Early Childhood Research Quarterly 44 (2018) 161–169

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Measuring the quality of teacher–child interactions at scale: Comparing research-based and state observation approaches

Virginia E. Vitiello*, Daphna Bassok, Bridget K. Hamre, Daniel Player, Amanda P. Williford

University of Virginia, United States

ARTICLE INFO

Article history:
Received 14 April 2017
Received in revised form 27 February 2018
Accepted 5 March 2018
Available online 20 April 2018

Keywords:
Classroom observations
Qris
Preschool

ABSTRACT

Use of observational measures to monitor preschool quality is growing rapidly. Although a large body of research has examined the validity of classroom observation tools within the context of researcher-conducted studies, little research to date has examined the extent to which the observations conducted as a part of state accountability efforts correspond to observations collected by research teams. This paper examines the degree of agreement between local and research rater teams using an observational measure of preschool classroom quality. It also explores the extent to which ratings predicted gains in children's literacy, math, and self-regulation skills. Local ratings were conducted as a part of Louisiana's quality rating and improvement system. Both rating teams observed 85 classrooms offering publicly funded preschool programs using the Pre-K CLASS, and 820 children from these classrooms (average age = 52.6 months, SD = 3.6 months) were directly assessed in the fall and spring. Results indicated correlations between local and research teams' scores on corresponding domains, ranging from $r = .21$ to .43. Both teams' scores were significantly but modestly related to children's learning gains, although patterns of association differed. Results are discussed in the context of policies that require observational measures at scale.

Early childhood education (ECE) programs can yield short and long-term benefits for children (Phillips et al., 2017). However, many children in the United States attend ECE programs that do not offer high quality environments (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Dowsett, Huston, Imes, & Gennerian, 2008). Lower-quality programs are less likely to benefit children in terms of developing school readiness skills than are higher-quality programs (Karoly, 2014; Sabol & Pianta, 2014, 2015). This has led to substantial public investments in improving ECE quality. Early childhood accountability systems have become one increasingly prominent policy lever. Spurred by Federal funding from the Race to the Top Early Learning Challenge, today nearly all states have developed and are expanding Quality Rating and Improvement Systems (QRIS; The Build Initiative and Child Trends, 2016). QRIS are accountability systems, typically administered at the state level that define quality benchmarks for ECE programs and seek to improve quality both through supports and incentives for programs and by providing parents with information about ECE program quality, to help them make informed choices.

It is not yet clear whether public investments in QRIS are leading to meaningful system-wide improvements in ECE program quality. One concern is that the rapid design and rollout of states' QRIS systems has outpaced the research base around accurately measuring quality. In order to lead to quality improvements – and, ultimately, better child outcomes – QRIS must begin by accurately measuring the features of program quality that affect child learning (Cannon, Zellman, Karoly & Schwartz, 2017).

However, despite decades of research on efforts to measure quality in ECE settings, many questions remain about how to do this accurately at scale (Burchinal, 2017). For instance, most states include classroom observations as a component of their QRIS (The Build Initiative and Child Trends, 2016), in part because a large body of evidence demonstrates positive, though modest, associations between these classroom observations and children's learning. However, there is relatively little evidence about the use of classroom observations at scale for policy applications like QRIS (Goffin & Barnett, 2015). It is not yet clear whether measures of classroom quality collected as a part of large-scale policy initiatives capture

* Corresponding author at: University of Virginia, Center for Advanced Study of Teaching and Learning, P.O. Box 800784, Charlottesville, VA 22904, United States.
E-mail address: vev9m@virginia.edu (V.E. Vitiello).

https://doi.org/10.1016/j.ecresq.2018.03.003
0885-2006/

child outcomes as well as do measures collected by researcher-based teams, especially when quality ratings are tied to stakes.

Given the United States' growing investments in classroom observations, which can be costly and time-consuming to collect, it is important to address this gap. Using data from Louisiana, we compare classroom observations collected by local raters to observations conducted by independent data collectors using a standard research protocol. Both teams observed classrooms using the Classroom Assessment Scoring System (CLASS), a widely used measure of teacher–child interactions (Pianta, La Paro, & Hamre, 2008). Their approaches differed, however, in that the research team observed each classroom more frequently, and was also more explicitly focused on strategies to ensure rater reliability. The goal of this study is to assess to what extent the observations conducted according to these different approaches lead to similar conclusions about program quality. The results of this exploratory analysis raise considerations for policy makers determining how to include classroom observations into accountability systems.

1. QRIS & efforts to measure and improve quality at scale

By articulating a clear definition of quality, measuring programs' performance relative to that definition, and providing programs with incentives and supports, QRIS aim to create a culture of improvement (Goffin & Barnett, 2015; Zellman, Perlman, Le, & Setodji, 2008). Because QRIS are relatively new and because in many states they are not implemented at scale, there has not yet been research on the effects of these accountability systems on children's learning.

Several studies do provide encouraging evidence that QRIS can foster program-level changes. For instance, one small randomized control trial in Washington State demonstrated that programs participating in a QRIS with coaching supports demonstrated increases in quality as measured using a widely used observational tool (Boller et al., 2015). A recent study of North Carolina's QRIS system found that quasi-random assignment to a lower quality rating led programs to make notable improvements on a multi-faceted measure of classroom quality (Bassok, Dee, & Latham, 2017).

These studies suggest that at least in some contexts, ECE programs are responsive to the incentives and supports embedded in QRIS. However, for QRISs to foster meaningful change, it must be the case that they define and measure quality in a way that is closely aligned with children's development.

