although a number of states with newer systems have engaged in systematic piloting efforts that are beginning to address important system questions (see Lahti, 2013, for a discussion of efforts in four states). Second, each state's system is different. This makes it challenging for QRIS designers in one state to use findings from pilot or implementation efforts in other states. Third, a decision to launch a QRIS is often a political one. Especially in the early years when QRISs were not widely accepted, some advocates believed that there was no time to waste on pilot studies in case the political climate changed and support ebbed before the system could be widely implemented (Zellman & Perlman, 2008). Given the pressure to design and implement a QRIS, states typically decide to employ the same measures being used in other states' QRISs.

### *Pre-implementation data*

Nevertheless, some states have collected data to guide QRIS design. For example, Elicker, Langill, Ruprecht, and Kwon (2007) conducted a pre-implementation scientific review of the quality standards contained in Indiana's QRIS to assess the "scientific validity" of those standards. The goal of this work was to ensure that there was "substantial evidence" that the indicators for each of the four QRIS program quality standards: health and safety, environmental supports for learning, planned curriculum, and national accreditation, support children's development, learning, or well-being in child care. Results indicated that 75% of the quality indicators showed moderate to strong associations with child outcomes. Thornburg et al. (2011) used pilot data from Missouri's QRIS as well as state registry data to determine how to measure specific quality elements (e.g., how to establish standards for education and training levels for staff in different positions), how to aggregate quality measured at the classroom level to a center-level measure, and how many classrooms to assess in centers with multiple rooms.

### *System data*

As QRISs have matured, states increasingly have used the data they collect in the course of their program rating process to validate their QRISs and examine the functioning of the elements and scoring rubrics they employ to assess quality (see Zellman and Fiene (2012) for a discussion of four approaches to QRIS validation). One focus of these studies is to examine whether the elements that are part of the state's QRIS are working in expected ways, e.g., how program scores on given elements are distributed, the extent to which scores on different elements are correlated, and whether scores on individual elements and overall ratings distinguish programs known to differ in quality. These studies have found that distributions across the several elements of a QRIS typically vary enormously (Sabol & Pianta, 2012; Tout et al., 2011); that correlations among element scores range from low to moderate, depending on the specific elements included (Sabol & Pianta, 2012); and that some elements show no variability across programs known to vary in quality (Tout et al., 2011; Zellman et al., 2008).

A second objective of this literature has been to examine relationships between QRIS ratings and other measures of quality; the most commonly used criterion measure is the set of ERSs (Infant-Toddler Environment Rating Scale-Revised or ITERS-R, Early Childhood Environment Rating Scale-Revised or ECERS-R and Family Child Care Environment Rating Scale-Revised or FCCERS-R) developed by Harms et al. (2005, 2006, 2007). These studies