Evaluation of Rhode Island's BrightStars Child Care Center and Preschool Quality Framework

Kelly Maxwell, Amy Blasberg, Diane Early, Winnie Li, and Nadia Orfali

Child Trends

December 2016

Suggested citation:

Maxwell, K.L., Blasberg, A., Early, D.M., Li, W., and Orfali, N. (2016). *Evaluation of Rhode Island's BrightStars Child Care Center and Preschool Quality Framework*. Chapel Hill, NC: Child Trends.

Acknowledgments:

Many individuals helped make this project a success. We would like to thank Karen Beese, Deb Anthes, and Jael Lopez from the Rhode Island Department of Human Services for their leadership and support of our work. We are also grateful to Lisa Hildebrand and the BrightStars team at the Rhode Island Association for the Education of Young Children for their assistance in answering questions and providing needed information. This project could not have been completed without the skilled individuals who collected data for us: Amanda Ayotte, Heather Chapman, Erinn Huetteman, Heather Jackson, and Shari Leskowitz. We would also like to thank Kathryn Tout, our colleague at Child Trends, for sharing her expertise about QRIS evaluation and reviewing this report and Jennifer Abrams of Child Trends for her significant contributions to the data collection effort. We appreciate the work of August Aldebot-Green and Catherine Nichols, from the Child Trends communications team, who designed the final version of this report.

We extend a special thank you and express our gratitude to the program directors and teachers who participated in the study and to the parents who granted permission for their children to participate in this study. We hope that this evaluation report will be used to support the continued improvement of BrightStars and the programs that serve young children in Rhode Island.

This project was funded by the Rhode Island Race to the Top-Early Learning Challenge grant, and we would like to thank the Early Learning Challenge grant leaders for their guidance and support of this evaluation.

Publication #2016-60

drop us a line

We'd love to hear your thoughts on this publication. Has it helped you or your organization? Email us at feedback@childtrends.org

childtrends.org

Table of Contents

List of Tables

List of Figures

BrightStars, Rhode Island's Quality Rating and Improvement System (QRIS), was launched in 2009 and provides a quality framework for licensed child care centers and preschools, family child care homes, and school-age programs. As stated in its mission (Rhode Island Association for the Education of Young Children, 2016), BrightStars:

- helps families access quality child care, early learning, and school-age programs;
- helps child care providers learn about best practices in early learning and apply them to the care children receive; and
- recognizes program quality and gives parents information to make choices about their children's care and education.

Rhode Island early childhood leaders have grounded the development of BrightStars in research. Criteria were selected that research suggests are important for early care and education quality and children's development, and all three of its frameworks (centers/preschools, family child care, and school-age programs) were pilot-tested before statewide launch (Maxwell, 2008; Rhode Island Department of Children, Youth & Families, Rhode Island Department of Human Services, and Rhode Island Department of Education, 2014). Rhode Island leaders have continued their commitment to a data-based approach in conducting this evaluation of BrightStars.

The current BrightStars QRIS reflects two major efforts as part of Rhode Island's Early Learning Challenge (ELC) grant.

1. Rhode Island's Early Learning Council and ELC leadership worked to strengthen alignment across multiple aspects of the early learning system to improve efficiency and develop a shared vision of program quality. BrightStars is one of three sets of standards—child care licensing and Rhode Island Department of Education Comprehensive Early Childhood Education Program Approval Standards are the other two—that provide the framework for Rhode Island's early childhood system. The ELC Program Standards Alignment Subcommittee spent the first two years of the ELC grant reviewing and revising all three major sets of standards to strengthen alignment. The 2013 revision of BrightStars reflects this work to more closely align the various standards and ensure that each set addresses criteria that are important in supporting children's development and learning.

2. As of April 2014, Rhode Island early childhood leaders moved BrightStars from a completely voluntary system to a system that required BrightStars participation for all programs participating in the Department of Human Services' Child Care Assistance Program (CCAP).

The purpose of this evaluation was to collect and analyze data regarding BrightStars to help Rhode Island leaders continue to refine and strengthen its QRIS. This evaluation provides important data for determining whether the revised framework produces rating levels that meaningfully differentiate higher quality programs from lower quality programs. As required in the Early Learning Challenge Grant, this evaluation also examines the extent to which there are relationships between measures of children's growth and BrightStars ratings and standards.

Description of BrightStars

The 2013 BrightStars framework for child care centers and preschools includes 23 criteria grouped into 10 standards across 6 domains. (See Table 1 for an overview.) BrightStars uses a block rating structure to create a program rating that ranges from 1 star to 5 stars. In a block structure, programs must meet all criteria at a certain level to be awarded that level. Standards and criteria are not averaged or summed together. To earn a 2-star rating, for instance, programs must meet 100 percent of the criteria at the 2-star level. Across the country, 48 percent of QRIS use a block structure; 15 percent use a point structure (i.e., programs earn points for meeting various criteria and points are summed to determine the rating); and 38 percent use a hybrid structure that relies on both blocks and points for some criteria or levels (QRIS Compendium, 2016).

Table 1. BrightStars child care center and preschool standards[1]

Standards	Criteria
Health, safety and nutrition	
1. Learning environment	Regulatory compliance, Learning environment training
Enrollment and staffing	
2. Minimum staff-child ratio	Staff-child ratio, Ratio posted
3. Maximum group size	Group size, Group size posted
Staff qualifications and ongoing professional development	
4. Teacher qualifications	Professional development plan, Formal education, Rhode Island Early Learning and Development Standards (RIELDS) training
5. Program leadership	RIELDS training-administrator, RIELDS training-education coordinator
Administration	
6. Continuous quality improvement	Program self-assessment, Quality improvement plan
Early learning and development	
7. Curriculum	Written curriculum, Teaching and interaction training
8. Child assessment	Developmental screening info, Child assessment
9. Inclusive classroom practices	Written program philosophy, Staff release time
Family engagement	
10. Family communication and involvement	Family communication, Family-teacher conference, Advisory board

[1] Adapted from p. 3 of BrightStars 2013 Child Care Center and Preschool Quality Framework.

As of July 2015, the time we started recruiting centers and preschool programs to participate in this study, 261 centers and preschools were part of BrightStars. Of those, 226 were rated using the 2013 BrightStars framework and were eligible to participate in the evaluation. The ratings of these programs are provided in Table 2.

Table 2. BrightStars ratings for all participating centers and preschools

Rating level	# of programs rated under 2009 and 2013 frameworks (July 2015)	# of programs rated under the 2013 framework (August 2015)
1-star	110 (42%)	103 (46%)
2-star	59 (23%)	64 (28%)
3-star	41 (16%)	32 (14%)
4-star	36 (14%)	23 (10%)
5-star	15 (6%)	3 (1%)
total	**261**	**226**

Source: BrightStars administrative data from the Rhode Island Association for the Education of Young Children (July 2015; August 2015).

