Validating Virginia’s quality rating and improvement system among state-funded pre-kindergarten programs

Terri J. Sabol a,*, Robert C. Pianta b

a Northwestern University, School of Education and Social Policy, Walter Annenberg Hall, 2120 Campus Drive, Evanston, IL 60208, United States
b University of Virginia, Curry School of Education, 405 Emmet St S, Charlottesville, VA 22904, United States

A R T I C L E   I N F O

Article history:
Available online 8 April 2014

Keywords:
Quality Rating and Improvement Systems (QRIS)
Validation
Pre-literacy skills
Pre-kindergarten
Early childhood education quality

A B S T R A C T

Early childhood education programs’ quality ratings are the key output of Quality Rating and Improvement Systems (QRIS), yet there is little empirical evidence as to whether these ratings relate to child outcomes. The present study examines associations between state-funded targeted pre-kindergarten programs’ ratings in Virginia’s QRIS and children’s growth in pre-literacy skills through kindergarten (n = 2448 children in 71 pre-k programs). Children in higher-rated pre-kindergarten programs had sharper literacy growth in the preschool year compared to children in lower-rated pre-kindergarten programs. There was no difference in children’s growth from the spring of preschool to the fall of kindergarten or during the kindergarten year as a function of pre-kindergarten programs’ ratings. Implications of these findings are discussed toward understanding the potential of QRIS to assess quality in pre-kindergarten programs.

© 2014 Elsevier Inc. All rights reserved.

At the state and federal level, policymakers are struggling to improve the quality of early childhood education settings. The increased effort to improve the quality of early childhood care partly is motivated by the strong evidence that high-quality non-parental care substantially improves outcomes for young children (Camilli, Vargas, Ryan, & Barnett, 2010). Yet, a substantial number of children are attending early care that is mediocre at best, even among regulated pre-kindergarten programs (Barnett, Carolan, Fitzgerald, & Squires, 2012). Moreover, disadvantaged children are disproportionally exposed to programs with lower quality, often participating in programs with larger class sizes, less educated teachers, and classrooms typified by lower relational quality (Burchinal & Cryer, 2003).

Quality Rating and Improvement Systems (QRIS) address the need for systematic and widespread improvement of quality by pushing beyond basic licensure requirements within individual states. QRIS not only provide the metrics to assess quality, but also provide levers to facilitate improvement in the quality of early childhood education. To do so, they clearly delineate quality standards, assess and monitor performance through ratings, and incentivize and support quality improvement based on programs’ ratings. The logic model for QRIS posits that early childhood education program ratings will create a local market for high-quality care with parents having greater demand for higher-rated programs. This local demand will motivate providers to attempt to improve their rating by improving the quality of their program, consequently incentivizing parents to select their program. The increased motivation and support of early childhood education programs to improve quality will expose more children to higher quality and lead to better outcomes for children (Tout et al., 2010; Tout, Zaslow, Halle, & Forry, 2009; Zellman, Perlman, Le, & Setodji, 2008).

Almost every state in the U.S. is developing or implementing a QRIS (Tout et al., 2010; QRIS National Learning Network, 2013). Despite the rapid expansion of QRIS, there is virtually no empirical evaluation of their operation or the alignment of QRIS ratings to empirical evidence of their effects (Tout et al., 2009). Because empirical evaluations of the QRIS are still in their nascent stages, a good foundation for work on the effectiveness of QRIS will rely
on the validation of the QRIS components. Better understanding whether the rating structure utilized in the QRIS measure aspects of quality that relate to children’s functioning and development will help ensure the use of valid measures and rating structures, as well as support future empirically sound investigations on the effectiveness of QRIS for improving school readiness. The present study validates the rating system in Virginia’s QRIS, the Virginia Star Quality Initiative (VSQI). We examine the association between star ratings and children’s literacy skills upon school entrance, as well as associations with children’s growth in literacy skills across preschool and kindergarten.

1. Background on Virginia’s QRIS: The Virginia Star Quality Initiative (VSQI)

Virginia’s QRIS, the Virginia Star Quality Initiative, offers a useful setting for QRIS validation work as it defines ratings using quality standards with strong empirical support, uses a unique rating structure to create star ratings, and is a relative newcomer to the QRIS landscape. Piloted in 2007 and officially implemented in 2008, the main goal of the VSQI is to provide a consistent way to distinguish the level of quality in early education programs within the Commonwealth of Virginia. Similar to other QRIS, the VSQI is a volunteer system in which communities elect to participate in the system and early childhood education programs volunteer to participate within that community (Tout et al., 2010).

The quality standards and rating structure used in Virginia’s rating system have some important differences and similarities to other QRIS. Programs are assessed on four quality indicators: (1) education, qualifications, and training of staff; (2) teacher-child interactions; (3) structure (i.e., staff-to-child ratio); and (4) environment and instruction (see Appendix Table A.1). Every QRIS in the country uses some form of staff qualifications as a quality standard and the VSQI is no exception. Additionally, almost all states, including the VSQI, use scales from Environment Rating Scales to measure environment and instruction (Harms, Clifford, & Cryer, 2005). Half of all QRIS measure staff-to-child ratio or group size, although the cut-points vary substantially between states. Differing from other QRIS, the VSQI measures the quality of teacher-child interactions through use of the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), which is gaining popularity as a measure of quality, but is a much less used observational tool compared to the Environmental Rating Scales.

In the VSQI, the four quality standards are converted to a star rating (out of five stars) using a point system, in which points for each standard are summed. Ratings are calculated based on the centers’ total points falling within the range for a specific rating level with no required minimum score for individual quality standards. Only four other states use a point system to generate program ratings (Tout et al., 2010). A much more common approach to converting indicators of quality to ratings is a building block approach in which states prescribe a number of mandatory standards that a program must meet in order to achieve a certain level or rating.

2. Validation of QRIS rating system and quality standards

There are many potential outcomes of interest of QRIS, such as providing a quality framework, unifying early childhood education programs under one system, or promoting professionalism of the early educator workforce. This paper focuses on the ways in which QRIS ratings relate to children’s development, building off the assumption that child outcomes are a key outcome of QRIS.

Past research provides little evidence on whether we may expect that the aggregate star ratings relate to child outcomes. Most validation investigations on other states have examined correlations among QRIS quality standards, or correlations among QRIS quality standards and outside measures of quality. These methods focus on internal concurrent validity of measures rather than linkages to outcomes. For example, studies in Oklahoma and North Carolina found that a program’s ratings based on the structural elements of the program, such as licensure status and group size, were related to independent observational assessments of structural and process quality (Bryant, 2001; Norris, Dunn, & Eckert, 2003). On the other hand, a study conducted in Colorado’s QRIS found few associations between quality standards that measured structural quality and independent measures of process quality (Zellman et al., 2008).

To date, a few studies have linked quality standards to child outcomes, with mixed findings (Ma et al., 2013; Norris et al., 2003; Thornburg, Mayfield, Hawks, & Fugger, 2009). For example, Thornburg et al. (2009) found an association between star ratings and children’s socioemotional development in Missouri’s QRIS. On the other hand, Zellman et al. (2008) evaluated Colorado’s QRIS using a variety of child outcomes, including academic performance, language skills, and classroom behavior, and failed to find evidence that individual QRIS quality standards were associated with children’s preschool performance. Additionally, they found no consistent relations between star ratings and child outcomes. However, due to relatively high levels of child attrition, it is difficult to make any broad generalizations to other states’ systems.

2.1.1. Relation between VSQI quality standards and child outcomes

Despite the lack of research exploring relations between QRIS aggregate ratings and child outcomes, the extant literature has found links between the individual quality standards employed in the VSQI and a range of positive developmental outcomes for children, including improved cognitive ability, socioemotional skills, and language development (Burchinal, Kainz, & Cai, 2011; Hamre & Pianta, 2005; NICHD ECCRN, 2002). Yet, the magnitude of effects varies substantially across quality standards. Below we review findings that examine the relations among the four quality standards employed in the VSQI and outcomes.

2.2. Standard 1: education, qualifications, and training

The VSQI determines the level of quality for the first standard using the education level, years of experience, and professional development/training of directors, teachers, and assistant teachers within the center. Research linking structural features of teacher qualifications to child outcomes has often yielded equivocal findings. Early et al. (2006) failed to find that teacher education, particularly having a Bachelor’s degree or higher, was directly related to children’s achievement gains. However, there may be indirect relations between teachers’ qualifications and outcomes. McDonald-Connor, Son, Hindman, and Morrison (2005) found that teachers with more years of education were associated with teachers with greater warmth. In turn, teachers with more warmth had students with stronger vocabulary skills. Teachers’ years of experience and their participation in targeted professional development experiences can also predict teacher quality and effectiveness (Cloftel, Ladd, & Vigdor, 2007). In terms of director qualifications, directors with more education often provide more effective support for teachers, particularly when teachers have lower levels of education. There is little evidence that director or assistant teacher qualifications directly relate to child outcomes (Zellman & Perlman, 2008).
2.3. Standard 2: interactions

The VSQI uses the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, et al., 2008) to measure the quality of interactions within the classroom. The CLASS assesses the emotional support, classroom organization, and instructional support between teachers and children. The quality of teacher-child interactions is related to children’s language development, general cognition and executive functions, socioemotional development, and approaches to learning (Howes et al., 2008; Maier, Vitiello, & Greenfield, 2012; Mashburn et al., 2008) and may even have a compensatory influence for children with known social and familial risk (Hamre & Pianta, 2005).