Most QRIS rate programs based on a complex set of factors including structural features (e.g. class size, ratios, teacher credentials), classroom observations (e.g. using Classroom Assessment Scoring System (CLASS; Pianta et al., 2008) or the Early Childhood Environment Rating Scales (ECERS; Harms, Clifford, & Cryer, 1998)), and a host of other measures (e.g. family engagement, administration and business practices, measures of curriculum and assessment use, etc.). States typically use some formula to combine these disparate metrics into a single quality rating, which is typically broken into 3–5 quality levels. Programs scoring above high-level thresholds are publicly recognized as high quality programs in ways that are intended to drive greater enrollment. They often also receive fiscal rewards. Programs scoring at very low levels may receive additional professional development and/or have more punitive sanctions such as a reduction in subsidies.

Ensuring alignment between quality ratings and child outcomes is so central to the QRIS theory of change that, to date, the vast majority of QRIS research has focused on rating validation (Goffin & Barnett, 2015). Existing research suggests that many of the individual metrics included in QRIS are weak predictors of children's learning in ECE settings (Early, Maxwell, Ponder, & Pan, 2017; Mashburn et al., 2008) and that they are not systematically related to child outcomes when grouped together and used to create program ratings (Sabol, Hong, Pianta, & Burchinal, 2013). Further, a growing body of QRIS validation studies has generally found no or inconsistent associations between QRIS ratings and children's outcomes (Cannon et al., 2017; Karoly, 2014).

Across 15 recent reports, four found no differences by QRIS rating (Sirinides, 2010; Tout et al., 2010; Tout, Starr, Albertson-Junkans, Soli, & Quinn, 2011; Zellman et al., 2008), and the rest found small, non-linear associations, that are typically significant for just one skill domain (e.g. Elicker, Langhill, Ruprecht, Lewsader, & Anderson, 2011; Sirinides, Fantuzzo, LeBoeuf, Barghaus, & Fink, 2015; Soderberg, Joseph, Stull, & Hassairi, 2016; Thornburg, Mayfield, Hawks, & Fuger, 2009; Tout et al., 2016).

This lack of predictive validity is a serious threat to the utility of QRIS, and has led to calls for new ways of measuring quality in ECE settings, especially at scale (Burchinal, 2017; Cannon et al., 2017; Karoly, 2014). These calls have focused on the need to simplify quality ratings by focusing on fewer measures that have consistent, demonstrable links with children's learning (Sabol et al., 2013; Sabol & Pianta, 2015).

2. Classroom observations as a tool for quality measurement

One potentially promising quality measure for large-scale accountability systems is the CLASS, a widely used observational measure of teacher–child interactions that assesses effective interactions across ten dimensions divided into three broad domains: Emotional Support, Classroom Organization, and Instructional Support (Pianta et al., 2008). Currently 45% of QRIS systems use the CLASS (The Build Initiative and Child Trends, 2014) and it is also included in Head Start's monitoring system, the Designation Renewal System (DRS; Administration of Children and Families (ACF), 2011). A substantial research base shows a positive relationship between CLASS scores and gains in child outcomes, although these relationships are typically small.

For instance, research using the CLASS indicates that when teachers offer warm, supportive, and responsive interactions, children develop stronger social and emotional skills (e.g., Johnson, Seidenfeld, Izard, & Kobak, 2013). Children in classrooms with strong behavior management and classroom organization demonstrate stronger growth in self-regulation skills (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). Further, teachers' daily provision of cognitively stimulating instruction and conversation appears to be a critical ingredient in fostering academic learning (e.g., Howes et al., 2008). Most compellingly, a recent experiment that randomized children to classrooms within schools showed that young children make greater gains in language, math, and executive functioning skills in classrooms where teachers were more highly rated on the CLASS (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016).

Although the positive relationship between CLASS and various child outcomes has been documented widely – and has motivated many states to include the measure in their QRIS – many questions about the role of CLASS within large-scale accountability remain. First, the associations between CLASS scores and child outcomes tend to be modest. Araujo et al. (2016) found that a standard deviation increase on the CLASS was associated with .07–.11 standard deviation increases in child outcomes; Keys et al. (2013) used *meta*-analytic techniques across multiple studies using various observation measures of quality, including the CLASS, and found an average standardized main effect of .05 on child outcomes. This is not a problem unique to CLASS, however; as a field, we do not yet have measures of quality that are systematically related to moderate or large increases in child outcomes. Still, the relatively

small associations between CLASS scores and child outcomes may raise concerns about the utility of the tool in the policy arena, especially given the time and resources needed to collect classroom observations (Keys et al., 2013).

Second, the research studies that validated the CLASS used researcher-trained raters who typically observed classrooms multiple times and received substantial support to maintain their reliability after initially passing the test. These supports include calibration sessions, during which raters code and discuss additional master coded video, as well as double-coding a subset of observations with a coding partner and then debriefing on any discrepancies that arise. Although the CLASS authors recommend a minimum of four CLASS cycles on a single day of observation (Pianta et al., 2008), research teams often go beyond this collecting multiple days of observation in order to obtain as stable an estimate as possible.

These efforts to ensure the validity of quality measures are important in light of a growing body of research highlighting measurement challenges related to classroom observations. For instance, several recent papers have called attention to the significant rater effects present in these systems (Gargani & Strong, 2014; Mashburn, Downer, Rivers, Brackett, & Martinez, 2014). In one study using the CLASS in elementary schools, rater effects explained between 4% (Classroom Organization factor) and 18% (Instructional Support) of the variance in scores. These studies also demonstrated substantial variability in CLASS scores from day to day or even observation cycle to observation cycle within a single classroom (Gargani & Strong, 2014; Mashburn et al., 2014). These studies echo a much larger body of research demonstrating similar challenges reliably classroom quality through observations in the K-12 sector (Hill, Charalambos, & Kraft, 2012; Ho & Kane, 2013; Kane & Staiger, 2012).