QRIS Validation

This evaluation project is framed within the larger QRIS validation context. A 2012 Quality Rating and Improvement System validation brief sponsored by the Office of Planning, Research and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services (Zellman & Fiene, 2012) defines QRIS validation as a: "multi-step process that assesses the degree to which design decisions about program quality standards and measurement strategies are resulting in accurate and meaningful ratings." As suggested in the phrase "multi-step process," QRIS validation is not a one-time event but rather a series of activities that support the continuous improvement of the QRIS. Validation studies do not yield a yes or no answer (i.e., "Yes, the QRIS is valid") but rather provide detailed information used to support the continuous improvement and refinement of the QRIS.

Zellman and Fiene's 2012 validation brief describes four interrelated phases of QRIS validation:

1. examining the concepts included in a QRIS to ensure that they reflect the key aspects of quality,
2. examining the strategies used to measure each concept in the QRIS,
3. determining the extent to which the overall program ratings accurately reflect differences in quality, and
4. determining the extent to which the program ratings are associated with measures of children's development.

Typically, states address different aspects of validation at different times in the development of a QRIS. Each of the four phases is described briefly below and tied to the Rhode Island evaluation.

Examining key concepts in the QRIS. The first phase addresses whether the QRIS includes standards and criteria that are important components of quality and that are related to children's growth and development. This aspect of validation is typically done during the development and design phase of a QRIS and can include expert review as well as a review of the research. It is a critical step to ensure that the QRIS measures what is most important. Because Rhode Island has already addressed the validity of the key concepts included in its BrightStars framework, this was not a major emphasis of the current evaluation project (Rhode Island Department of Children, Youth & Families, Rhode Island Department of Human Services, and Rhode Island Department of Education, 2014). However, information about the relationship between particular standards and other measures of quality and children's development will provide useful information in considering the extent to which BrightStars includes key concepts of greatest importance to the state.

Measurement strategies. This phase of validation examines the strategies used to measure the standards. This includes examining whether the measurement strategies are rigorous and working as intended. Because program ratings are based on the information gathered for each standard, it is important to ensure that each standard is measured well and functions as intended. If the measurement strategies are not strong, then the ratings will not likely do their job of sorting high- and low-quality programs. In this report we use the data collected by BrightStars staff in determining the rating for the programs participating in our study to examine the extent to which the standards are measuring a single quality construct and the extent to which the standards are related to one another.

Program ratings as indicators of quality. The third validation phase looks closely at the QRIS ratings and the extent to which the ratings reflect meaningful differences in quality. Addressing this question is one of the two requirements regarding QRIS evaluation in the ELC grant and was addressed in this evaluation. This is an important question because program ratings are made public and are often tied to resources (e.g., access to technical assistance, tiered subsidy reimbursement). Thus, policymakers and the public should have confidence that the overall program ratings are meaningful and reflect quality. We address this question by presenting data comparing programs' BrightStars ratings with scores on another measure of quality that is not used in determining BrightStars ratings.

Relationship between ratings and measures of children's development. The fourth validation question, also required by the ELC, addresses the extent to which ratings are related to measures of children's development and learning. In other words, are development and learning greater for children who attend higher rated programs as compared to those who attend lower rated programs? Although QRIS were developed to identify high-quality programs and support quality improvement that will benefit children, they may not include the precise quality factors that support specific areas of development (e.g., language) or identify high thresholds of quality necessary to influence children's development. We examine this question in the evaluation to provide data to inform Rhode Island leaders' understanding of the current BrightStars framework and to inform future revisions. Impact statements or a determination that the QRIS caused children to learn more are not possible because they would require random assignment of children to programs. However, information about the relationship between program ratings, standards, and measures of children's development are presented in this evaluation.

Primary Evaluation Questions

This evaluation was designed to address two main questions.

1. Is there evidence that the BrightStars rating and standards are measuring quality? Specifically, is there a relationship between the BrightStars rating and standards with the Classroom Assessment Scoring System (CLASS Pre-K: Pianta, LaParo, & Hamre, 2008; CLASS Toddler: LaParo, Hamre, & Pianta, 2012), another quality measure focused specifically on teacher-child interactions?

2. Is there evidence that the BrightStars rating and standards are related to measures of children's development and learning?

We also examined the relationship among the 10 BrightStars standards for programs that participated in the evaluation and the extent to which the relationship between star rating and children's development differed by family income.

Study Design and Procedures

Highlights

- The evaluation included 71 BrightStars child care and preschool programs (21 one-star, 21 two-star, 15 three-star, 11 four-star, and 3 five-star).
- Independent classroom quality data were collected using the Classroom Assessment Scoring System (CLASS) Pre-K in all 71 participating programs; CLASS Toddler data were collected in a subset of 32 of the programs.
- Fall and spring assessments of early academic, executive function, and social-emotional skills were collected from 332 children in 65 of the participating programs.
- Parents of participating children completed questionnaires about their family and child's demographic characteristics.
- Directors and teachers completed questionnaires about their demographic characteristics, their programs, and their impressions of BrightStars.

This section contains key information about the study design and procedures. See Appendix A for a more detailed description.

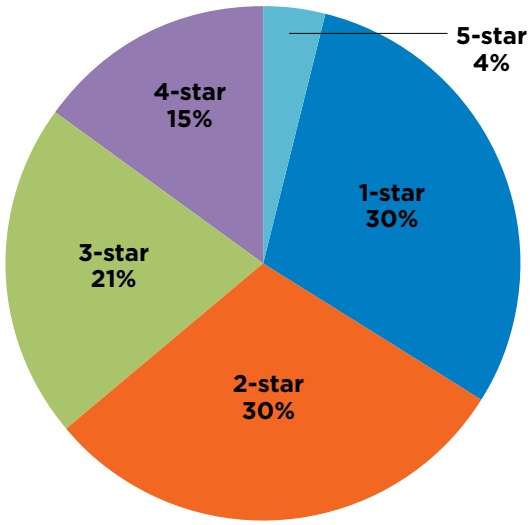
Recruitment of programs and classrooms

We attempted to recruit 242 center-based programs rated under the BrightStars 2013 Child Care Center and Preschool Quality Framework. Programs rated under this framework included child care centers, Head Start programs, and school-based programs. Child Trends obtained contact information for all eligible programs from the Rhode Island Department of Human Services. The 242 programs we attempted to recruit included 226 that were rated by August and were recruited for the full study plus 16 three-, four-, and five-star programs that were recruited in the winter to augment the sample at the higher star levels. Those recruited in the winter participated only in the classroom observation portion of the study. The sample of programs in this evaluation included child care centers, preschool programs, Head Start programs, and school-based programs.