2.4. Standard 3: staff-to-child ratio and group size

The third standard is measured based on staff-to-child ratio and group size. Experimental and quasi-experimental studies suggest an important link between lower class size and staff-to-child ratio and children’s performance in early childhood and elementary school (Chetty et al., 2011; NICHD ECERS, 2002). In addition, lower class size in early elementary school yields positive effects on later life outcomes, such as high school graduation, college enrollment, and likelihood of homeownership (Chetty et al., 2011; Dynarski, Hyman, & Schanzenbach, 2011; Schanzenbach, 2014). Evidence from quantitative and qualitative studies suggests that reduction in class size is particularly effective when teachers’ uses smaller classes to strengthen the quality of the interactions with their students and ensure students’ mastery of content and skills (Graue, Rauscher, & Sherfinski, 2009; Schanzenbach, 2014).

2.5. Standard 4: learning environment & instructional practices

The VSQI uses the Early Childhood Environmental Rating Scale-Revised (ECERS-R; Harms et al., 2005) to assess the learning environment in the center. The ECERS-R measures the structural components of the classroom, such as the physical environment and basic care of children, as well as more process-oriented components, such as the interactions among staff, children, and parents. Burchinal, Kainz and Cai (2011) examined the relations between the ECERS-R and outcomes for low-income children in several large datasets conducted in the late 1990s to early 2000s (Cost Quality and Outcomes Study; National Center for Early Development and Learning (NCEDL) Multi-State Study of Pre-Kindergarten). They found that the ECERS-R was significantly related to children’s outcomes, but these associations were modest. In contrast, evidence from a nationally representative longitudinal study, the Early Childhood Longitudinal Study (ECLS-B), suggests that the ECERS-R is not a consistent predictor of child outcomes in the current child care landscape (Sabol & Pianta, 2014). In addition, the quality of teacher-child interactions may be a stronger predictor of children’s learning than the overall learning environment measured by the ECERS-R (Mashburn et al., 2008).

3. Using quality standards in a QRIS context

Although there is a general consensus that certain quality standards relate to child outcomes, albeit often modestly, there is little understanding of whether there are certain points in which the magnitude of associations are greater, or points in which associations begin to level off either at the low or high end of the quality distribution (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). This lack of evidence is concerning given that QRIS apply cut-points to individual quality standards to determine quality ratings. Most quality standards were originally developed to identify the features of care that best contribute to children’s development rather than for potentially high-stakes policy contexts, and as such, offer little empirical regarding where to determine cut-points (Halle, Vick Whittaker, & Anderson, 2010).

Consequently, states have had to rely on their own methods or preferences for designing the rating structure of the QRIS. As an example, 20 out of the 26 statewide QRIS use the ECERS-R as a quality standard, yet the cut-points that states use to delineate top-rated programs varies greatly from state to state. The ECERS-R is based on a 1–7 scale, with developers identifying 1 as inadequate quality, 5 as good quality, and 7 as excellent quality. In order to receive the top rating in the QRIS, a program must receive at least: 4.0 in the District of Columbia, 5.1 in Mississippi, 5.5 in Florida, and 6 in Kentucky, with varying cut-off points and minimum required scores across the subscales (Tout et al., 2010). Although individual states may have varying levels of quality which may explain the different location of cut-points, the question remains whether there are implications of these cut-points within a given state regarding the links to children’s learning and development.

Extant evidence suggests that individual quality indicators vary in terms of magnitude of associations with child outcomes. As a result, it is unclear whether cutting, weighting, and averaging these quality standards will relate to child outcomes in a meaningful way. In the VSQI, the rating structure relies on cut-points based on developers’ identified ranges of quality or alternatively, seemingly arbitrary demarcations. The extent to which the VSQI rating structure reflects differential associations with outcomes remains an important empirical question, not only for Virginia, but for other states as they seek to design and structure their systems.

4. A focus on pre-kindergarten programs

In the current study, we only focus on children who attended Virginia’s state-funded targeted pre-kindergarten program, the Virginia Preschool Initiative. Virginia’s targeted pre-kindergarten program provides education to at-risk children who are not served by another program (e.g., Head Start). A child is eligible to participate in the pre-kindergarten if they have at least one demographic risk, including if the child lives in poverty, is homeless, comes from a family with low levels of education, or is an English Language Learner (Huang, Invernizzi & Drake, 2012). Given the current national focus on expanding state pre-kindergarten programs, and explicit emphasis on quality, results may have implications for the validity of QRIS within pre-kindergarten programs, particularly for at-risk children.

Centers in Virginia’s pre-kindergarten program have to adhere to several benchmarks, such as requiring teachers to specialize in early childhood, and having class sizes that are 20 or lower. According to the State of Preschool report from 2007, which is the same year the pre-kindergarten programs were assessed in the current study, Virginia’s pre-kindergarten program served around 12% of 4-year-old children in the state, ranking in the middle of states at the time that had pre-kindergarten programs (23 out of 38; Barnett, Hustedt, Friedman, Boyd, & Ainsworth, 2008). Virginia’s preschool initiative also ranks in the middle in terms of resources allocated per child enrolled ($3577; 17 out of 38). Therefore, our findings may have implications for other states on the utility of ratings for pre-k programs to regulate beyond minimum structural requirements.

5. The current study

The rapid proliferation of QRIS has led to a need for empirical evaluation of how QRIS operate within a potentially high stakes policy context. The QRIS ratings are the key output of QRIS, yet there is very little empirical evidence on whether states’ rating structures capture unique aspects of quality that relate to children’s outcomes.
The present study validates Virginia’s pilot QRIS among children who attended state-funded, targeted pre-kindergarten programs. We capitalize on data collected by the state regarding children’s literacy performance, and investigate the extent to which the rating structure in the Virginia Star Quality Initiative relates to children’s growth in literacy across preschool and kindergarten.

6. Method

6.1. Sample description

Analyses draw upon a database constructed from the following datasets: (1) child care facility information from the Virginia Department of Social Services; (2) block group census data from the 2000 U.S. Census; (3) quality standard scores and star ratings on the VSQI from 2007 to 2009; and (4) pre-kindergarten and kindergarten performance on the Phonological Awareness Literacy Screening (PALS), and center and child characteristics from the Virginia Department of Education and the University of Virginia PALS Office from 2007 to 2010. Each of the datasets had a different constraint on the sample which we describe below.

The first dataset included information on 6472 child care facilities in Virginia that were in operation from 2007 to 2009, which was provided by the Virginia Department of Social Services. This file included names, addresses, facility type (e.g., Head Start or religious exempt program), and licensing status for all child care programs at the time. Using GIS mapping software, we geocoded all child care facilities. We then determined whether the program was eligible to apply for the VSQI based on (1) program type (i.e., licensed and/or center-based program); and (2) geographic boundaries. Fortyt-five percent (n = 2928) of the child care facilities in Virginia were ineligible to apply to the VSQI because there were not center-based or licensed. These included home-based care (n = 1802; 24%), unlicensed programs (n = 1041; 16%), and after school programs (n = 85; 1%). There were 3544 center-based programs that were eligible to apply to the VSQI including licensed child care programs, and/or pre-kindergarten, Head Start, religious exempt, or military-based programs.

Programs were ineligible to apply to the VSQI if they were not located in communities that elected to participate. From 2007 to 2009, 14 communities voluntarily participated in the VSQI: Arlington/Alexandria, Central Virginia, Charlottesville/Albemarle, Fairfax, Greater Richmond/Petersburg, Newport News, Prince William County, Rappahannock County, Roanoke, Shenandoah Valley, South Hampton Roads, Western Tidewater, Williamsburg, and New River Valley. We used defined boundaries set by the state to distinguish among programs that were eligible to apply to the VSQI. Of the 3544 center-based programs, 1891 were located in 14 communities that volunteered to participate in the VSQI and thus eligible to participate.

The second data set is from the U.S. Census in 2000. We use census data from 2000 because those were the data available at the time of analysis. We linked to census data to each geocoded child care facility at the block group level. A block group is the smallest geographic area for which census data are available and thus provides the best representation of the neighborhood context in which the child care center is located. We aggregated block census data within a two-mile radius of the center in order to capture the heterogeneity within a neighborhood. These data provided key covariates for all of our analyses.

The third dataset was provided by the Virginia Early Childhood Foundation, and the Virginia Office of Early Childhood Development, and included center-based program ratings in Year 1 (2007–2008) and Year 2 (2008–2009) of the VSQI, which included 237 centers. Thus, only 13% (n = 237 out of 1891) of the centers in Virginia that were eligible to participate in the QRIS during the first two years participated. The relatively low penetration rate may be due to the fact that Year 1 was the pilot year of the VSQI and was purposefully rolled out with a small number of centers. Moreover, because the VSQI was voluntary and new, centers may not have been aware of the program and thus did not apply.

Our outcome data, provided in the fourth dataset, come from a state-wide individually administered assessment of early literacy, the Phonological Awareness Literacy Screening (PALS). The PALS is intended to help teachers and schools identify children requiring early literacy intervention (Invernizzi, Sullivan, Meier, & Swank, 2004). Although the PALS is a diagnostic tool, the PALS has been used in several empirical studies as an outcome as in the case for this study (Huang et al., 2012; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). All preschool programs and elementary schools receiving at least partial funding from the state were required by the Early Intervention Reading Initiative (mandated by the 1997 Virginia Acts of Assembly) to administer a diagnostic assessment of student’s reading skills (Invernizzi, Justice, Landrum, & Booker, 2004). As a result, the PALS was used by over 99% of Virginia’s state-funded pre-kindergarten programs, as well as public elementary schools, through a contract with the University of Virginia (Huang et al., 2012).