In a scaled-up policy context, getting quality measures "right" may be doubly important given the increasingly "high-stakes" context in which QRIS ratings are publicized and oftentimes tied to monetary incentives. However, the same emphasis on reliability, calibration, double-coding and multiple visits that has been common in research may not be as feasible in this context, due to costs or other constraints.

To date there has been little research comparing observations conducted by researcher-trained observation teams with those collected in practice in large-scale initiatives. One notable exception is Derrick-Mills et al. (2016) which did so in the context of Head Start's DRS.

The DRS uses CLASS scores and administrative data to identify underperforming programs and requires those programs to re-compete for their funding. Researchers compared DRS coders' CLASS scores with CLASS scores collected independently by the research team. CLASS scores assigned by the DRS coders tended to be significantly higher than those assigned by the research team, and correlations between the two teams' domain scores were small to moderate, ranging from $r = .04$ to .53. These findings underscore concerns about the use of observational measures for monitoring in scaled-up settings.

3. Quality rating and improvement system in Louisiana

Louisiana provides a unique opportunity to explore the use of CLASS within a large-scale accountability system because all publicly funded ECE programs, including Head Start, state pre-kindergarten and subsidized child care must participate in the QRIS. Every toddler and pre-k classroom in every publicly funded ECE program is now observed in person using the CLASS at least twice a year. In addition, Louisiana's early childhood accountability system has all the defining components of a QRIS, including quality ratings and financial incentives for programs, supports for improvement, and public information campaigns for parents. Unlike other states, where observational measures of quality are one of a host of quality measures included in the QRIS, in Louisiana, CLASS observations are the only quality measure that is currently used to calculate program ratings. For this reason, the accuracy of the CLASS measures in Louisiana is particularly important.

CLASS observations are coordinated and collected separately in each Community Network, which is typically a group of all the publicly funded early childhood programs in a particular parish (similar to a county) that is coordinated by a single lead agency (e.g. a school district). The state provides funding to lead agencies to conduct CLASS observations and requires that all observers attend a CLASS training and pass the certification test. In the year the current study was conducted, all local networks were asked to aim for two days of observation per classroom and to collect four observation cycles each day. Beyond that, local networks have substantial flexibility with respect to the ways in which they provide ongoing support to raters. Louisiana rating system provides four quality categories, which correspond to ranges of CLASS scores. These categories, rather than the underlying CLASS scores, are the consequential component of their accountability system.

4. Current study

The goal of this study was to assess the extent to which classroom observations conducted "at scale" by local observers correspond to those conducted by independent data collectors using a standard research protocol. To our knowledge, it is the first study to compare local and researcher observations of the same classrooms. The fairness and ultimate impact of QRIS depends upon whether the assumption that the observation methods used in the field are as reliable and valid as those collected for research studies is supported, making this an important area for research.

We explored three related research questions. First, to what extent are classrooms' averaged CLASS scores from local and research-trained teams associated with each other? To address this question we examined correlations and mean differences across the teams' scores. Second, we asked to what extent programs' rating (e.g. unsatisfactory, approaching proficient, proficient, and excellent) differed depending on the rater type to assess whether the two types of raters would lead to different category placement. Finally, we asked whether observations conducted by local and research-trained teams predict child outcomes to the same extent.

Without strong prior research in this area, these questions were largely exploratory. We hypothesized low to moderate correlations across the coding teams and moderate agreement on QRIS categorization. We based this hypothesis on the fact that the rater teams were in classrooms on different days and received different types of support to maintain reliability. In predicting to child outcomes, we expected the research team would show stronger associations than the local raters due to their conducting more days of observation per classroom and receiving a higher level of support from experienced CLASS coders.

5. Methods

5.1. Participants

Primary data occurred during the 2014–2015 school year, when Louisiana was piloting their new accountability system. Our study included 90 programs from five Louisiana parishes, with one classroom serving primarily four-year-olds randomly selected from each program. At the end of the academic year, the Louisiana Department of Education provided the research team

Table 1
Participant characteristics.

	Percent	Mean	SD
Teachers			
Female (n = 85)	98.4		
Race/ethnicity (n = 84)			
White	54.1		
Black/African American	41.2		
Hispanic	2.4		
Other ethnicities	2.3		
Education (n = 83)			
Associate's degree	4.7		
Bachelor's degree	41.2		
Bachelor's degree plus additional coursework	22.4		
Master's degree	21.2		
Beyond a Master's	4.7		
Other	4.7		
–			
Years experience (n = 84)		9.5	8.9
Children			
Female (n = 812)	49.6		
Age (months) (n = 812)		52.6	3.6
Race/ethnicity[a] (n = 673)			
White	20.8		
Black/African American	70.4		
Hispanic	2.7		
Other ethnicities	6.1		
Family income[a] (n = 568)			
$25,000 or less	67.4		
$25,001–55,000	23.9		
$55,001 or more	8.6		
Parent education[a] (n = 657)			
No high school diploma	14		
Diploma or GED	13		
Some college, no degree	30.9		
Associate's degree	11.3		
Bachelor's degree or higher	13.2		

[a] Available for children whose parents completed demographic items on a parent questionnaire.

with CLASS data from the local coding teams. Local coders visited 85 of the 90 classrooms observed by the research coding team; we limit the current analysis to those 85 classrooms. The classrooms were located in public schools (45.9%), private child-care centers (12.9%), Head Start centers (18.8%), public charter schools (11.8%), and private schools and centers receiving local subsidies (10.6%). Directors reported that their programs served diverse student populations, with an average of 73.6% black/African American students (SD = 32.7%, min = 0%, max = 100%), 26.1% white students (SD = 34.1%, min = 0%, max = 100%), and 3.6% Hispanic students (SD = 9.0%, min = 0%, max = 55.3%). All children who were four years of age and had no IEP (except for IEPs related to language delays) were eligible to participate. The child sample included 820 predominantly low-income four-year-old children who were assessed in both the fall and spring. Teacher and child demographic characteristics are presented in Table 1.