Recruitment of the main cohort took place over a four-month period in the summer and fall of 2015. Original benchmarks were to recruit 50 one- and two- star programs and 50 three-, four-, and five-star programs. That goal was met for the lower star ratings, but not for the higher star ratings. Out of 167 total 1- and 2-star eligible programs, 42 participated (25 percent). Out of 74 three-, four-, and five-star programs, 29 participated (39 percent). In total, 71 programs (29 percent) took part in the study; 65 for the full study and 6 for the winter only. Figure 1 shows the distribution of programs in the final sample by star rating: 21 one-star, 21 two-star, 15 three-star, 11 four-star, and 3 five-star.

If there was more than one 4-year-old classroom at the center, the recruiter randomly selected a 4-year-old classroom to participate during the recruitment call. In programs where all of the preschool-aged classrooms served mixed ages, preference was given to the classroom that served the most children who would be entering kindergarten the next year.

Figure 1. Distribution of star ratings in the program sample



Source: Child Trends recruitment data and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015).

The study sample was contacted in spring 2016, concurrent with post-test child assessment visits, to see if programs wanted to participate in an additional classroom observation in a toddler classroom. Thirty-two programs agreed to participate, 19 programs did not have toddler classrooms, and the remaining programs either declined or did not respond to our attempts to contact them.

Recruitment of children and families

Children were selected at random for participation from all those who were eligible. Eligible children were those who 1) had parent permission, 2) were at least 3 years of age, 3) attended the program at least 9 hours per week. All assessments were administered in English.[2]

In the 65 preschool classrooms, teachers reported that there were a total of 959 children enrolled. Overall, 497 parents (52 percent) gave permission for their child to participate in the study (an average of 8 per classroom, with a range of 3 to 17). The final sample includes 332 children with both fall and spring data. Within classrooms, the average number of participating children was 5 (range 2 to 8).

Information collected

Child assessments

All participating children were given a brief set of assessments in the fall and spring, designed to test their expressive language, pre-literacy, math, and executive function skills. The assessments battery included:

- Three subscales of the Woodcock Johnson Tests of Achievement, 3rd edition (WJ-III; Woodcock, McGrew, & Mather, 2001):
    - Picture Vocabulary: measures expressive language
    - Letter-Word Identification: measures pre-literacy skills
    - Applied Problems: measures math skills
- The Pencil Tap Task (Smith-Donald, Raver, Hayes, & Richardson, 2007), a measure of executive function (i.e., attentional skills that help him/her focus, remember, and plan).

[2] There were only four instances where a parent had indicated on the demographic questionnaire that a child's English language skills were *poor*. In those cases, the data collector spoke with the teacher. In all cases the teacher confirmed that the child could be assessed in English, and the child was assessed.

Assessments were administered to 387 children in the fall between the months of September and November. Spring assessments were administered to 332 of the original 387 children between the months of April and June. Children were assessed in the spring an average of 6.6 months (range 5.8 to 7.5 months) after their fall assessment. The 55 children who took part in the fall assessment but not the spring were no longer enrolled in that program.

Teacher report of social-emotional and behavior skills

In the fall and spring, at roughly the same time as the child assessments, teachers were asked to report on children's social-emotional and behavior skills, using the standardized tools described below:

- Social Competence and Behavior Evaluation short form (SCBE-30; LaFreniere & Dumas, 1996), which measures:
    - social competence (emotionally mature, pro-social behaviors),
    - anger aggression (oppositional behaviors, poor frustration tolerance), and
    - anxiety withdrawal (anxious, depressed).
- Preschool Learning Behavior Scale (PLBS; McDermott, Leigh, & Perry, 2002), which measures learning behavior such as motivation, persistence, and attitude toward learning.

These teacher reports were completed for 361 (93 percent of 387) children in the fall and 320 (96 percent of 332) children in the spring.

Classroom observation measures

The *Classroom Assessment Scoring System Pre-K* (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008) is an observational tool that assesses the quality of the interactions between children and teachers in preschool classrooms (ages 3 to 5 years). The CLASS Pre-K has 10 subscales that fall under three domains: Emotional Support, Classroom Organization, and Instructional Support. Each subscale is scored on a 7-point scale with scores of 1 and 2 considered low-quality, 3-5 considered mid-range quality, and 6 and 7 considered high-quality. Observers rate the classrooms and teachers on each of the dimensions every 30 minutes throughout the observation morning.

The *Classroom Assessment Scoring System Toddler* (CLASS Toddler; La Paro, Hamre, & Pianta, 2012) is an observational tool that assesses the quality of the interactions between children and teachers in toddler classrooms (ages 15 to 36 months). The CLASS Toddler has eight dimensions that fall under two domains: Emotional and Behavioral Support and Engaged Support for Learning. The scoring for the CLASS Toddler matches that of CLASS Pre-K.

Our team completed a total of 71 CLASS Pre-K observations as part of this study. The majority of these observations (n = 67) were conducted in January, February, and March. Four additional observations were conducted in April or May due to lead teacher absences. CLASS Toddler observations (n = 32) were conducted in May, June, and July.

Questionnaires

We developed questionnaires for parents, teachers, and directors. The parent questionnaire included 11 items focusing on child and family demographics including child gender, date of birth, parental education, and household income and size. Parent questionnaires were returned for all 332 children who participated in the fall and spring assessment and are in the current analyses.

The teacher questionnaire included demographic items such as highest level of education, major area of study, and years of experience. Teachers were also asked about their opinion of BrightStars. Sixty-eight (96 percent of 71) preschool teachers and 31 (97 percent of 32) toddler teachers completed the questionnaire.

The director questionnaire collected the same demographic information as the teacher questionnaire. In addition, directors were also asked questions about their program including

total capacity, enrollment numbers by age group, teacher turnover, and whether or not the center accepts children receiving child care subsidies. The last set of questions focused on participation in BrightStars including why the director decided to participate, what supports the center used, and the director's overall impressions of BrightStars. Seventy (99 percent of the 71) directors completed the questionnaire.

Staff hiring and training

In preparation for a year-long fielding effort, we hired three part-time data collectors. Data collectors were trained on the protection of the rights of research participants, child abuse and neglect protocol, the child assessment battery, and other data collection procedures during a 2-day training conducted in Rhode Island by an experienced member of the project team. After training and practice, they submitted videotapes of themselves administering the assessment to ensure that they were following all procedures.

In December 2015, the data collection team was trained to conduct CLASS Pre-K observations by a certified Teachstone[3] affiliate trainer. Following the training, all three data collectors achieved reliability on the CLASS Pre-K in alignment with Teachstone's requirements. To ensure ongoing interrater reliability on the CLASS, a calibration video was coded by the team in mid-February; each data collector's scores were within one point of the master codes 85 percent of the time, on average. Observations were also coded in pairs throughout the fielding window. During nine double-coded observations, the team's scores were within one point of one another 91 percent of the time, on average.