Among the 237 centers that participated in Year 1 and 2 of the VSQI, 71 were pre-kindergarten programs. We matched the 71 VSQI-rated programs to PALS data using district, school, and program names. All pre-kindergarten programs that were in the VSQI had PALS data. Independent sample t-tests indicated that VSQI pre-kindergarten programs (n = 71) had similar characteristics to non-VSQI pre-k programs eligible to apply to the VSQI (n = 135). For example, children’s literacy performance in the fall of the preschool year (baseline) was nearly identical (M = 33.79, SD = 6.65 versus M = 33.86, SD = 7.42) between the two groups. The two groups were also similar in terms of percentage of children who received Title I funding, received special education, were English language learners, and were Latino. VSQI pre-k programs did have significantly fewer children enrolled (M = 41.58, SD = 63.11) compared to non-VSQI pre-k programs (M = 104.04 children, SD = 108.12).

Of the 2805 children in the 71 pre-k programs with PALS data, 2448 were enrolled in the center during the fall and spring of pre-kindergarten. We did not include the 357 children who were not in the center during both time points due to uncertainty of children’s exposure to the center. Based on demographic characteristics, there were slightly more African American children in the analytic sample (40%; n = 2448) than the full sample (30%; n = 2805; χ2 = 6.53, p < .05). All other characteristics (e.g., Hispanic, disability, or gender) were relatively similar.

6.2. Participating communities, pre-kindergarten programs, and children

Our analytic sample included 2448 children in 71 state-funded pre-kindergarten programs in 14 communities. The average income of residents in the 14 communities was approximately $59,000 with 10% of families living below poverty.1 All 71 centers in our sample were in Virginia’s targeted state-funded pre-kindergarten program. Among the 71 pre-k programs in the

---

1 Communities in the VSQI do vary from one another. For example, there is a relatively large disparity in the average income across communities, ranging from approximately $43,000 in Williamsburg to $99,000 in Fairfax. Independent sample t-tests indicated no significant differences between the average characteristics of communities (e.g., income, race, education level of residents) in the VSQI compared to communities not in the VSQI. For instance, 10% of the residents in the VSQI communities lived below poverty compared to 9% in the rest of Virginia.
Table 1
Descriptive statistics of center characteristics by pre-k program star rating.

<table>
<thead>
<tr>
<th>Analytic sample</th>
<th>By Virginia QRIS star rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Star 2 (SD)</td>
</tr>
<tr>
<td>M or % (n = 71)</td>
<td>M or % (n = 8 (11%))</td>
</tr>
</tbody>
</table>

QRIS
1: Staff Qual.
- 23.48 (6.41)
- 13.67 (3.85)
- 24.06 (5.48)
- 25.57 (3.92)

2: Interactions
- 42.93 (9.72)
- 32.50 (6.91)
- 39.53 (7.50)
- 51.04 (7.21)

3: Group Size
- 22.55 (7.39)
- 16.25 (5.18)
- 22.10 (6.55)
- 24.83 (5.82)

4: Environment
- 23.32 (7.39)
- 13.25 (5.75)
- 21.13 (4.74)
- 28.40 (5.11)

Center characteristics
- Black (%): 32.77 (2.71)
- Hispanic (%): 15.06 (5.02)
- LEAP (%): 8.16 (3.93)
- Avg. FP Sum: 33.86 (7.42)
- Title I (%): 4.41 (5.11)
- Disability (%): 9.69 (7.58)
- Enrollment: 41.58 (63.11)

Neighborhood
- Household income: 54,909.91 (16,557.20)
- Below poverty (%): 11.04 (7114.25)
- Black (%): 23.85 (16,610.29)
- Hispanic (%): 4.45 (14,493.42)
- High school (%): 49.67 (11.09)
- Some college (%): 24.57 (11.09)
- Rural/No farm (%): 31.32 (12.78)
- Rural/Farm (%): 1.42 (0.54)

Note: Avg. FP Sum = The average performance of children in a center at fall of pre-kindergarten.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Significantly different from 2-star pre-k programs, p &lt; .05.</td>
</tr>
<tr>
<td>b</td>
<td>Significantly different from 3-star pre-k programs, p &lt; .05.</td>
</tr>
<tr>
<td>c</td>
<td>Significantly different from 4-star pre-k programs, p &lt; .05.</td>
</tr>
</tbody>
</table>

Table 1 presents center-level descriptive information on the full analytic sample. On average, centers had approximately 41.58 children enrolled with a fair amount of variation (SD = 63.11). Centers were composed of almost equal percentages of White (38%) and African American children (33%) followed by Hispanic children (15%). In terms of the neighborhood characteristics of areas surrounding the centers, pre-k programs were predominately located in urban neighborhoods (70%), with 25% of the residents having a Bachelor’s degree or more, and relatively high percentages of White residents (67%), followed by African American (24%) and Hispanic (4%) residents. The average income of neighborhood residents was $54,909 with 11% living below the poverty level.

Table 2 presents descriptive information on the 2448 children in our sample. Children in the sample were on average 57 months of age in the fall of the preschool year (baseline). Fifty-four percent of the children were non-white, 50% were boys, 7% had disabilities at pre-kindergarten entry, 8% had limited English Proficiency (LEP) and 12% of children received Title-I funding.

7. Measures

7.1. Star quality ratings

The VSQI assessed center quality on four performance standards: (1) staff education and qualifications; (2) teacher-child interactions; (3) structure (i.e., staff-to-child ratio); and (4) environment and instruction. Star Quality Raters assessed classrooms through documentation and observations. Star Quality Raters were extensively trained and were tested for inter-rater reliability once every seven visits. In 2007–08 (VSQ Year 1), Star Raters observed one out of every three classrooms. In 2008–09 (VSQ Year 2), the state changed this procedure and raters observed one toddler classroom, one 3-year-old classroom, and one 4-year-old classroom when available.

After the Star Quality Raters assessed a classroom, the points from each performance standard were summed and weighted. Standard 1 (Staff Qualifications) and Standard 4 (Environment) was out of 40 possible points, and Standard 3 (staff-to-child ratio and group size) was out of 30 points. Because the stakeholders were particularly concerned with the quality of teacher-child interactions, the VSQI weights Standard 2 (Interactions) more than the other three standards (60 points). The total score is out of 170 possible points. The total scores were converted to a star rating based on the following cut-points: Star 1, 34–50 points; Star 2, 51–84; Star 3, 85–118; Star 4, 119–152; and Star 5, 153–170, with 1 indicating low quality and 5 indicating superior quality (Kirby, Boilier, & Tout, 2010). See Appendix Table A1 for a full list of all indicators and associated points (a complete matrix of how points are awarded within each indicator is available upon request).

7.1.1. Standard 1: Education, qualifications, and training (40 points)

Programs provided documentation related to education, qualifications, and training for directors, teachers, and assistant teachers to demonstrate their performance on Standard 1. Documentation included credentials, transcripts, training plans, and evidence of participation in professional development or enrollment in formal coursework. Additionally, centers provided documentation of annual clock hours of approved training for staff, including mentorship, professional conferences and college coursework, and years of experience. Outside evaluators examined the documentation and assigned a total point score based on eight indicators (5 points each): (a) director, teacher, and assistant teacher qualifications; and (b) all staff, director, all instructional staff, teacher, and assistant teacher professional development activities. Each indicator has its own set of requirements to determine points awarded. For example, in order to receive 5 points for director qualifications, the director must have at least a Master’s degree and a minimum of 2 years of experience with 1 year in a supervisory capacity in a child-related environment. All eight indicators were then summed for a total of 40 points.

7.1.2. Standard 2: Interactions (60 points)

The pre-kindergarten Classroom Assessment Scoring System (CLASS) was used to measure the interactions between teachers and children (Pianta, Mashburn, et al., 2008). CLASS organizes teacher-child interactions into three broad domains: (1) Emotional Support, (2) Classroom Organization, and (3) Instructional Support. Star Quality Raters then assign a global rating score from 1 to 7 to each domain based on developers’ recommendations for different quality levels, with 1–2 indicating a low classroom score, 3–5 indicating that the classroom is in the mid-range, and 6–7 indicating the classroom is high on that domain. The dimensions are averaged within each domain to create a domain-level composite score. Each domain score corresponds to a specific star rating. For example,
Table 2
Descriptive statistics of child characteristics by pre-k program star rating.

<table>
<thead>
<tr>
<th></th>
<th>Analytic sample</th>
<th>By Virginia QRIS rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 2448</td>
<td>n = 168</td>
</tr>
<tr>
<td></td>
<td>M or % (SD)</td>
<td>M or % (SD)</td>
</tr>
<tr>
<td><strong>Child</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-k age (in months)</td>
<td>57.34 (3.85)</td>
<td>56.93c (4.30)</td>
</tr>
<tr>
<td>Boy (%)</td>
<td>50.61</td>
<td>53.57</td>
</tr>
<tr>
<td>Black (%)</td>
<td>39.99</td>
<td>66.67d,e (40.73)</td>
</tr>
<tr>
<td>Hispanic (%)</td>
<td>13.57</td>
<td>4.17b,c</td>
</tr>
<tr>
<td>LEP (%)</td>
<td>8.21</td>
<td>1.78c</td>
</tr>
<tr>
<td>Pre-k disability (%)</td>
<td>7.72</td>
<td>7.73</td>
</tr>
<tr>
<td>Title I Funds (%)</td>
<td>12.37</td>
<td>16.80ae</td>
</tr>
<tr>
<td><strong>Fall Pre-K (FP)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP alphabet knowledge</td>
<td>16.04</td>
<td>19.80a,b,e</td>
</tr>
<tr>
<td>FP phonological awareness</td>
<td>9.84</td>
<td>10.62c</td>
</tr>
<tr>
<td>(5.69)</td>
<td>(4.70)</td>
<td>(5.71)</td>
</tr>
<tr>
<td><strong>Days between assessments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall and spring pre-k</td>
<td>206.29</td>
<td>194.47</td>
</tr>
<tr>
<td>(11.96)</td>
<td>(12.08)</td>
<td>(11.97)</td>
</tr>
<tr>
<td>Spring pre-k and fall kindergarten</td>
<td>157.72</td>
<td>163.72</td>
</tr>
<tr>
<td>(10.66)</td>
<td>(11.09)</td>
<td>(9.93)</td>
</tr>
<tr>
<td>Fall and spring kindergarten</td>
<td>209.90</td>
<td>205.64</td>
</tr>
<tr>
<td>(11.97)</td>
<td>(11.96)</td>
<td>(10.74)</td>
</tr>
</tbody>
</table>

**Note:** Fall and spring pre-k = Number of days between fall and spring PALS-PreK literacy assessment.

* Significantly different from 2-star pre-k programs, p < .05.
* Significantly different from 3-star pre-k programs, p < .05.
* Significantly different from 4-star pre-k programs, p < .05.