5.2. Measures

Classrooms were observed using the CLASS (Pianta et al., 2008). The CLASS is coded in multiple 30 min cycles which include 20 min to observe and record classroom interactions and 10 min to code the CLASS dimensions; codes from each cycle are averaged together to arrive at a single set of classroom scores. Dimensions are coded on a seven-point scale with detailed behavioral descriptors of interactions at the low (1–2), mid (3–5), and high (6–7) ranges of effectiveness. Fifteen percent of observations conducted by the research coding team were double-coded by two data collectors. Intraclass correlations indicated high levels of agreement between coders (Emotional Support, ICC = .812; Classroom Organization, ICC = .878, Instructional Support, ICC = .883, total score, ICC = .902). Internal consistency estimates were strong across both rater teams, with Cronbach's alphas ranging from .77 to .96.

Children were directly assessed using tests of language, literacy, math, and executive functions. The Peabody Picture Vocabulary Test-4th edition (PPVT-IV) was used to measure children's receptive vocabulary skills (Dunn & Dunn, 2007). Children are shown four pictures and are asked to identify the picture that corresponds to a word stated by the tester. The psychometric properties of the PPVT-IV are well established, with evidence for strong reliability and validity (Dunn & Dunn, 2013).

The Woodcock–Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) is a widely used achievement battery that can be used with individuals from age 2 to adulthood. This study used three subtests: Picture Vocabulary, which measures expressive vocabulary, and Applied Problems and Quantitative Concepts, which measure math knowledge and reasoning, including problem solving, analysis, and vocabulary. The WJ-III has high split-half reliabilities and shows strong concurrent validity with other tests of achievement (Schrank, McGrew, & Woodcock, 2001).

The Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen,& Rashotte, 2007) is an assessment battery designed to assess preschool children's emergent literacy skills. Two of the three TOPEL subtests were used in the assessment battery: the Phonological Awareness subtest, which assesses word elision and blending ability, and the Print Knowledge, which assesses children's knowledge of the alphabet, written language conventions, and writing form. TOPEL subtests have shown good internal consistency (Cronbach's alphas ranging from .78 to .89) and concurrent validity ranging from .41 to .43 (Lonigan, Keller, & Phillips, 2004).

Children's executive functions were assessed using two tasks. The Pencil Tap test measures inhibitory control and asks children to tap a pencil once when the assessor taps twice, and vice versa (Smith-Donald, Raver, Hayes, & Richardson, 2007). The Head Toes Knees Shoulders task (HTKS; Ponitz et al., 2008) is a measure of inhibitory control, working memory, and attention, in which the child must do the opposite of what the assessor says (e.g., touch their head when the assessor says "Touch your toes"). Both tasks have been widely used to assess preschool children (Smith-Donald et al., 2007; Ponitz et al., 2008).

5.3. Procedure

In collaboration with the state department of education, the research team selected five Louisiana parishes to participate in the study that captured the geographic and demographic diversity of the state. The Department of Education provided the research team with a list of all preschool programs receiving public funds in each of the five parishes, from which the researchers randomly selected 90 programs, stratified by parish and program type. Ten programs declined to participate and were replaced, with 6–36 programs per parish (the number of programs per parish was determined based on the size of the parish) and program acceptance rates by parish ranging from 84 to 100%.

Within each program, all teachers of classrooms serving primarily four-year-olds and typically developing children were randomly ordered and the first teacher from each program was contacted. Six teachers declined to participate or were later found to be ineligible because their classrooms were not serving a majority of typically developing four-year-olds, so the next eligible teacher on the randomized list was contacted. If there were no other eligible teachers at the program, the program was dropped and another program was contacted as a replacement. Four teachers left their classrooms

during the year and were replaced with the teacher who took over teaching responsibilities for that classroom.

All research and local raters completed the Teachstone CLASS Observation Training and passed the reliability test with a minimum score of 80% agreement within one point of master codes. The research coding team was required to complete one day of live coding practice with an experienced coder, two calibration sessions during data collection, and frequent double-coding and debrief sessions with fellow raters that comprised 15% of observations. The research coding team did not personally know the teachers that they observed outside of the context of the research study and were not involved in Louisiana's QRIS initiative.

Parishes varied in their approaches to supporting local raters. As a part of this study, we worked with LDOE to determine parishes' training and calibration procedures. Three parishes reported that they did some proportion of double coding, and four of the five reported that raters sometimes, often, or almost always knew the teachers they observed. The research coding team visited each classroom an average of 3.94 times (SD = .28); local raters visited an average of 1.47 times (SD = .59). Both teams coded four CLASS cycles per visit. Internal consistency estimates were strong across both coding teams, with Cronbach's alphas ranging from .77 to .96.

Direct assessments were completed by the research team data collectors in the fall and spring after a two-day in-depth training. Children were assessed individually in a single session lasting approximately 45 min in a quiet location outside of the classroom, as free from distractions as possible.

5.4. Analytic approach

We mirror Louisiana's process for calculating QRIS by averaging CLASS scores for each team across cycles and days to arrive at a single score per classroom per team. Louisiana's QRIS currently uses the following cut points for the CLASS total score to categorize programs: Unsatisfactory: 1–2.99; Approaching Proficient: 3–4.49; Proficient: 4.50–5.99; Excellent: 6–7. We used the same cut points to compare alignment across raters.

To analyze the child assessment data, we used MPlus to construct latent variables representing fall and spring math, literacy (including language scores), and executive functions. We also tested an overall composite of child achievement with all assessments loaded onto single fall and spring factors. Separate models were run for each outcome measure, first using the local CLASS scores, and then again using the researchers' CLASS scores. We predicted spring achievement controlling for fall scores and clustering standard errors at the classroom level and used the Wald test to determine whether the strength of associations was stronger for one team than the other.