In April 2016, we hired additional data collectors to conduct CLASS Toddler observations. As with the CLASS Pre-K, the CLASS Toddler observers were trained and tested using Teachstone's standards. The process for maintaining interrater reliability on CLASS Toddler was similar to the process for CLASS Pre-K. The team conducted observations in pairs for their first visits. The two team members who conducted the majority of the observations (94 percent) were 96 percent reliable with each other, on average, across three double-coded observations.

Data entry and verification

All data were double-entered to ensure accuracy. When both rounds of data entry were complete and the files were in equivalent formats, a Child Trends staff member used computer software to find inconsistencies between the two files. All inconsistencies were then checked against the hard copies and corrected to create a final data file for each source. The last step of the process was to spot-check 10 percent of each data file to ensure accuracy with the hard copies.

Data available for analysis

This report includes all director, teacher, and parent questionnaire data, CLASS observations, and child assessment data. Child assessment and parent questionnaire data are included for children who participated in both the fall and spring assessments. Table 3 shows the sample sizes for each type of data.

[3] Teachstone is the company that oversees training and reliability testing on the CLASS tools.

Table 3. Data available for analysis

Type of data[4]	Valid data
Director questionnaire	70
Preschool teacher questionnaire	68
Toddler teacher questionnaire	31
CLASS Pre-K	71
CLASS Toddler	32
Assessment of early academic and executive function children's skills	332 children in 65 classes
Teacher reports of children's social-emotional skills (SCBE, PLBS)	320 children in 63 classes
Parent questionnaire	332 children in 65 classes

Source: BrightStars Evaluation (Child Trends, 2015-2016).

Findings

This section begins with descriptive information about the programs, teachers, and children who took part in the study. That section is followed by a description of children's development during the study year. Next, we turn to information about the BrightStars standards. We provide information about the extent to which the various standards that are used to determine the BrightStars rating measure a single underlying construct of quality. The descriptive information is followed by analyses of the two major research questions: Do BrightStars ratings and standards differentiate levels of quality? And are star ratings and standards related to children's academic and social development? We end with a summary of director and teacher reasons for enrolling in BrightStars and their impressions of BrightStars.

What are the characteristics of programs, directors, and teachers in the study?

To better understand the study sample, we start by presenting descriptive information about the programs, directors, and teachers who took part. As seen in Table 4, centers and preschools varied widely in enrollment and capacity. On average, just over one third of the children in these programs received a child care subsidy. Almost all programs served preschoolers, with smaller numbers serving younger and older children.

Lead teacher turnover[5] in the past 12 months was just over 20 percent, again with a large range. The median lead teacher turnover was 15 percent, meaning that half the programs were higher and half were lower than 15 percent. Only a small portion of the programs received Head Start, Early Head Start, or State Pre-K funds, and the majority were for-profit.

[4] For teacher reports of children's social-emotional skills, these values reflect only children with valid data in the fall and spring because those are the only data in the current analyses. The same is true for the values reported for one-on-one assessments of children. For parent questionnaires (which were administered in the fall only) these values include only children who had a parent questionnaire and valid fall and spring data on the teacher report or one-on-one child assessments, again because those are the children in the analyses.

[5] To measure teacher turnover, the director questionnaire asked how many teachers were employed in their center and how many left and had to be replaced in the past 12 months. The number who left and had to be replaced was divided by the total number of teachers.

Table 4. Characteristics of participating centers and preschools

	Percent or mean (SD)	Median	Range
Star rating (n = 71)	2.35 (1.18)	2	1-5
Enrollment (n = 65-68)			
Licensed capacity	78 (55.60)	60	18-100+
Current enrollment	69 (44.10)	59	12-100+
Had a waiting list	74%		
Accepted subsidy (CCAP; n = 69)	87%		
Proportion of children in the program who received a subsidy (n = 56)	36% (31%)	28%	0%-100%
Age groups served (n = 67)			
Infants (birth to 18 months)	49%		
Toddlers (19-36 months)	75%		
Preschoolers (37 months to kindergarten entry)	99%		
School-aged children (kindergarten or above)	34%		
All ages (infants, toddlers, preschoolers, and school-aged)	19%		
Teachers turnover (past 12 months; n = 66-68)			
Lead teacher	21% (27%)	15%	0%-100%
Assistant teacher	25% (27%)	20%	0%-100%
Funding (can receive more than one source; n = 69)			
Head Start	12%		
Early Head Start	10%		
State Pre-K	7%		
CCAP	78%		
None of the above	6%		
For-profit status (n = 69)	59%		

Source: Child Trends director questionnaire and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).

The center directors in this study were highly experienced, with an average of almost 19 years in early childhood education and almost 7 in their current position. Most had a Bachelor's or Master's degree and most of their degrees were in early childhood (see Table 5).

As described in the Study Design and Procedures section, we randomly selected one preschool classroom and teacher for participation in the study. The participating teachers were highly experienced and educated. On average, preschool teachers had been in the field for almost 16 years and had been in their current positions over 7 years. The toddler teachers had over 13 years of experience in ECE and almost 6 in their current position. Most preschool teachers had a Bachelor's degree or more and over three-quarters majored in early childhood education. All the toddler teachers had attended at least some college and over half had an Associate's degree or higher. Among those with a degree, almost all were in early childhood education. All participating teachers were women.

Eighty-nine percent of preschool teachers reported that they used a curriculum. Of those preschool teachers who used a curriculum, Creative Curriculum was the most common (63 percent). Ninety percent of toddler teachers reported that they used a curriculum. Of those toddler teachers, 71 percent reported using Creative Curriculum.

Table 5. Characteristics of participating directors and teachers

	Directors	Preschool teachers	Toddler teachers
Number of respondents	70	67	31
Average years in ECE	18.88	15.69	13.28
Average years in current position	6.87	7.03	5.84
Education			
Some college	3%	16%	45%
Associate's degree (AA)	11%	18%	32%
Bachelor's degree (BA/BS)	39%	28%	16%
Some graduate work	16%	18%	0%
Master's degree (MA)	31%	19%	6%
Major area of study			
Early childhood	70%	78%	93%
Elementary education	11%	9%	0%
Other education	3%	4%	3%
Other	16%	9%	3%

Source: Child Trends director questionnaire, Child Trends preschool teacher questionnaire, and Child Trends toddler teacher questionnaire (2016).

What are the characteristics of the children in the study?

Table 6 shows the demographic characteristics of the 332 children in this study with both fall and spring data. On average, the children in this study were 4 years and 4 months old (52.8 months old; range 36.5 to 63.7 months) in the fall. The study included an even number of boys and girls. About one quarter of the families of children in the study received a subsidy to attend child care. Household income was relatively high, with over one third of families reporting an income over $80,000, which was the highest category on the questionnaire. The majority of families earned over 250 percent of the federal poverty guideline for their family size, using United States Census Bureau 2015 guidelines.