A CLASS score of six or higher on all three domains – Emotional Support, Classroom Organization, and Instructional Support – is awarded a 5-star rating for each domain, whereas a CLASS score of two or lower across all three domains is awarded a 1-star rating. The star ratings for each of the three domains are multiplied by four to obtain a total possible score of 60 points.

7.1.3. Standard 3: staff-to-child ratio and group size (30 points)

Staff-to-child ratio and group size were assessed through documentation and observation from the facility’s staffing plan. Staff-to-child ratio and group size requirements for each star level are differentiated between classrooms serving 2-, 3-, and 4-year-old children. At the high end of the scale, a classroom with no more than a 1:7 staff-child ratio for 3-year-old children and 1:8 for 4-year-old children is awarded a 5-star rating. At the low end, a classroom with greater than 1:9 staff-child ratio for 3-year-old children and 1:10 for 4-year-old children is awarded a 1-star rating. The star rating for staff-to-child ratio is multiplied by four (20 possible points) and the rating for group size is multiplied by two (10 possible points) for a total of 30 possible points.

7.1.4. Standard 4: learning environment & instructional practices (40 points)

Star Quality Raters assessed the overall learning environment using the Early Childhood Environment Rating Scale-Revised Edition (ECERS-R; Harms et al., 2005) for preschool-aged classrooms. The ECERS-R included 43 items that measured the quality and safety of the physical environment, as well as the quality of interactions in the classroom. Each item was rated on a 7-point scale and averaged to create the ECERS-R total score, ranging from 1 (low quality) to 7 (excellent quality). In the VSQI, individual classroom ECERS-R scores were averaged within a center, with each classroom needing a minimum score in order to receive a particular rating. For instance, in order to receive a 5-star rating, the average ECERS-R scores must be at or above a six, with no classroom scoring below a five. All observers achieved 80 percent reliability within one point before conducting field observations.

Additionally, Standard 4 included ratings on programs’ practices in preparing children to transition to kindergarten or to another setting (e.g., special education context). In order to receive a 5-star rating on this indicator, programs must document that they have conducted transition-related training for all teaching staff involved in child’s transition and individual meetings with families to share specific information regarding the child’s transition to another setting. The rating for transition practices is weighted by two (10 possible points) and the ECERS-R is weighted by six (30 possible points) for a total of 40 points.

7.2. Child outcomes

7.2.1. Children’s pre-literacy skills

Children’s pre-literacy skills were assessed by teachers at four time points: fall and spring of pre-kindergarten using the Phonological Awareness Literacy Screening pre-kindergarten (PALS-PreK; Invernizzi, Sullivan, et al., 2004) and fall and spring of kindergarten using the Phonological Awareness Literacy Screening-Kindergarten (PALS-K; Invernizzi, Juel, Swank, & Meier, 2013). The PALS-PreK and PALS-K are criterion-referenced assessments that take teachers approximately 20–25 min to administer with no formal training. Items and tasks in the PALS-PreK and PALS-K were derived and modified based on analyses of scores of more than 500,000 students from 1997 to 2004, demonstrating sound concurrent and predictive validity. For example, children’s letter knowledge and phonological skills measured by the PALS-K had a positive and significant correlation with another widely used measure of literacy skills, the Stanford Achievement Test, measured at the same time point (r = 0.79) and later in elementary school (see technical manual for further details on validity for PALS-PreK and PALS-K; Invernizzi, Sullivan, et al., 2004; Invernizzi et al., 2013).
The PALS-PreK includes seven subscales: letter recognition, letter sound awareness, rhyme awareness, beginning sound awareness, print and word awareness, name writing, and nursery rhyme awareness. The seven subscales of the PALS-K include letter recognition, letter sound awareness, rhyme awareness, spelling, concept of a word, and word recognition in isolation. We selected four subscales (76 items) that were consistently measured across all four time points allowing for longitudinal analysis: letter recognition and letter sound awareness (Alphabet Knowledge), and beginning sounds and rhyming skills (Phonological Awareness).

For Alphabet Knowledge, letter recognition was assessed by asking children to name each letter in the alphabet. Inter-rater reliabilities for the 26 items have been consistently high ($r = .99$, $p < .01$), and have high test-retest reliability is (.92). For letter sound awareness, children were presented with each of the letters and asked to tell the sound that the letter makes. Inter-rater reliabilities ($r = .98; p < .01$) and test-retest reliability (.88) are also high for the 26 letter sounds. We summed letter recognition and letter sound awareness to create a composite score for Alphabet Knowledge (range = 0–52).

For Phonological Awareness, beginning sound awareness examined children’s ability to identify two words that share a common initial phoneme from a series of four words represented by pictures. The subscale is out of 10 items with higher scores indicating better sound awareness. The sound awareness subscale demonstrates strong psychometric properties (inter-rater reliability = .99, $p < .01$; test-retest reliability = .78). Rhyme awareness assessed children’s ability to identify a rhyming match for a target word. There are 10 items in the subscale, with higher scores indicating stronger rhyming skills (inter-rater reliabilities ranged from .96 to .99, $p < .01$; and test-retest reliability = .81; for a summary see Invernizzi, Sullivan, et al., 2004). We summed beginning sound awareness and rhyme awareness to generate the composite for Phonological Awareness (range = 0–20) in order to create more variance and potentially mitigate the ceiling effect of the composites. Cronbach’s alpha ranged from .69 to .87 for the two composites across the four time points.

7.2.2. Community and neighborhood characteristics

The community and neighborhood characteristics from the U.S. Census were as follows: (1) median household income; (2) percent of households living below the poverty level; (3) percent of African American residents; (4) percent Hispanic residents; (5) percent of residents who are another race; (6) percentage of residents with a Bachelor’s degree or higher (% B.A. or higher); (7) percent of single-mother households; and (8) percent of the population that live in a rural area (farm or nonfarm). We also calculated the number of centers within a two-mile radius for each center, as well as the proportion of programs participating in the VSQI relative to non-participating programs. We included these demographics as covariates to account for parental choice into pre-kindergarten programs, and the potential correlation with children’s learning.

7.2.3. Additional covariates

An additional set of center-level and child-level covariates were included to account for factors that may correlate with child outcomes, and potentially correlate with selection into the child care program. Child characteristics included age in months at the time of testing, gender, race/ethnicity, limited English proficiency (LEP), fall pre-kindergarten disability status, and whether the child received Title-I Funds (as a proxy for income). Because there was a slight range in the intervals between assessments, we controlled for the months between each assessment (e.g., months between the fall and spring pre-kindergarten assessment). We also controlled for the number of children in kindergarten that were in the same pre-kindergarten to account for potential peer influences. Additionally, we created center-level averages of the child measures described above, including the proportion of African American students, proportion of Hispanic children, proportion of children with disabilities, enrollment size, proportion of children who were LEP, the average performance on the PALS in Fall pre-kindergarten, and proportion of children who qualified for Title I.

8. Analytic plan

A central tenet to developmental theory is that individual children’s development is strikingly heterogeneous across time. Examining the association between child care quality and children’s performance using two time points (e.g., predicting fall kindergarten performance while controlling for fall pre-kindergarten performance) does not account for the variability in growth within individual children. Indeed, there is substantial evidence that accounting for within-child variability yields a more accurate portrayal of children’s functioning (Singer & Willett, 2003; Votruba-Drzal, Li-Grining, & Maldonado-Carreno, 2008). In addition, the influence of child care quality is often nonlinear, with stronger gains in preschool than in later years (Puma et al., 2012).

In order to account for individual-specific growth trajectories, we employ hierarchical linear modeling techniques, which combine longitudinal and multilevel features to adjust for variation between and within children and schools. To account for within-child variability in trajectories, we estimate time in two different ways. First, we treat time continuously, making the assumption that growth across pre-kindergarten and kindergarten is linear. In the second set of models, we estimate a piecewise growth model in order to account for the potential non-linear growth across early childhood and to estimate separate growth curves at each time point (Chou, Yang, Pentz, & Hser, 2004).

For both sets of three-level models, Level 1 included children’s literacy scores, which were nested within individual children at Level 2, who were nested within pre-kindergarten programs at Level 3. Therefore, Level 1 represents time (i.e., the four repeated assessments of the constructs of interest for each child), Level 2 represents the variation across children in growth parameters (i.e., initial status and growth rate) within a center, and Level 3 represents variation in initial status and growth rate between programs. Unconditional models suggested that 38%–62% of the total variance was attributed to mean differences between children (child-level ICC = .38 for Alphabet Knowledge and 0.62 for Phonological Awareness), and 12–14% was attributable to between center variation (center-level ICC = 0.14 for Alphabet Knowledge and 0.12 for Phonological Awareness).