6. Results

6.1. Associations between local and research-trained coders' CLASS scores

6.1.1. Means

Domain and total score means from each coding team are presented in Table 2.

Paired samples *t*-tests suggested that Total CLASS scores were higher among local raters compared to the research team ($p = .01$). Local raters' Instructional Support scores were roughly three quarters of a point higher than the research team's ($p < .001$). For Classroom Organization and Emotional Support, there were no statistically significant differences between how local raters and the research team coded classrooms. Notably, local raters showed greater variability in their scores compared with the research team (SDs ranged from .68 to 1.19 for the local raters, and .61–.86 for the research team raters).

Table 2
CLASS score means and standard deviations ($n = 85$).

	Local raters				Research team raters				
	Mean	Std. dev	Min	Max	Mean	Std. dev	Min	Max	
Emotional Sup.	5.78	.68	3.00	7.00	5.72	.61	3.63	6.78	
Classroom Org.	5.32	.82	2.75	6.92	5.43	.75	2.81	6.78	
Instructional Sup.	3.73	1.19	1.42	6.46	2.97	.86	1.33	6.19	***
Total score	4.94	.80	2.73	6.65	4.71	.66	2.62	6.55	**

** $p < .01$.
*** $p < .001$.

Table 3
Correlations between domain and total scores across coding teams ($n = 85$).

		Local raters			Research team			
		ES	CO	IS	Total	ES	CO	IS	Total
Local raters	ES	–							
	CO	.77	–						
	IS	.64	.65	–					
	Total	.87	.89	.90	–				
Research team	ES	.27	.35	.17	.29	–			
	CO	.24	.43	.30	.36	.82	–		
	IS	.29	.38	.21	.32	.69	.56	–	
	Total	.30	.44	.26	.36	.92	.89	.87	–

Note: All correlations are statistically different from zero with $p < .05$ except the correlation between the Research Team Emotional Support and Local Rater Instructional Support which has a $p = .11$. ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support; Total = Total CLASS Score.

6.1.2. Correlations

Domain and total score correlations across teams are presented in Table 3. Within each coding team, domain scores were moderately to highly correlated with each other, ranging from $r = .56$ to $r = .82$ for the research team and $r = .64$ to $r = .77$ for the local raters. Across coding teams, correlations of domain scores were lower but were significantly correlated with one exception: the correlation between the research team's Emotional Support and the local raters' Instructional Support did not reach significance ($r = .17$, $p = .11$). Overall CLASS scores were correlated at $r = .36$. The remaining correlations ranged from $r = .21$ to $r = .43$.

6.2. Agreement on program ratings

CLASS total scores from both rater teams were divided into quality categories using the Louisiana QRIS cut points (Table 4). Results indicated that 55.3% of classrooms placed into the same category by both teams (Chi-squared statistic = 22.5, *p*-value < .01). In approximately 27% of classrooms the local team rated the classroom in a higher category than the research team; the reverse was the case in 18% of classrooms. The lowest agreement was in the highest and lowest categories. Local coders categorized three programs as "Unsatisfactory," research coders categorized two, and there was overlap on one program. There was no agreement on membership in the "Excellent" category; local coders placed seven classrooms in the "Excellent" category and the research team placed two, but no classroom was rated "Excellent" by both.

6.3. Prediction to child gains

Results from regression analyses predicting child learning outcomes are presented in Table 5. Both teams' CLASS scores showed significant prediction to some child gains. However, the pattern of significance differed. For the local raters, Classroom Organization and Instructional Support were significantly associated with chil-

Table 4
QRIS category frequencies and agreement across raters ($n = 85$).

		Research Team				
		Unsatisfactory	Approaching Proficient	Proficient	Excellent	Total
Cut points		1 – 2.99	3 – 4.49	4.50 – 5.99	6 - 7	
Local Coders	Unsatisfactory	1 (1.2%)	2 (2.4%)	0 (0%)	0 (0%)	3 (3.5%)
	Approaching Proficient	1 (1.2%)	5 (5.9%)	11 (12.9%)	0 (0%)	17 (20.0%)
	Proficient	0 (0%)	15 (17.65%)	41 (48.2%)	2 (2.4%)	58 (68.2%)
	Excellent	0 (0%)	0 (0%)	7 (8.2%)	0 (0%)	7 (8.2%)
	Total	2 (2.4%)	22 25.9%)	59 (69.4%)	2 (2.4%)	85 (100%)

Chi-squared statistic = 22.51, p-value = .007.
Note: Percentages represent the percent of the total that falls into each cell. Categories are based on the cut points chosen for Louisiana's accountability system.

Table 5
Associations between ratings and child achievement gains ($n = 820$).

Local Raters	Math		Literacy		Executive Function		Achievement Average	
Emotional Sup.	.065		.063		.045		.057	
	(.046)		(.039)		(.061)		(.041)	
Classroom Org.	.110	*	.093	*	.083		.092	*
	(.050)		(.037)		(.052)		(.043)	
Instructional Sup.	.177	***	.105	**	.107	*	.137	**
	(.047)		(.040)		(.051)		(.042)	
CLASS Total	.139	**	.100	*	.091		.112	*
	(.048)		(.040)		(.055)		(.043)	
Research Team								
Emotional Sup.	.078		.004		.123	**	.049	
	(.045)		(.033)		(.043)		(.039)	
Classroom Org.	.177	***	.067		.162	***	.129	**
	(.048)		(.035)		(.042)		(.041)	
Instructional Sup.	.004		-.017		.017		-.013	
	(.047)		(.031)		(.042)		(.037)	
CLASS Total	.091	*	.016		.109	*	.056	
	(.044)		(.032)		(.043)		(.038)	

Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. Shaded cells reflect coefficients that are significantly different for the research team compared to the local raters.

dren's math, literacy, and achievement average gains. Only local raters' Instructional Support scores were associated with gains in executive functions. For the research team, Classroom Organization predicted math, executive functions, and average achievement gains. Emotional Support predicted only executive functions, and Instructional Support scores were not associated with any of the child outcomes. For both groups we observed that the CLASS total score, which is oftentimes the relevant measure in monitoring and accountability systems, was associated with children's learning gains. For the local raters, this measure was positively linked to math, literacy and average achievement gains. For the research group, the total score was significantly related to math and executive function gains.