Table 6. Characteristics of participating children

	n	Percent
Gender	332	
Female		50%
Male		50%
Received a subsidy	332	26%
Household income	311	
Under $15,000		8%
$15,000 to under $30,000		18%
$30,000 to under $50,000		16%
$50,000 to under $80,000		18%
$80,000 or more		41%
Poverty level	300	
Under 100%		17%
Between 100% and 185%		16%
Between 185% and 250%		12%
250% or more		55%
Parental education	324	
Some high school		2%
High school diploma / GED		12%
Some college		20%
Two-year college degree		10%
Bachelor's degree		30%
Graduate degree		25%

Source: Child Trends parent questionnaire (2016).

How did children's skills change during the year?

To understand children's development during the study year, we compared children's fall and spring scores on the child assessments (see Table 7 for a summary; Appendix B presents additional details). The Woodcock Johnson (WJ) scores have been standardized, using a national sample of children of the same age, so that the national mean is 100. Furthermore, when children learn the amount over time that is expected for children their age—again using a national sample as a comparison—their scores remain the same because the standard scores are adjusted to reflect age. When children learn *more* than children typically learn during that same timeframe, their scores go up.

Turning to the children in this study, all the fall and spring mean WJ standard scores for math, pre-literacy and expressive vocabulary were over 100, indicating that the skills of the children in this sample were somewhat higher than national norms. The final column of Table 7 indicates that the changes from fall to spring were non-significant, meaning that these children learned a typical amount over this period.

The Pencil Tapping Test measuring executive function and the teacher ratings of social-emotional skills (SCBE and PLBS) are not standardized, meaning that we anticipate that children's scores will increase as they age and develop. The findings indicate that indeed, executive function skills, social competence, and learning behaviors (i.e., motivation, persistence, and attitude toward learning) went up during the preschool year. Anxiety-withdrawal and anger-aggression remained the same.

Table 7. Children's fall and spring assessment scores

	Fall			Spring			Change
	Mean	Min	Max	Mean	Min	Max	p-value
Math (WJ: Applied Problems)	107.42	68	139	107.56	62	134	n.s.
Pre-Literacy (WJ: Letter Word)	101.71	67	186	102.45	74	176	n.s.
Expressive Language (WJ: Picture Vocabulary)	104.47	71	149	104.26	77	145	n.s.
Executive Function (Pencil Tapping)	9.71	0	16	12.41	0	16	<0.001
Social Competence (SCBE)	39.00	14	60	41.00	17	60	<0.01
Anxiety-Withdrawal (SCBE)	16.32	10	48	15.71	9	46	n.s.
Anger-Aggression (SCBE)	18.24	10	58	17.76	10	59	n.s.
Learning Behaviors (PLBS)	50.37	7	61.67	52.29	22.33	61.67	<0.01

Source: Child Trends child assessment data collection (2015-2016).

How do the BrightStars standards and ratings relate to one another?

BrightStars includes a variety of quality constructs in the star rating. It is important to understand whether and how these quality dimensions relate to each other and if they are measuring a single underlying construct. The odds that the star rating will be related to a more narrowly-defined measure of quality, such as the CLASS, or to children's skills, increases if the star rating measures a single construct. (Burchinal, Tarullo, & Zaslow, 2016). To learn if the 10 standards measured in BrightStars are all measuring a single, underlying construct of program quality, we looked at the correlations among the standards. Also, we calculated Cronbach's Alpha—a statistic used by researchers to understand the extent to which various components are measuring related information—for the 10 standards.

Each standard is significantly correlated with the other nine and with the final star rating. Cronbach's alpha was .97 for the 10 standards, which is very high and provides support for the idea that that all standards are measuring a single construct. See Appendix C for descriptive information about the star rating and standards and correlations among the star rating and all standards.

What is the distribution of standards met for programs at each rating level?

As noted previously, the block structure requires programs to meet all standards at the star rating level they are assigned, which means that programs earn an overall rating based on the lowest level met across the 10 standards. This section provides information about the percentage of programs at each star rating level that met standards at a higher level. As indicated in Table 8, the data we were able to collect from the rating application from participating programs indicated that more than half of all programs in the study with a star rating of 2, 3, or 4 met the requirements for a higher rating level for at least half (i.e., five or more) of the 10 standards. Table 9 indicates the percentage of programs that met a higher level on each of the standards, ordered from most to least common. As seen in Table 9, more than two thirds of programs exceeded their eventual star rating on three standards: minimum staff-child ratio, maximum group size, and inclusive classroom practices. The ratio and group size standards are rated as either a 1 or 5 due to the way BrightStars is constructed, so it is not surprising that many programs met a higher level for these two standards. More than half of the programs met a higher level requirement for 8 of the 10 standards. These findings suggest that programs vary more widely in meeting the quality standards than is reflected in the overall rating earned using a block structure.

Table 8. Percentage of higher standards met for study programs at each rating level (n = 68)[6]

Star rating	% programs that met higher rating level for half or more standards
1	19%
2	62%
3	93%
4	100%
5	N/A

Source: BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).

Table 9. Percentage of study programs that met a higher rating level (than their overall rating earned) for each standard (n = 68)

Standard	% programs that met higher level than their overall Star Rating Earned
Standard 2: Minimum staff-child ratio	71%
Standard 9: Inclusive classroom practices	71%
Standard 3: Maximum group size	68%
Standard 5: Program leadership	62%
Standard 6: Continuous quality improvement	59%
Standard 7: Curriculum	59%
Standard 8: Child assessment	56%
Standard 10: Family communication and involvement	51%
Standard 4: Teacher qualifications	37%
Standard 1: Learning environment	24%

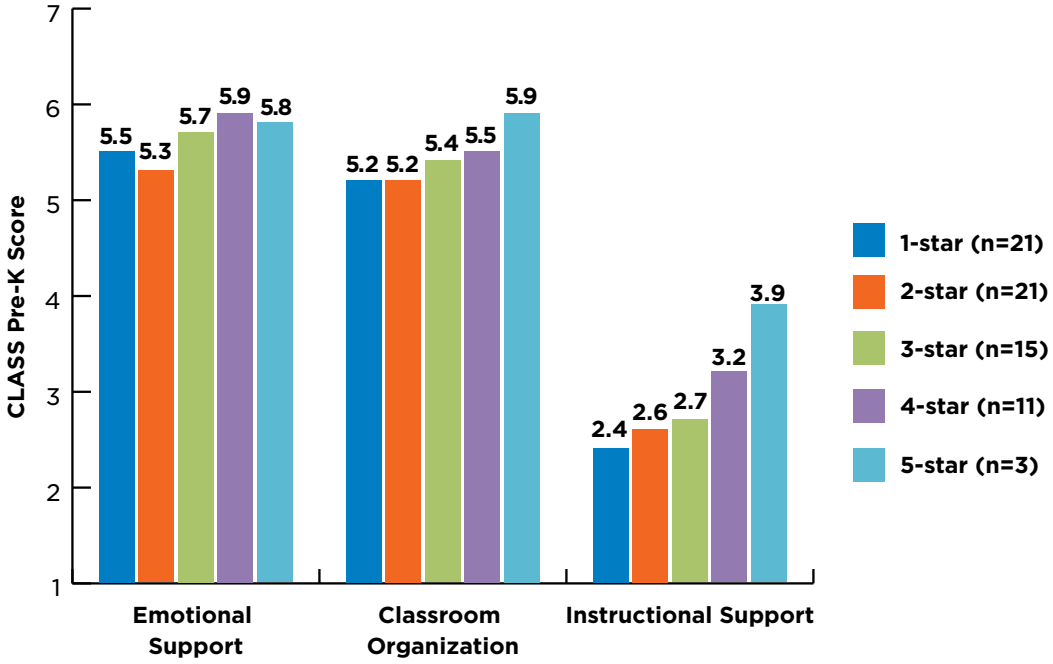
Source: BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).