To model growth, we used two composites from the PALS-PreK and PALS-K, Alphabet Knowledge and Phonological Awareness, which have consistent items across fall and spring pre-kindergarten, and fall and spring kindergarten. On average, there were eight months between the fall and spring of pre-kindergarten assessments (Time 1), six months between the spring of pre-kindergarten and the fall of kindergarten (Time 2), and eight months between the fall and spring of kindergarten (Time 3). We use these three time points as the cut-points for three-level piecewise models.

Conditional models were estimated to explain the heterogeneity in literacy skill growth. For parsimony, we discuss the models and equations only for the piecewise regressions below. Eq. (1) depicts the Level-1 model for the piecewise linear model, which specifies that a literacy score $Y$ at time $t$ for a child $i$ in center $j$ is a function of literacy scores of child $i$ in center $j$ at the fall of pre-kindergarten ($x_{0j}$; when $t = 0$), the growth rate from the fall of pre-kindergarten to spring pre-kindergarten ($x_{1j}$), the child’s growth rate from the spring of pre-kindergarten to the fall of
kindergarten ($\pi_{2ij}$), the growth rate from fall kindergarten to spring kindergarten ($\pi_{3ij}$) and the residual associated with an individual child’s score at a specific time point ($\zeta_{ij}$).

$$Y_{ij} = \tau_{0ij} + \pi_{1ij}(\text{time1})_{ij} + \pi_{2ij}(\text{time2})_{ij} + \pi_{3ij}(\text{time3})_{ij} + \zeta_{ij} \quad (1)$$

Eq. (2) was used at Level 2 to model variability in the Level 1 parameters. In this level, the initial level (i.e., slope if $p = 1, 2,$ or 3 intercept if $p = 0$) of children’s literacy performance was modeled as a function of an intercept ($\beta_{0p0}$), vector of time-invariant child-level covariates ($\beta_{p0j}$) the varying time between assessment points ($\beta_{p2j}$), and estimation error between children within children ($\zeta_{pj}$). The three slope parameters ($p = 1, 2, 3$) represented growth between the four time points, which were estimated with the same set of covariates as the intercept.

$$\pi_{pj} = \beta_{0pj} + \beta_{p1j}(\text{child})_{ij} + \beta_{p2j}(\text{assess})_{ij} + \zeta_{pj} \quad p = 0, 1, 2, 3 \quad (2)$$

In Level 3, depicted in Eq. (3), we included dummy codes for VSQI rating, indicating whether the center received a 3-star ($\gamma_{p01}$) or 4-star quality rating ($\gamma_{p02}$; omitted group = 2-star rating) to predict initial levels in Level 2. Level 3 also included a host of center characteristics ($\gamma_{p03}$), neighborhood characteristics for each center ($\gamma_{p04}$), dummy codes for each community (i.e., community fixed effects; $\gamma_{p05}$) and a dummy code for the year that center’s VSQI rating occurred ($\gamma_{p06}$).

$$\beta_{pj} = \gamma_{p00} + \gamma_{p01}(3-\text{star})_{j} + \gamma_{p02}(4-\text{star})_{j} + \gamma_{p03}(\text{center})_{j} + \gamma_{p04}(\text{neighborhood})_{j} + \gamma_{p05}(\text{community})_{j} + \gamma_{p06}(\text{year})_{j} + \varphi_{p0j} \quad p = 0, 1, 2, 3 \quad (3)$$

Our strategy for estimating the relation between programs’ ratings and children’s growth while mitigating selection bias included four main approaches: (1) controlling for initial performance; (2) including community and year fixed effects; (3) employing a set of child and center controls; and (4) accounting for potential ceiling effects of the outcome measures using a tobit regression. We employ combinations of these approaches across four models, which are displayed in Table 3 for Alphabet Knowledge and Table 4 for Phonological Awareness.

In order to calculate effect sizes for the three-level growth models, we used procedures adapted from Snijders and Bosker (1999) and employed by Jones, Brown, and Aber (2011) by dividing the rating estimate by an estimated standard deviation of the outcome. The outcome standard deviation was calculated from a baseline unconditional three-level model as the square-root of the total variance.

### Table 3

Unstandardized estimates of association between pre-k program star rating and growth in alphabet knowledge from fall pre-kindergarten to spring kindergarten, 3-level piecewise model ($n = 2448$).

<table>
<thead>
<tr>
<th>Start of FP</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE)</td>
<td>$-0.40$ ($2.25$)</td>
<td>$-0.76$ ($2.82$)</td>
<td>$-1.28$ ($1.64$)</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE)</td>
<td>$-0.81$ ($2.57$)</td>
<td>$-1.45$ ($2.19$)</td>
<td>$-2.37$ ($1.54$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FP to SP</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE)</td>
<td>$0.67$</td>
<td>$0.12$</td>
<td>$0.35$</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE)</td>
<td>$1.34$ ($3.56$)</td>
<td>$1.94$ ($2.05$)</td>
<td>$2.83$ ($1.66$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SP to FK</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE)</td>
<td>$-0.81$ ($3.04$)</td>
<td>$-1.50$ ($1.38$)</td>
<td>$-1.74$ ($1.66$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FK to SK</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE)</td>
<td>$-0.15$</td>
<td>$-0.18$</td>
<td>$-0.11$</td>
</tr>
</tbody>
</table>

**Note:** Child-level controls = child’s gender, race, age, and time between assessments. Center-level controls = racial and ethnic make-up of center, average student performance in fall of preschool, percentage of children who are English Language learners, percentage of children who receive Title 1 funding, center enrollment, and percentage of children who have disabilities, as well as neighborhood characteristics including average household income, percentage of households living below the poverty line, education level of head of household, percentage of households led by single mothers, and urbanicity.

* The omitted group in all models were children in 2-star pre-k programs. The comparisons among children in 3-star and 4-star pre-k programs are based on post-estimation tests of significance.

* $p < .05$.

** $p < .01$.

*** $p < .001$. 

---

### Table 4

Unstandardized estimates of association between pre-k program star rating and growth in phonological awareness from fall pre-kindergarten to spring kindergarten, 3-level piecewise model (n = 2448).

<table>
<thead>
<tr>
<th>Start of FP</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE) = -1.45(0.85)</td>
<td>-1.18 (0.84)</td>
<td>-1.08 (0.54)</td>
<td>-1.44* (0.56)</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE) = -1.27</td>
<td>-0.22</td>
<td>-0.20</td>
<td>-0.27</td>
</tr>
<tr>
<td>3-Star to 4-star</td>
<td>B(SE) = -1.33(0.89)</td>
<td>-0.87(0.88)</td>
<td>-0.32(0.56)</td>
<td>-0.32(0.58)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FP to SP</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE) = 2.92(1.21)</td>
<td>2.63 (1.20)</td>
<td>2.00 (0.85)</td>
<td>2.84* (1.09)</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE) = 2.84(1.26)</td>
<td>2.37 (1.25)</td>
<td>1.15(0.89)</td>
<td>1.20(1.08)</td>
</tr>
<tr>
<td>3-Star to 4-star</td>
<td>B(SE) = -0.07(0.78)</td>
<td>-0.26(0.76)</td>
<td>-0.90(0.51)</td>
<td>-1.52* (0.64)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SP to FK</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE) = 0.27(0.38)</td>
<td>0.26(0.38)</td>
<td>0.19(0.37)</td>
<td>0.01(0.68)</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE) = -0.17(0.41)</td>
<td>-0.17(0.42)</td>
<td>-0.09(0.40)</td>
<td>-0.22(0.73)</td>
</tr>
<tr>
<td>3-Star to 4-star</td>
<td>B(SE) = -0.51(0.91)</td>
<td>-0.69(0.9)</td>
<td>-1.19 (0.58)</td>
<td>-1.75 (0.65)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FK to SK</th>
<th>Model 1 Fall pre-K</th>
<th>Model 2 Fixed effects</th>
<th>Model 3 Center &amp; child</th>
<th>Model 4 Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Star to 3-star</td>
<td>B(SE) = -0.25(0.43)</td>
<td>-0.26(0.43)</td>
<td>-0.37(0.39)</td>
<td>-0.24(0.74)</td>
</tr>
<tr>
<td>2-Star to 4-star</td>
<td>B(SE) = -0.14(0.43)</td>
<td>-0.14(0.45)</td>
<td>-0.63(0.42)</td>
<td>-0.45(0.81)</td>
</tr>
<tr>
<td>3-Star to 4-star</td>
<td>B(SE) = -0.39(1.1)</td>
<td>-0.57(1.08)</td>
<td>-1.45 (0.68)</td>
<td>-1.96 (0.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specification</th>
<th>OLS regression</th>
<th>Fall pre-k covariate</th>
<th>Community and year fixed effects</th>
<th>Child and center-level controls</th>
<th>Tobit regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: Child-level controls - child’s gender, race, age, and time between assessments. Center-level controls - racial and ethnic make-up of center, average student performance in fall of preschool, percentage of children who are English Language learners, percentage of children who receive Title 1 funding, center enrollment, and percentage of children who have disabilities, as well as neighborhood characteristics including average household income, percentage of families living below the poverty line, education level of head of household, percentage of households led by single mothers, and urbanicity.

* The omitted group in all models was children in 2-star pre-k programs. The comparisons among children in 3-star and 4-star pre-k programs are based on post-estimation tests of significance.