Overall, for 10 of the 16 relationships considered, the coefficients were not statistically distinguishable across the two groups. There were 6 cases where the coefficients did diverge, 4 of which suggested a stronger association for the local team and 2 suggesting a stronger association for the research team. Specifically, associations between Instructional Support and child outcomes tended to be stronger for the local raters, while Classroom Organization asso-

ciations were stronger for the research team. Notably, across teams, the standardized effects tended to be small in magnitude, with the largest at .177.

7. Discussion

With the rapid expansion of QRIS nationwide, and the increasing use of observational measures within them, there is a need for research that can guide decision-making in ways that help ensure the fairness, reliability, and accuracy of data (Lahti, Elicker, Zellman, & Fiene, 2015). The use of local observers to conduct classroom observations offers several notable potential benefits (e.g., saving money, gaining local buy-in), but their use may also create unintended consequences if they produce biased or unreliable scores. The goal of this paper was not only to examine similarities but also to find any differences between classroom ratings based on local observations and those based on observations conducted by a research team. To address this question, we examined correlations and mean differences across the teams' scores. Second, we asked to what extent programs' rating (e.g. unsatisfactory, approaching proficient, proficient, and excellent) differed depending on the rater type to assess whether the two types of raters would lead to different category placement. Finally, we asked whether observations conducted by local and research-trained teams predict child outcomes to the same extent.

The results provide a mixed picture. On one hand, the average CLASS scores from the local and research observers were statistically indistinguishable for two of the three domains – Emotional Support and Classroom Organization. The pattern of correlations suggests that the two teams were broadly tapping into a common set of behaviors and interactions in assigning codes. On the other hand, local raters gave programs systematically higher ratings on Instructional Support, and in turn the overall scores from the local raters were somewhat higher and more variable.

The correlations across teams were smaller than would be expected. For context, correlations between the research teams' domain scores on double-coded cycles were $r = .66$ (Emotional Support), .76 (Classroom Organization), and .75 (Instructional Support), while the cross-team correlations were .27, .43, and .21, respectively. Importantly, the double-coded data were from two observers coding the same 20 min side-by-side while the cross-team ones were from observations on different days. Our low cross-team correlations are similar to the cross-team correlations found in the Head Start DRS evaluation, which had a range from $r = .04$ to $r = .53$ (note that they examined correlations within programs above and below the DRS threshold and found stronger associations above the threshold; Derrick-Mills et al., 2016). Still, the relatively small associations could be a cause for concern and highlight the need for more study.

From a policy perspective, the most notable divergence was that nearly half of classrooms were put into different QRIS categories, and there was no concordance in the highest rating category. Notably, assigning the scores that landed classrooms in these "extreme" categories was very uncommon for both rater teams – and is uncommon in practice, as well (Ho & Kane, 2013). Given these low rates, as well as the small number of classrooms examined in this study, the lack of alignment across rating team on the lowest and highest categories should be viewed cautiously. Still, it is worth noting that this pattern mirrors results from a recent evaluation of the Head Start DRS, which found that programs recommended for sanctions under the DRS system – recommendations that were made largely based on CLASS scores – did not differ significantly from other programs on the CLASS when CLASS data were drawn from research team observations (Derrick-Mills et al., 2016). In other words, external observers found no significant differences between programs above and below the thresholds set by the Office for Children, despite the fact that the thresholds were highly consequential for programs.

Like the current study, the Head Start DRS evaluation also noted a tendency for the monitoring team to assign higher scores than the research team (Derrick-Mills et al., 2016). This finding is in line, as well, with prior research on teacher evaluation in K-12, which indicated that same-school administrators tended to assign higher scores to classrooms compared with administrators from other schools (Ho & Kane, 2013).

There are several possible explanations for this pattern, which will require further research to unpack. One is that observers, who believe that their scores are consequential may be reluctant to assign low scores or tend to inflate the scores they assign. This interpretation is in line with the finding from the Head Start DRS study, in which neither team of observers knew the teachers personally, but the monitoring team was aware that their scores had consequences for teachers and programs.

Another explanation is that observers who are more familiar with teachers bring some of that knowledge into their coding. For example, an observer may watch a 20-min cycle during which students are drawing fish on a banner. A trained external observer might watch for questions that prompt higher-order thinking, or conversations that elicit children's prior knowledge, and give the cycle low Instructional Support scores if those interactions are not present. A local coder might know that the teacher is doing a long, project-based unit on the ocean and, therefore, see this activity as something that taps into what the children have learned about underwater ecosystems, and therefore assign higher codes.

This is interpretation is in line with the Ho and Kane (2013) finding, since neither of their teams' scores were consequential for teachers but same-school coders still tended to assign higher scores. Alternatively, another possible explanation for the higher ratings among more local ratings may be that their existing relationships with teachers or administrators may make it difficult to be impartial. More research is needed into these mechanisms to better understand how and why this is occurring.