[6] These percentages exclude 5-star programs because, by definition, those programs achieved a 5 on all 10 standards.

Do BrightStars ratings and standards differentiate levels of quality?

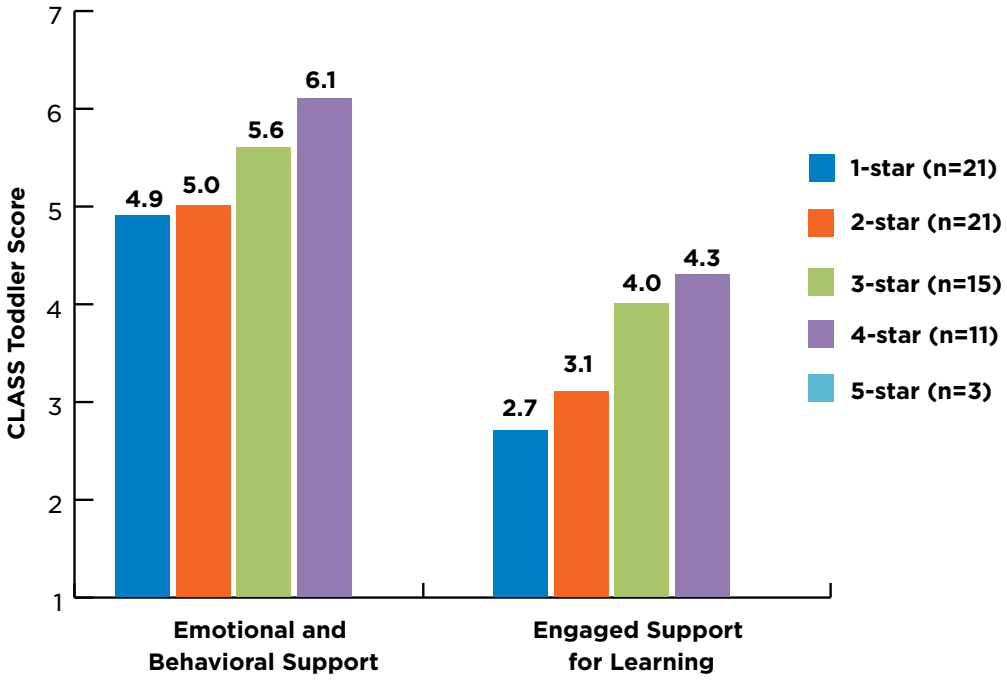
To provide descriptive information, Figures 2 and 3 illustrate the average CLASS score in each domain. As seen in these figures, higher star ratings were generally associated with higher CLASS scores. See Appendix D for additional descriptive information about the CLASS.

Figure 2. Average CLASS Pre-K scores by BrightStars level (n = 71)



Source: Child Trends classroom observation data (2016).

Figure 3. Average CLASS Toddler scores by BrightStars level (n = 32)



Source: Child Trends classroom observation data (2016).

As noted earlier, this study included few programs at the upper star rating levels. This meant that we could not conduct statistical tests comparing each star rating to each other rating. There was, however, enough power to test if there was a link between star rating and our independent measure of quality, the CLASS. To test if star ratings and standards statistically differentiate levels of quality, we conducted regression analyses, a type of statistical tests that considers linear trends. These analyses test the question: Did programs that received higher star ratings, or that had higher scores on the individual BrightStars standards, also score higher on the Pre-K and Toddler CLASS domains?

Table 10 summarizes all statistically significant ($p < .10$) associations from these analyses. We focus our interpretation, however, on those associations that are significant and where the effect size corresponds to the What Works Clearinghouse (WWC; 2014) definition of *substantively important*, meaning an effect size of 0.25 or higher. Table 10 uses a plus sign to indicate that the association was statistically significant and positive[7] and a check mark to denote which are large enough for interpretation. Blank cells indicate that the association was non-significant. See Appendix D for the rationale for this definition, as well as all coefficients, standard errors, and effect sizes for all associations.

As seen on the first row of Table 10, as star rating went up, classrooms scored higher on Emotional Support, Classroom Organization, and Instructional Support scores on the CLASS Pre-K. The remaining rows on Table 10 represent the associations between each of the BrightStars standards and the three CLASS Pre-K domains. Findings indicate that learning environment, teacher qualifications, program leadership, curriculum, inclusive classroom practices, as well as family communication and involvement were associated with all three CLASS Pre-K domains. Additionally, ratios, group size, continuous quality improvement, and child assessment were related to Instructional Support. The ratio and group size standards are rated as either a 1 or 5 due to the way BrightStars is constructed; this limited variability means that reaching statistical significance is less likely.

Table 10. Star ratings and standards as predictors of CLASS Pre-K (n = 71)

	Emotional Support	Classroom Organization	Instructional Support
Star rating	+✓	+✓	+✓
S1. Learning environment	+✓	+✓	+✓
S2. Minimum staff-child ratio			+✓
S3. Maximum group size			+✓
S4. Teacher qualifications	+✓	+✓	+✓
S5. Program leadership	+✓	+✓	+✓
S6. Continuous quality improvement			+✓
S7. Curriculum	+✓	+✓	+✓
S8. Child assessment			+✓
S9. Inclusive classroom practices	+✓	+✓	+✓
S10. Family communication and involvement	+✓	+✓	+✓

Source: Child Trends classroom observation data and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).
Notes: Each row represents three regression analyses—one per outcome. A plus sign indicates that the association is significant ($p < .10$) and positive. A check mark indicates that the effect size for a one-star difference is at least .07, which WWC considers important. See Appendix D for a full description of the analyses and all coefficients, standard errors, p values, and effect sizes.

[7] This table includes only plus signs because all significant associations are positive. A later table includes minus signs to denote negative associations.

The same set of analyses were repeated using the two domains of the CLASS Toddler, and a similar pattern emerged (See Table 11 for a summary and Appendix D for details). Programs with higher star ratings had higher scores on Emotional and Behavior Support as well as on Engaged Support for Learning. Furthermore, all 10 standards were associated with Engaged Support for Learning and all except ratios and group size were significantly associated with Emotional and Behavioral Support.