8.1.1. Selection issues

The difficulty of validating QRIS ratings lies in adequately accounting for selection into the QRIS. In Virginia, there are three potential threats to validity. First, VSQI requires communities to volunteer in order to participate, thereby limiting participation to early childhood education programs within the voluntary communities, and potentially limiting generalizability and internal validity. For example, if some communities provide more resources to children than others, such as after school programs, then it may be these resources that are driving the effects, resulting in, more than likely and upward estimate of the effect of quality. The second level of bias derives from early childhood programs’ selection into the QRIS. Programs that volunteer may believe that they will have higher ratings than programs that choose not to volunteer. Selection of programs is of particular concern because an evaluation of any component of a voluntary QRIS can only ensure validity if participants reflect the relevant market population (Cautier & van der Klaauw, 2012).

Third, there is the concern of parental selection into certain programs, a potential source of bias for estimates of quality. Children are enrolled in programs with certain characteristics, backgrounds, and experiences. If these experiences, characteristics, and backgrounds outside of school are related to their academic skills and their parents’ selection of preschools, then a failure to take these important contexts into account when considering relations between star ratings and outcomes may bias results. As such, the very indicators that are used to determine the ratings may be what drive parental choice, and without adequate controls, the model may estimate the relation to differential parental selection rather than quality.

The present study attempts to improve external validity and mitigate selection bias by employing a multilevel growth model that controls for previous performance. Controlling for pre-test performance differences out disparities in levels of outcomes among children who attend high-quality programs and children who attend lower-quality programs (Raudenbush, 2001). This approach will reduce bias if the associations between unobserved characteristics and children’s development are constant over time.

Results may still be biased if the models do not adequately account for (1) differences among centers that select into the VSQI and among families that select into certain types of programs; (2) differences in demographics and resources across communities, and (3) time trends. Consequently, the analyses include child- and center-level controls, and community and time fixed effects. Child-level (Level-2) covariates may further account for individual differences in children’s performance, as well as background characteristics that may explain parents’ selection of program, such as whether they receive Title 1 funding. Center-level (Level-3) controls attempt to account for characteristics that may be associated
with the quality of care, such as average student performance in fall of preschool. Level-3 models also include neighborhood characteristics that may be associated with quality of care, as well as children’s performance, such as whether the head of household had a Bachelor’s degree or higher (see Section 7 for a complete list).

Community fixed effects were included to account for between-community differences that may affect student’s literacy, including the quality of early childhood care available in the area, demographics of families, and other local community specific variables. These fixed effects account for both observed and unobserved community characteristics that may explain student achievement. Therefore, results that include fixed effects can be interpreted as the within-community variance of children’s literacy growth. Additionally, we include dummy codes for the year in which the rating took place (2007 or 2008). We only explore the relations between program ratings and child outcomes in the first year the center was rated, reducing the potential that parents are differentially selecting programs based on the center rating.

The multilevel growth model with controls approach is consistent with numerous rigorous examinations of relations between pre-kindergarten program quality and children’s functioning (Barnett, Lamy, & Jung, 2005; Gormley, Gayer, Phillips, & Dawson, 2005). Although these analytic techniques may reduce selection, the approaches may be unable to control for time-varying differences among children who attended higher-compared to lower-quality centers. For example, if parents who are more motivated send their children to high-quality programs, and this motivation increases over time, then our models may misattribute the effects of parent motivation to center quality. Thus, the present study is a non-causal, descriptive exploration of the relation between ratings and child outcomes.

8.1.2. Other considerations

Of the 2448 children in the sample, there were missing data for child outcomes in spring pre-kindergarten (8% missing; n = 196), fall kindergarten (30%; n = 734), and spring kindergarten (17%; n = 416). More traditional approaches to handling missing data, such as list-wise deletion, have been criticized for biasing estimates and reducing power to detect effects (Widaman, 2006). Therefore, we imputed missing data using the chained equations command (ICE) in Stata. The multilevel analyses were conducted for each of the five imputed data files, and coefficients and standard errors resulting from each analysis were corrected to account for the multiple imputations to provide estimates of the associations between children’s estimated developmental outcomes and star ratings. Enders (2011) highlights the validity of imputing outcomes in the context of longitudinal data given advances in missing data methodology. We also acknowledged the counterargument for not imputing outcomes and conduct the analyses without imputing outcomes (Carlin, Li, Greenwood, & Coffey, 2003). Any differences between imputed and non-imputed estimates are noted in the explanation of findings.

Additionally, our outcomes demonstrated the potential for ceiling effects. For instance, Alphabet Knowledge is based on knowing the letters and the sounds of letters in the alphabet. If two students know all 26 letters in the alphabet, they may be equal according to our scale, but might not truly be equal in aptitude. Indeed, a number of cases were potentially censored at the right end of the distribution. At the final time point, the spring of kindergarten, 37% of the children received the perfect score on the Alphabet Knowledge composite and 68% received a perfect score for Phonological Awareness. To account for the potential ceiling effect of these composites, we employed tobit regressions which are designed to estimate linear relationships when there is censoring in outcome variables. Tobit regressions keep the predictors in the model but censor the outcome when the predicted value is outside of the expected range, thereby predicting the probability of being censored and the expected value of the uncensored portion based on observed data (Breen, 1996; Kalil, Ryan, & Corey, 2012).

9. Results

9.1. Differences between center characteristics

Table 1 presents comparisons of center characteristics among 2-star, 3-star, and 4-star pre-k programs. One-way between-subjects ANOVAs indicated a number of differences between 2-star, 3-star, and 4-star pre-k programs with regard to pre-k programs’ neighborhood characteristics. Two-star programs tended to be located in areas with more African American residents and more single-mother households compared to 3-star and 4-star programs. Four-star programs, and to a lesser extent 3-star programs, were located in areas with most Hispanic residents, and more rural areas. Table 2 presents differences in the characteristics of children across 2-star, 3-star, and 4-star pre-k programs. One-way between-subjects ANOVAs indicated higher proportions of African American children in 2-star programs compared to 3-star and 4-star pre-k programs, and higher proportions of Hispanic and LEP children in 4-star programs compared to 2-star and 3-star pre-k programs. Children were also slightly older in 4-star programs and more children received Title-I funding in 3-star programs (16%) compared to children in 2-star (none) and 4-star (5%) centers.

Most surprisingly, children in the 2-star pre-k programs had higher performance on measures of literacy performance at the start of pre-kindergarten compared to 3-star and 4-star pre-k programs. This may be due to the greater proportion of children who had limited English proficiency in higher-rated programs. Overall, the differences in center and child characteristics and performance demonstrate that children enrolled in 2-star, 3-star, and 4-star pre-k programs did not have the same characteristics, posing a challenge in yielding unbiased comparisons of performance among programs based on ratings.

9.2. Predicting literacy growth in early childhood

We examined whether pre-k programs’ ratings were associated with children’s literacy growth in early childhood. We estimated two sets of conditional models of child trajectories from the fall of pre-kindergarten to the spring of kindergarten for two outcomes, Alphabet Knowledge and Phonological Awareness, with varying controls and specifications. In all of our models, the reference category was children who attended the lowest-rated VSQI programs (i.e., 2-star programs). We then estimated the difference in performance between children who attend 3-star and 4-star programs using post-estimation tests. In the first set of growth analyses, we examine associations between star ratings and children’s linear growth and in the second set, we examine associations with children’s nonlinear growth.

First, we treated time as a continuous variable, and examined the linear trajectory of children’s pre-literacy development. Across all models and specifications, there were no significant differences in growth from the fall of pre-kindergarten to the spring of kindergarten between 2-star, 3-star, and 4-star pre-k programs for either Alphabet Knowledge and Phonological Awareness, although the coefficients were all in the expected directions.

The lack of significant findings for the linear growth models is not too surprising given the shape of children’s trajectories. Fig. 1 presents the approximate growth trajectories in early childhood after allowing for the slope to vary among the three time points: fall to spring pre-kindergarten (8 months), spring pre-kindergarten...
to fall kindergarten (6 months), and fall to spring kindergarten (8 months). The figure presents children's average values at each time point after adjusting for the full set of controls. The nonlinear slopes across the three time points suggests that distinguishing among academic-year growth and summer growth for individual children may be necessary to detect associations with star quality ratings.

Given that children's trajectories appeared to be nonlinear, we estimated a series of three-level piecewise growth models to better understand whether the slopes between each time point vary as a function of star rating. We present findings for these models in Table 3 for Alphabet Knowledge and Table 4 for Phonological Awareness. The most basic specification, Model 1, adjusted for performance in the fall of pre-kindergarten. Children in 3-star and 4-star pre-k programs had significantly lower Alphabet Knowledge and Phonological Awareness skills at the start of the preschool year, starting at least half a standard deviation behind children in 2-star pre-k programs. These results suggest that there are different groups of children in lower-rated programs compared to higher-rated programs.

Results from Model 1 indicated a large effect of 3-star and 4-star pre-k programs on children's growth in literacy skills during the preschool year. Yet, we may be concerned about the validity of these findings for several reasons. First, we may be concerned that unobserved community differences are accounting for the center quality effects on children's learning. For example, if certain communities are dedicating more resources for children, such as health care services, then these resources may explain children's learning rather than the effects of quality. Failing to account for differences across communities, particularly if more children in these higher-income areas are attending higher-rated programs, may lead to potentially inflated estimates.

Therefore, Model 2 included community and year fixed effects. The fixed effects account for both unobserved and observed community-level characteristics. The year fixed effects account for time trends. Results can be interpreted as explaining within-community and within-year variances. Results in Model 2 continue to demonstrate a significant effect of quality on children's gains in literacy in the preschool year; however, the effect size is slightly smaller than in Model 1 (ES = 0.63 in Model 2 compared to 0.76 in Model 1). Additionally, though not displayed, a number of the individual coefficients for communities were significant, suggesting that communities may systematically differ on performance.