The current study suggests that classroom observations of teacher–child interactions do demarcate important elements of children's classroom experiences. Classrooms scoring higher on the CLASS, as assessed by either local or research raters, had children that made greater gains across domains of development. Thus, classrooms that receive higher CLASS ratings through the Louisiana QRIS appear to be supporting better outcomes for students, although substantial differences in CLASS scores were associated with small differences in child gains. Looking at the research team's ratings, a difference of one standard deviation on the CLASS was associated with child gains ranging from −.02 to .18 standard deviations, with a median association of about .09. This is largely in line with the results of a recent *meta*-analysis, which found that classroom quality observations had an association of .05 with children's language outcomes, .03 on math, and a non-significant .02 association with social skills (Keys et al., 2013).

The pattern of prediction to child outcomes differed across the two rating teams. Most notably, although the local teams rated classrooms significantly higher on Instructional Support than the research team did, their scores were consistently associated with greater achievement gains and the research team's were not. The research team's ratings of Classroom Organization were more strongly associated with gains in math and executive functions than the local raters' codes. These differences are not straightforward to explain and may again require more research into the reasons that different teams of raters code differently. The Instructional Support findings may lend support to the idea that the local raters knew something about teachers' instruction that the research team did not, leading to higher scores and stronger associations with child

outcomes. In a sense, this would make local raters "less reliable" since raters are explicitly trained not to consider anything outside of the direct observation in assigning scores. It does raise interesting questions about the use of local or knowledgeable raters versus independent raters, though, where there may be tradeoffs between reliability as we traditionally understand it and validity in predicting scores.

8. Limitations

Several caveats about this work are worth noting. These data were collected during a year in which Louisiana was piloting the use of CLASS. The CLASS observations were not tied to incentives or consequences in that year, though they currently are. This context has two important implications. First, the pressure to assign higher codes may be more pronounced when the observations are attached to greater stakes. Second, conducting the CLASS observations was a new responsibility for lead agencies during the pilot year. For the most part, local raters were new to the CLASS measure. Further, at that time, there were no requirements that local community networks ensure the reliability of observations beyond requiring the initial certification. It is worth noting that in the years since Louisiana has put into place a number of measures to help ensure reliability. Most importantly, the state now sends, independent, "third-party" observers commissioned by the state to conduct observations in classrooms at every site, and uses those third-party observations when there is a significant divergence with local ratings. The state also developed stringent guidelines for ensuring the reliability of these third-party observers, which may improve the accuracy of those ratings and prevent coding drift over time (Karoly, Zellman, & Perlman, 2013). It will be important to continue examining these issues in Louisiana as they move into a more high-stakes version of their accountability system.

An additional set of limitations relate to the way data were collected for this study. First, the sample size is small: future studies will need to evaluate the reliability and validity of local ratings using the full set of state observations which now include multiple CLASS ratings from every classroom serving toddlers and preschoolers in publicly funded early childhood programs. Second, reliability and validity were assessed here at the classroom level and only in classrooms primarily serving four-year-olds. Most QRIS (including Louisiana's) assign ratings at the program level and include ratings for infant and toddler classrooms as well.

Finally, our study compared local and researcher ratings, but did not disentangle what factors drove differences across groups. The two teams' ratings likely differed for a variety of reasons including (1) the different levels of supports and practice the teams received to ensure reliability; (2) differences in the make-up of the teams themselves, including education and experience; (3) the different number of days each team observed; and the fact that (4) each team was observing the classroom on different days (Casabianca et al., 2013). Isolating the role of these (and other) factors is essential for determining policy implications (Casabianca et al., 2013).

9. Conclusion

Louisiana's decision to develop an ECE accountability system, focused on teacher–child interactions, was motivated by existing research that demonstrated positive (though small) associations between CLASS scores and child outcomes, as well as research demonstrating a lack of association between other states' ratings and child outcomes. The findings from the current study suggest this approach has promise. It is also encouraging that the scores from local raters were associated with child outcomes.

At the same time, the inconsistent alignment between coding teams and the small associations between CLASS scores and gains in outcomes, suggest caution is warranted when incorporating local observation scores into QRIS ratings. Strong data quality procedures should be in place to ensure the best possible data, including reliability testing procedures, calibration opportunities during data collection, and frequent checks on the data to make sure scores from different teams are well aligned. The small associations between CLASS scores and child gains suggest that the field should continue to look for other markers of effective programs that might be incorporated into QRIS, such as the use of evidence-based curricula (Burchinal, 2017).

For researchers, this work highlights the need for more research that can inform QRIS development and decision-making. There are many new practices being included in these systems that have not been adequately studied, and there are significant policy implications for this work. Specific to classroom observations, future work may focus on understanding the factors that affect reliability such as the timing of observations, the number of classroom visits, the calibration processes in place, and how raters are assigned to classrooms.

Conflict of interest

Bridget Hamre is the cofounder and part owner of Teachstone Training Inc., a company that was founded to help implement the CLASS and aligned professional development programs. Dr. Hamre complies with all university policies regarding managing conflict of interest.

Acknowledgements

This research was supported by a grant from the Institute of Education Sciences (R305A140069). Opinions reflect those of the authors and do not necessarily reflect those of the granting agency. We thank the Louisiana Department of Education for their willingness to share data for this project, and the children, teachers, and families who generously agreed to participate in this study.

References

Administration for Children and Families (ACF). (2011). Head Start program performance standards, 45 CFR chapter XIII, part 1304. *Subpart B — designation renewal,*. Retrieved from http://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/part-1304-federal-administrative-procedures.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics, 131*(3), 1415–1453.

Bassok, D., Dee, T., & Latham, S. (2017). *The effects of accountability incentives in early childhood education. NBER Working Paper No. 23859*. Cambridge, MA: National Bureau of Economic Research.

Boller, K., Paulsell, D., Del Grosso, P., Blair, R., Lundquist, E., Kassow, D. Z., et al. (2015). Impacts of a child care quality rating and improvement system on child care quality. *Early Childhood Research Quarterly, 30*, 306–315.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25*, 166–176.