Table 11. Star ratings and standards as predictors of CLASS Toddler (n = 32)

	Emotional and Behavioral Support	Engaged Support for Learning
Star rating	+✓	+✓
S1. Learning environment	+✓	+✓
S2. Minimum staff-child ratio		+✓
S3. Maximum group size		+✓
S4. Teacher qualifications	+✓	+✓
S5. Program leadership	+✓	+✓
S6. Continuous quality improvement	+✓	+✓
S7. Curriculum	+✓	+✓
S8. Child assessment	+✓	+✓
S9. Inclusive classroom practices	+✓	+✓
S10. Family communication and involvement	+✓	+✓

Source: Child Trends classroom observation data and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).
Notes: Each row represents two regression analyses—one per outcome. A plus sign indicates that the association is significant ($p < .10$) and positive. A check mark indicates that the effect size for a one-star difference is at least .07 and considered substantive. See Appendix D for a full description of the analyses and all coefficients, standard errors, p values, and effect sizes.

Are star ratings and standards related to children's academic and social development?

To test the extent to which star ratings and standards are associated with children's early academic and social development, we conducted statistical analyses that tested associations between the star rating and individual standards and each of the spring child assessments. We employed a statistical technique called hierarchical linear modeling (HLM) to account for the fact that multiple children were in the same classrooms and children within the same classroom are likely to be more similar to one another than children from different classrooms.

These analyses controlled for children's skills in the fall, as well as family poverty (above or below 185 percent of the poverty line for the family's size). Families select child care for a variety of reasons, and families with more education and income may be more likely to send their children to higher quality care (Torquati, Raikes, Huddleston-Casas, Bovaird, & Harris, 2011). By controlling these variables, we statistically remove these preexisting differences among children from the analyses and increase our confidence that any associations with rating or standards in the spring are not being driven by differences that existed prior to the children enrolling in the program.

Table 12 shows the results for early academic skills, executive function, and social-emotional development. Again, we use a plus or minus sign to denote statistically significant associations ($p < .10$), but we focus our interpretation on those where the effect size reaches WWC definition of substantively important. A check mark indicates which associations reach this level. See Appendix D for additional details about the rationale for this definition and Appendix E for all coefficients, standard errors, and effect sizes for all associations. As seen in Table 12, there are some small but statistically significant negative associations on one measure, meaning that higher ratings or standards were associated with lower scores. Because they did not meet the criteria for *substantively important* we do not discuss their meaning.

As seen on Table 12, few of the associations were substantive. Star rating was not substantively associated with any of the child assessment outcomes. Several of the individual standards (teacher qualifications, program leadership, curriculum, inclusive classroom practices, as well as family communication and involvement) were associated with social competence. Additionally, curriculum was associated with math skills.

Table 12. Star ratings and standards as predictors of children's early academic skills and executive function

	Math (WJ: Applied Problems)	Pre-Literacy (WJ: Letter-Word)	Expressive Language (WJ: Picture Vocabulary)	Executive Function (Pencil Tap)	Social Compe-tence (SCBE)	Anxiety (SCBE)	Aggres-siveness (SCBE)	Learning Behaviors (PLBS)
Star rating			-					
S1. Learning environment								
S2. Minimum staff-child ratio			-					
S3. Maximum group size			-					
S4. Teacher qualifications	+		-		+✓			
S5. Program leadership			-		+✓			
S6. Continuous quality improvement	+		-					
S7. Curriculum	+✓		-		+✓			
S8. Child assessment			-					
S9. Inclusive classroom practices	+		-		+✓			
S10. Family communi-cation and involvement	+		-	+	+✓			

Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).
Notes: Each row represents eight HLM analyses—one per outcome. All analyses control for the fall measure of the same outcome and family poverty (above or below 185% of poverty for family size.) A plus sign indicates that the association is significant ($p < .10$) and positive. A minus sign indicates that the association is significant ($p < .10$) and negative.
A check mark indicates that the effect size for a one-star difference is at least .07, which is considered important by WWC. See Appendix E for a full description of the analyses and all coefficients, standard errors, p values, and effect sizes.

Is star rating more strongly related to children's academic and social development for some children than for others?

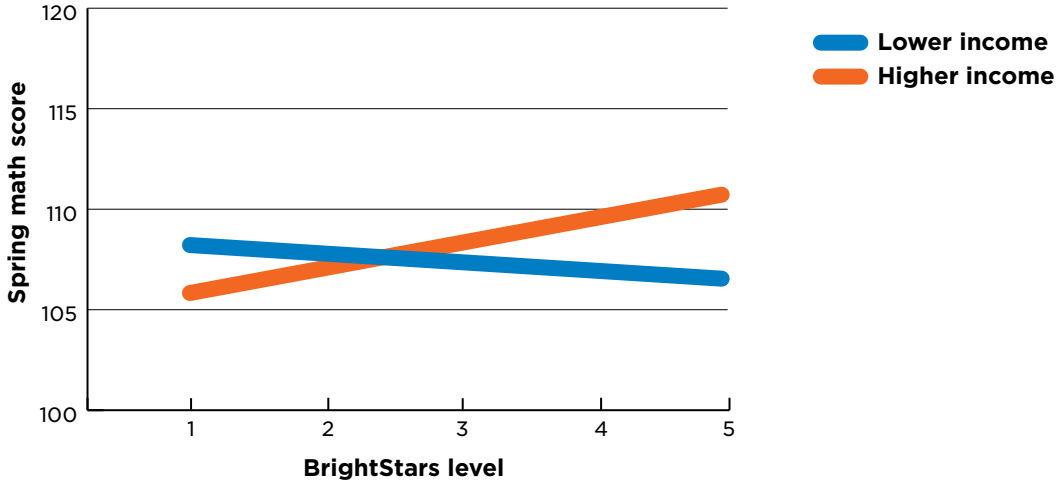
Although there were no substantive associations between star ratings and children's academic or social development overall, it is possible that the associations are stronger for some groups of children than for others.

For instance, there is some evidence that high-quality early childhood programs are especially important for promoting development of children from low-income families (Peisner-Feinberg et al., 2001; Votruba-Drzal, Coley, & Chase-Lansdale, 2004). Thus, it is possible that star rating is more strongly associated with children's development for children from lower income families as compared to those from higher income families. We tested this possibility and describe the findings below. Note that we limited these analyses to star rating only (not individual standards) to minimize the number and complexity of analyses. See Appendix F for full details on these analyses, as well as coefficients, standard errors and effect sizes for all interactions that were tested.