Yet, there may still be the concern that centers differentially selected into the VSQI within each community. For example, centers that are higher performing may volunteer to participate because they believe they will receive a high rating. All of the centers in our study are targeted pre-kindergarten programs, meaning they serve children who have at least one risk factor, which may reduce this concern. But it may still be the case that the highest performing centers that are serving at-risk children select into the VSQI.

In addition, there may be the concern that pre-test performance does not fully account for differential selection of parents into certain types of programs. Indeed, in Model 2, there remains a significant difference in children's literacy performance in the start of preschool among lower- and higher-rated programs. In addition, as demonstrated in Table 2, there are differences in the racial/ethnic makeup by quality level (i.e., more African American children in 2-star programs and more Hispanic children in 4-star programs).

Model 3 in Tables 3 and 4 reflect the addition of child-level and center/neighborhood-level covariates. The adjustment of center-level and child-level factors in Model 3 mitigated, and in some cases eliminated, the differences in performance at the start of pre-kindergarten. For instance, without any controls (Model 1), children in 3-star pre-k programs started almost four-tenths of a standard deviation behind 2-star pre-k programs in Alphabet Knowledge. After adjusting for all available factors (Model 3), only one significant difference remained at baseline; children in 3-star programs started around one third of a standard deviation behind in phonological awareness skills compared to 2-star programs at baseline. Thus, we were able to use our controls to attenuate the somewhat substantial differences in intercepts at the start of the pre-kindergarten year.

Overall, there appeared to be a significant difference in growth in the pre-kindergarten year among higher- and lower-rated programs, which were generally robust across multiple specifications (Models 1–4). In Model 3, our preferred model that controls for center- and child-level controls, Tables 3 and 4 demonstrates that children in 3-star pre-k programs grew over a third of a standard deviation more in pre-kindergarten in Alphabet Knowledge and Phonological Awareness compared to 2-star pre-k programs. Additionally, children in 4-star pre-k programs grew a third of standard deviation more in Alphabet Knowledge compared to 2-star pre-k programs. Children in 4-star pre-k programs grew slightly more in Phonological Awareness in pre-kindergarten compared to 2-star pre-k programs, yet the relation was no longer significant. There was no significant difference in children's literacy growth in pre-kindergarten among children who attended 3-star compared to 4-star pre-k programs.
Surprisingly, after adjusting for center-level factors, children in 3-star and 4-star programs had a somewhat steeper decline in alphabet knowledge skills over the summer between the end of pre-kindergarten and the start of kindergarten compared to 2-star programs. The addition of controls also magnified the differences in slopes between 3-star programs and 4-star programs for Phonological Awareness. Even in previous, less stringent models, the effect sizes demonstrated similar trends, albeit with non-significant effects.

9.2.1. Robustness checks

We also ran a series of robustness checks. In the final model, Model 4 in Tables 3 and 4, we included all possible controls (as in Model 3), but also attempted to mitigate the potential ceiling effects by employing a tobit regression. The terms from the tobit regression closely mirror the terms from the full model (Model 3), with modestly larger effect sizes, suggesting that our findings are robust even when we account for potential censoring.

Additionally, findings were tested on non-imputed data, which resulted in 1656 children in 56 pre-k programs who had data at all four time points. Findings were in the same direction as imputed data, but estimated fewer significant results. For instance, in the pre-kindergarten year, children in 3-star programs had half of a standard deviation ($ES = .50$) higher in literacy growth compared to children in 2-star pre-k programs after controlling for all possible factors. Although this effect size was larger than in the imputed data set, it was not significant, most likely due to the lack of power to detect effects.

Lastly, because 4-star pre-k programs had more children who were English language learners, we were concerned that the findings may have been due to differences in children’s language skills, rather than differences in the relation between quality and outcomes. As such, we omitted children who were Hispanic or limited English proficient ($n=390$). After controlling for program- and child-level controls (Model 3), the estimates were nearly identical to models run on the full analytic sample; the differences in the effect sizes for children’s growth on Alphabet Knowledge and Phonological Awareness between the full analytic sample and the omitted sample in the pre-kindergarten year ranged from -.05 to .01.

10. Discussion

Quality Rating and Improvement Systems are an increasingly popular policy approach to assess and improve the quality of early childhood education within a state context. Quality ratings are the central tenant of the QRIS; parental selection, improvement supports and financial incentives are posited to have strong ties to programs’ performance in the rating system. Yet, there is little empirical evidence on whether the complex weighting and aggregation structure used to derive programs’ ratings relates to constructs of interest, namely child outcomes. The present study investigated the extent to which the rating structure of Virginia’s QRIS, the Virginia Star Quality Initiative, related to child’s growth in literacy skills in early childhood. After attempting to adjust for the selection of programs and children in to the VSQ, we found that star ratings were associated with growth in the pre-kindergarten year, such that 3-star and 4-star pre-k programs had sharper growth in the pre-kindergarten year compared to children in 2-star pre-k programs.

10.1. Relation between star ratings and child outcomes

The general convergence in estimated effect sizes across the varying models suggests that child care quality is a modest, but reliable, predictor of growth in pre-literacy skills during pre-kindergarten. The results of this study suggest that attending a 3-star or 4-star program may have positive relations to children’s pre-literacy trajectories, at least in the short run; 3-star and 4-star programs were related to slightly steeper decline in alphabet knowledge over the summer, and by kindergarten, there were no differences in growth.

Importantly, our findings indicate the importance of controlling for variables that may be related to parental selection of care and child development. For instance, without any controls for selection, children in 3-star and 4-star pre-k programs performed around three-fourths of a standard deviation higher compared to 2-star pre-k programs in the pre-kindergarten academic year. This is equivalent to learning about 11 more letters or letter-sounds over the course of the preschool year. With the inclusion of program and child-level controls, the magnitude of the relations declined by over 40 percent, yet still remained significant. Children in 3-star and 4-star pre-k programs had one-third of a standard deviation higher literacy skills compared to 2-star pre-k programs, which equates to children learning roughly six more letters in higher-rated programs than in lower-rated programs.

Yet, we may still be concerned that an explanation for the relation among quality and literacy gains in the preschool year is that higher-rated programs served different types of children than lower-rated programs and our models may have lacked adequate analytic techniques to control for these selection patterns. Indeed, our descriptive portrayal indicated somewhat stark differences in the racial composition among 2-star and higher-rated programs, where 3-star and 4-star pre-k programs had higher more Hispanic children, and 2-star programs had more African American children. Researchers have found evidence on the effects of exposure to preschool differing across racial groups, with Hispanic children potentially benefiting more from preschool than African American children (Castro, Pérez, Dickinson, & Frede, 2011; Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007). For instance, in a quasi-experimental study on Oklahoma’s universal pre-kindergarten program, Gormley et al. (2005) found that Hispanic children improved over 50% on cognitive and language skills in the preschool year compared to African American children who improved 15.2% on language skills and 28.1% on cognitive skills. However, other studies have not found stronger preschool effects for Hispanic, or English language learners compared to African American children, particularly when controlling for household income (Bassok, 2010). In addition, findings hold in our study when we omit Hispanic and LEP children from the model, or when we control for race, suggesting that racial differences cannot completely explain the differences in growth between high-rated and low-rated programs.

The slower literacy growth among children in 2-star pre-k programs could also be due to ceiling effects associated with the testing instrument, which included a fixed number of items for each subscale. This hypothesis may particularly apply to children in 2-star pre-k programs because they started ahead of 3-star and 4-star pre-k programs and thus had less room to grow. Importantly, children in 2-star programs do not come close to the ceiling at the end of pre-kindergarten, and findings are generally robust when we attempt to mitigate the ceiling effects in the tobit regression analysis. However, there is still a valid concern that ceiling effects are lessening the precision of our estimates in the kindergarten year, where the averages of children across each of the three rating groups approach the maximum score of each subscale.

It may also be the case that exposure to care above a 3-star may indicate some sort of threshold of quality. Most of the work on quality standards and child outcomes typically examines linear relations rather than examining whether there are certain cutpoints where the relation to child outcomes becomes stronger or weaker (Burchinal et al., 2010). For instance, Burchinal et al.

(2010) found that quality of instruction measured by the CLASS predicted language skills for low-income children, but the magnitude was much stronger when the teachers provided moderate to high instructional quality compared to low quality. Yet, there is little evidence on thresholds of quality among combined measures of quality, as is the case in the VSQI. This type of threshold analysis is particularly difficult in that it assumes that each quality standard relates to outcomes, and that when combined and aggregated to levels or ratings, the levels distinguish between outcomes. Descriptively, 3-star and 4-star pre-k programs had similar scores on staff qualifications and group size; yet 4-star programs had significantly higher weighted scores on the CLASS the ECERS-R. Importantly, we did not test what quality standard was the driving force behind relations to child outcomes. The exact intervening mechanisms driving the relations between star rating and growth in pre-kindergarten remain to be explored.