Cannon, J. S., Zellman, G. L., Karoly, L. A., & Schwartz, H. L. (2017). *Quality rating and improvement systems for early care and education programs: Making the second generation better*. Washington, DC: Rand Corporation.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*, 757–783.

Derrick-Mills, T., Burchinal, M., Peters, H. E., De Marco, A., Forestieri, N., Fyffe, S., et al. (2016). . *Early implementation of the head start designation renewal system OPRE Report #: 2016-75a* (Vol I) Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Dowsett, C. J., Huston, A. C., Imes, A. E., & Gennetian, L. (2008). Structural and process features in three types of child care for children from high and low income families. *Early Childhood Research Quarterly, 23*(1), 69–93. http://dx.doi.org/10.1016/j.ecresq.2007.06.003

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test –4th edition*. Minneapolis, MN: Pearson Assessments.
Dunn, L. M., & Dunn, D. M. (2013). *PPVT-4 technical report*. Minneapolis, MN: Pearson Assessments.
Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher–child interactions: A randomized controlled trial of making the most of classroom interactions and my teaching partner professional development models. *Early Childhood Research Quarterly*, *38*, 57–70. http://dx.doi.org/10.1016/j.ecresq.2016.08.005
Elicker, J. G., Langill, C. C., Ruprecht, K. M., Lewsader, J., & Anderson, T. (2011). *Evaluation of—paths to QUALITY,‖ Indiana's child care quality rating and improvement system: Final report (Technical Report 3)*. West Lafayette, IN: Purdue University.
Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. *Journal of Teacher Education*.
Goffin, S. G., & Barnett, W. S. (2015). Assessing QRIS as a change agent. *Early Childhood Research Quarterly*, *30*, 179–182.
Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale*. New York, NY: Teachers College Press.
Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.
Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel research paper*. Seattle, WA: MET Project, Bill & Melinda Gates Foundation.
Johnson, S. R., Seidenfeld, A. M., Izard, C. E., & Kobak, R. (2013). Can classroom emotional support enhance prosocial development among children with depressed caregivers? *Early Childhood Research Quarterly*, *28*(2), 282–290.
Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Policy and practice brief. MET project*. Bill & Melinda Gates Foundation. Retrieved from. http://eric.ed.gov/?id=ED540962
Karoly, L. A. (2014). *Validation studies for early learning and care quality rating and improvement systems: A review of the literature*. Washington, DC: Rand Corporation.
Karoly, L. A., Zellman, G. L., & Perlman, M. (2013). Understanding variation in classroom quality within early childhood centers: Evidence from Colorado's quality rating and improvement system. *Early Childhood Research Quarterly*, *28*(4), 645–657.
Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., et al. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171–1190. http://dx.doi.org/10.1111/cdev.12048
Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*, 280–290. http://dx.doi.org/10.1016/j.ecresq.2014.04.005
Lonigan, C. J., Keller, K., & Phillips, B. M. (2004). Assessment of children's pre-literacy skills. In B. Wasik (Ed.), *Handbook on family literacy: Research and services*. Mahwah, NJ: Erlbaum.
Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of preschool early literacy (TOPEL)*. Austin, TX: Pro-Ed.
Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*(2), 146–155.
Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.
Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. P., et al. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects*. Washington, DC: Brookings Institute.
Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451.
Ponitz, C. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, *23*, 141–158. http://dx.doi.org/10.1016/j.ecresq.2007.01.004
Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental psychology*, *45*(4), 958.
Sabol, T. J., Hong, S. S., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-K programs predict children's learning? *Science*, *341*(6148), 845–846.
Sabol, T. J., & Pianta, R. C. (2014). Do standard measures of preschool quality used in statewide policy predict school readiness? *Education Finance and Policy*, *9*(2), 116–164.
Sabol, T. J., & Pianta, R. C. (2015). Validating Virginia's quality rating and improvement system among state-funded pre-kindergarten programs. *Early Childhood Research Quarterly*, *30*, 183–198. http://dx.doi.org/10.1016/j.ecresq.2014.03.004
Schrank, F. A., McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III Technical abstract*. Itasca, IL: Riverside.
Sirinides, P. M., Fantuzzo, J., LeBoeuf, W. A., Barghaus, K. M., & Fink, R. (2015). *An Inquiry into Pennsylvania's Keystone STARS*. Philadelphia, PA: Consortium for Policy Research in Education.
Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, *22*, 173–187. http://dx.doi.org/10.1016/j.ecresq.2007.01.002
Soderberg, J., Joseph, G. E., Stull, S., & Hassairi, N. (2016). *Early achievers standards validation study: Final report*. Olympia, WA: Washington State Department of Early Learning.
The Build Initiative & Child Trends. (2016). *A catalog and comparison of quality rating and improvement systems (QRIS) [Data system]*.. Retrieved from. http://qriscompendium.org/on 8/9/2016
Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009). *The Missouri Quality rating system school readiness study: Executive summary*. Kansas City, MO: Center for Family Policy & Research University of Missouri and the Institute for Human Development University of Missouri.
Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). Compendium of quality rating systems and evaluations: The child care quality rating system (QRS) Assessment. *Administration for Children & Families*.
Tout, K., Starr, R., Albertson-Junkans, L., Soli, M., & Quinn, K. (2011). *Evaluation of parent aware: Minnesota's quality rating system pilot: Final evaluation report*. Minneapolis, MN: Child Trends.
Tout, K., Cleveland, J., Li, W., Starr, R., Soli, M., & Bultnick, E. (2016). *Parent Aware: Minnesota's quality rating and improvement system: Initial validation report*. Bethesda, MD: Child Trends.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.
Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the qualistar early learning quality rating and improvement system as a tool for improving child-care quality [Product page]*.. Retrieved from. https://www.rand.org/pubs/monographs/MG650. html