We also found interesting patterns in children's literacy skills over the summer between preschool and kindergarten. On average, children's pre-literacy skills plateaued or slightly dropped between the spring of pre-kindergarten and fall of kindergarten. Given that Virginia's pre-kindergarten program does not run year round and typically follows an elementary school calendar, the pattern in our study matches past conceptual and empirical support on summer learning loss (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996). However, it was surprising that 3-star and 4-star programs had slightly larger summer learning loss compared to 2-star pre-k programs. One explanation may be due to differences in children's characteristics among higher- and lower-rating programs. Children in 2-star programs have higher performance at baseline than 3-star and 4-star programs. Moreover, 3-star and 4-star programs serve more Hispanic children, and 4-star programs in particular serve more limited English proficient children. Past evidence suggests that the lack of continuous learning over the summer appears to be particularly detrimental for children from lower socio-economic backgrounds and/or children who are English language learners (Kim, 2006). Potential explanations for summer learning loss include lack of exposure to English over the summer among English language learners (Entwisle, Alexander, & Olson, 1997). In our study, the dip in children's learning loss is rather small. However, it does highlight the complex relation among center quality, the demographics of its children, and children's development.

10.2. Limitations

The present study has several notable limitations. First, the study is a non-causal, descriptive exploration of the relation among VSQI ratings and preschool-aged children's literacy growth through kindergarten. Although the longitudinal approach with community and time fixed effects adjusted for observed and unobserved selection factors at each of the time points, there is a potential for the models to result in a biased estimate if key selection variables are missing or if the unobserved variables have a differential relation on early and later performance. A key concern is that more motivated parents may select higher-quality care for their children, making it difficult to untangle the effects of parents' skills and motivation from the effect of quality. Although this study does not examine the exact mechanisms in which parents select care, descriptive analyses suggested some evidence for differential parental selection of care. Two-star programs had a greater proportion of children from disadvantaged neighborhoods, yet also had higher performance at the start of pre-kindergarten compared to 3-star and 4-star programs. Further research is needed to understand parental selection of care and how to account for these complex selection patterns in QRIS validation work.

In addition, in order to account for differences in the neighborhood characteristics among centers, and to mitigate potential selection effects, we used data from the U.S. Census 2000, which was much earlier than the VSQI ratings took place (2007 and 2008). Therefore, our controls for centers’ neighborhood characteristics may be a bit outdated, and have the potential to lead to biased estimates if the characteristics we fail to account for are related to children's learning. Overall, validation work that employs quasi-experimental or experimental studies to determine the causal influence of programs' ratings and outcomes is greatly needed.

Our study also has limited external validity. Only 13% of the programs in the VSQI that were eligible to participate actually did so from 2007 to 2009. The low density of participation is similar to several other voluntary QRIS, including Minnesota (11%), Iowa (18%), New Hampshire (7%) and Kentucky (26%), suggesting that our results may generalize to other voluntary systems. Because the present study does not have quality ratings for all child care programs in the state, the question remains whether findings will hold if more programs, potentially with a greater range in quality, enter the system.

It is also a limitation of the paper that we are using a sample of pre-kindergarten programs that serve at-risk 4-year-old children and are subject to minimum state-regulations in terms of structural quality. QRIS often serve children from birth through the end of preschool, and are targeted to improve child outcomes across multiple domains. It is a limitation that our sample only includes 4-year-old children and does not include other age groups. In addition, the pre-kindergarten programs in our study were subject to minimum requirements in terms of structural features and staff qualifications. Therefore, although this dataset might be appropriate for considering the relation between QRIS ratings and children's learning, it might not represent the full range of programs at the lowest spectrum of quality.

According to the National Institute for Early Education Research (NIEER), Virginia's quality benchmarks are middle-of-the-road, suggesting that results may generalize to other state-funded pre-k programs. Indeed, in our dataset, there was still was a fair amount of variability among pre-kindergarten programs' quality, as evidenced by the range of quality within each indicator, and the range of 2-star, 3-star, and 4-star ratings.

In addition, our study also only included literacy as an outcome, and did not include other important components of children's functioning such as math skills, social competence, and self-regulation. We were limited to literacy as our sole outcome because the state only collected literacy performance among pre-kindergarten children using the PALS. The PALS measures also have several notable limitations. Although every effort was made to ensure that assessments at each data collection point (e.g., fall pre-k) were conducted close together, there was a range in when they were collected. For example, the time between the fall and spring assessment in the preschool year ranged from 140 days to 250 days, albeit the standard deviation was quite small (12 days) and there was no significant difference in intervals between assessments by star rating. Although we control for children's age at the time of assessment as well as the difference in the length of time between assessments, there is the potential that it could bias the results.

11. Future directions

Although we find some traction that pre-k star ratings relate to growth in the pre-kindergarten year, these relations fade out relatively quickly. The central question remains whether there is a better way to combine and weight the quality standards that predict development across early and later schooling. The aggregation technique in the VSQI, where multiple indicators of quality
are summed together, may explain part of the reason that relation among VSQI ratings and outcomes fades over time. It may be that some of the quality standards are driving the effects more so than others. For instance, in the VSQI, there is little evidence that several of the quality indicators, such as transition practices or director qualifications, relate to outcomes, and weak direct relations between other more structural indicators of quality, such as group size and teacher education levels, and outcomes (Early et al., 2006; NICHD ECCRN, 2002). Although other quality standards, such as teacher-child interactions, have larger, albeit somewhat modest, relations to outcomes, the inclusion of certain potentially non-significant quality standards in the overall rating may shrink rating effects.

The evidence that star ratings modestly predict growth in pre-kindergarten with the association fading out in the next year brings to question whether ratings could be structured in such a way that they predict both short- and longer-term outcomes. Future work on QRIS should include simulations of the weighting structure and aggregation techniques, as well as the relative contribution of each quality standard to outcomes. Some evidence suggests that more structural elements moderate the relation between process quality and outcomes, signifying that a more valid and comprehensive rating structure may account for this complex relation (Mashburn et al., 2008). Future work may explore the indicators of quality that most strongly relate to child outcomes and how to best structure their inclusion in a QRIS.

In some sense, validation work may be a bit premature without better understanding the implications of measurement and assessment procedures. For instance, research on the ideal number of classrooms to observe within each program, data management procedures, methods to assess quality standards (e.g., self-report versus observations), and measures to assess children’s development remain largely separate from validation work on QRIS. In order to ensure that QRIS are employed with the greatest validity, researchers should continue to explore the relation between quality standards and outcomes, as well as explore the extent to which findings vary as a function of aggregation techniques and measurement choices.

12. Conclusion

Quality Rating and Improvement Systems have dramatically expanded over the last decade in response to the growing need to improve the quality of early childhood education programs and outcomes for children. Despite the strong theoretical model for QRIS, there is little descriptive understanding of the issues related to QRIS or the alignment of QRIS ratings to empirical evidence of their effects. The evidence that higher star ratings are related to stronger growth in children’s pre-literacy skills in pre-kindergarten year is encouraging. Yet the magnitude of the association is attenuated after accounting for selection factors, and the relations to children’s literacy skills fade out by the kindergarten year, suggesting that much more work is needed in understanding the implications of measurement and rating choices on predicting child outcomes.

Ultimately, we are interested in whether QRIS improve outcomes for children, particularly for children who come from disadvantaged backgrounds and contexts. This paper only begins to address the research questions related to QRIS. Much more work is needed in validating ratings, examining selection, and understanding effects on parental selection. Given the demand and public attention for high-quality care, evaluations on the effectiveness of QRIS become ever more pressing. The ability of the field to address such questions can only improve as investigations continue to explore QRIS in practice.

Acknowledgments

The authors are grateful for the guidance and support of this study provided by the Virginia Early Childhood Foundation, the Virginia Office of Early Childhood Development, and the Virginia Department of Education. We would like to acknowledge Debra Jonas, Zelda Boyd, Maria Brown, Kathy Glazer, Jason Downer, Andrew Mashburn, Amanda Williford, and Yu Cao for their contributions to our work. The research reported here is supported by the Child Care Research Scholars Grant, Administration for Children and Families, U.S. Department of Health and Human Services. The study is also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040049 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education or the Virginia Department of Education. Research contained herein was conducted using data provided under a contract with the Virginia Department of Education. The content does not necessarily reflect the views or policies of the Virginia Department of Education, the Virginia Board of Education, or the Commonwealth of Virginia. Consequently, the Virginia Department of Education, the Virginia Board of Education, and the Commonwealth of Virginia are not responsible for the research brief’s content or any loss suffered due to the use of such content. Moreover, the mention of any trade names, commercial products or organizations in this research brief is not an endorsement of any of these entities by the Virginia Department of Education, the Virginia Board of Education, or the Commonwealth of Virginia.

Appendix.

Table A.1.

<table>
<thead>
<tr>
<th>Quality standard</th>
<th>Indicator</th>
<th>ECE program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard 1: Education, Qualifications, and Training</td>
<td>Director qualifications</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Teacher qualifications</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Assistant teacher qualifications</td>
<td>5</td>
</tr>
<tr>
<td>Subtotal: 40 points</td>
<td>All staff professional development</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Director professional development</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>All instructional STAFF professional development</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Teacher professional development</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Assistant teacher professional development</td>
<td>5</td>
</tr>
<tr>
<td>Standard 2: Interactions</td>
<td>Emotional support</td>
<td>20</td>
</tr>
<tr>
<td>Subtotal: 60 points</td>
<td>Classroom organization</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Instructional support</td>
<td>20</td>
</tr>
<tr>
<td>Standard 3: Structure</td>
<td>Staff-to-child ratio</td>
<td>20</td>
</tr>
<tr>
<td>Subtotal: 30 points</td>
<td>Group size</td>
<td>10</td>
</tr>
<tr>
<td>Standard 4: Environment and instruction</td>
<td>Overall learning environment subscales</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Transition practices</td>
<td>10</td>
</tr>
<tr>
<td>Subtotal: 40 points</td>
<td>Total Points</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Star Rating</td>
<td>5</td>
</tr>
</tbody>
</table>