To see if there are different associations between star rating and children's development for children from lower and higher income families, we repeated the star rating analyses that are summarized in Table 12, this time including an interaction between star rating and family poverty. As with other analyses, these analyses controlled for the fall score on the same measure. For these analyses, lower income was defined as children from families below 185 percent of the federal poverty line, and higher income was defined as children whose families are at or above 185 percent of the federal poverty line. The cutoff of 185 percent was selected because it is used to determine eligibility for several federal assistance programs, such as free/reduced price lunch (National School Lunch Program; USDA, 2013) and Women, Infants, and Children (WIC; USDA, 2016). In the current sample, 126 children (36 percent) were in the lower income group and 221 children (63 percent) were in the higher income group.

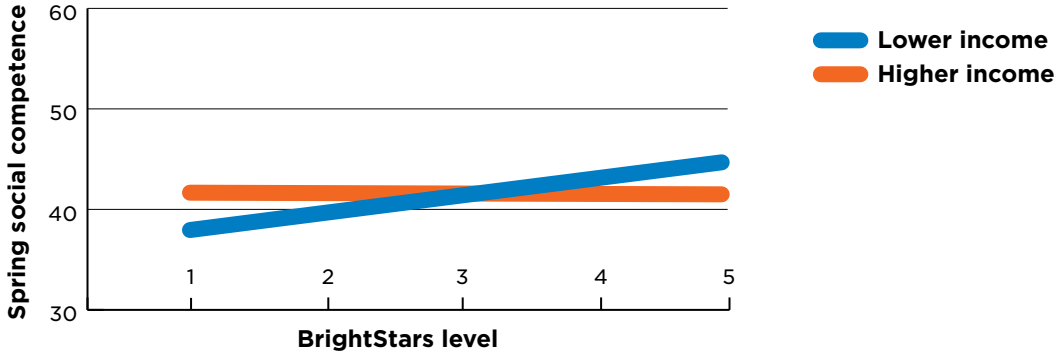
This interaction was statistically significant ($p < .01$) and substantively important for math, social competence, and learning behaviors. There was no association between star rating and spring math scores for children from lower income families, but there was a positive association for children from higher income families (See Figure 4). This contradicts some previous research findings (Peisner-Feinberg et al., 2001). For social competence and learning behaviors on the other hand, the association was in the expected direction. Among children from lower income families, higher star rating was linked to stronger social competence (See Figure 5) and learning behaviors (See Figure 6) in the spring. Among children from higher income families, there was no association between star rating and spring social competence or learning behaviors.

Figure 4. Associations between star rating and spring math scores for children from lower and higher income families



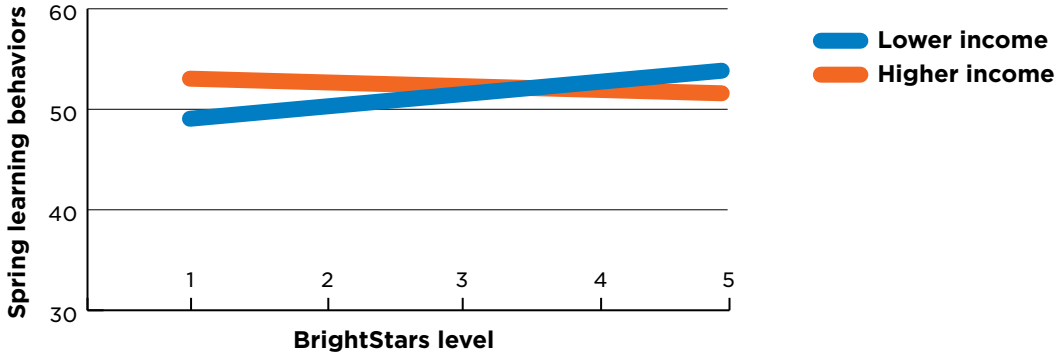
Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).

Figure 5. Associations between star rating and spring social competence for children from lower and higher income families



Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).

Figure 6. Associations between Star Rating and spring learning behaviors for children from lower and higher income families



Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).

In summary, we examined associations between two aspects of BrightStars—star rating and individual standards—and children's early academic and social-emotional skills. When looking at all children together, we found no substantive associations between star rating and children's skills, but 5 of the 10 standards were related to teacher reports of children's social competence. Additionally, the curriculum standard was positively associated with math skills.

When comparing children from lower and higher income families, we found a positive, significant and substantive relationship between star rating and math skills for children from higher income families but not for children from lower income families. We also found a positive, substantive relationship between star rating and social competence and learning behaviors for children from lower income families but not for children from higher income families.

Why did directors enroll in BrightStars and what supports did they access?

On the questionnaire, directors were asked why they decided to enroll in BrightStars. They were given three options to select, plus *other* and asked to check all options that apply. Their responses appear in Table 13. The most common response was that they wanted to be recognized as high-quality. The most common *other* response from directors was that they joined BrightStars to improve their program quality (7 out of 13).

Table 13. Directors' motivation for enrolling in BrightStars

	Yes
Wanted to be recognized as a high-quality program	84%
Wanted access to resources and supports	73%
Was required to participate	47%
Other	19%

Source: Child Trends director questionnaire (2016).

On that same questionnaire, directors were asked which BrightStars supports they had used and were again asked to check all options that apply. Table 14 shows the percent that checked each option, ordered from most to least common. The Center for Early Learning Professionals Professional Development/Technical Assistance and REILDS trainings were the most commonly used.

Table 14. BrightStars supports used by directors

	Yes
The Center for Early Learning Professionals (CELP) Professional Development or Technical Assistance	87%
RIELDS Trainings	87%
Teacher Education and Compensation Helps (T.E.A.C.H) Scholarships	51%
Local Initiatives Support Corporation (LISC) Child Care Facilities Fund Grants	45%
Rhode Island Early Childhood Education and Training Project at Community College of Rhode Island (CCRI)	45%
Department of Human Services (DHS) Quality Awards	31%
LISC Child Care Facilities Fund Technical Assistance (TA)	14%

Source: Child Trends director questionnaire (2016).

Directors were asked which BrightStars support they found most helpful of the ones they had selected. The Center for Early Learning Professionals (CELP) Professional Development/ Technical Assistance was the most frequently reported as most helpful (42 percent), followed by RIELDS trainings (17 percent), and LISC Child Care Facilities Fund Grants (15 percent).

What are center directors' and teachers' perceptions of BrightStars?

The director and teacher questionnaires gathered information about directors' and teachers' perceptions of BrightStars. Additionally, there was an open-ended question that asked them about their overall impressions. This section reports those results.

Directors were asked a series of questions about their impressions of BrightStars, using a five-point scale, where 1 indicated *strongly disagree*, 2 indicated *somewhat disagree*, 3 was *neutral*, 4 was *somewhat agree* and 5 was *strongly agree*. Their average responses are in Table 15.

Responses were generally in the mid-range, with the average score across the eight items just under 4 (somewhat agree). Directors tended to agree that they had made changes to their program as a result of participation in BrightStars. The least strongly endorsed statement was "The BrightStars application process was easy."

Table 15. Mean responses to questions about BrightStars

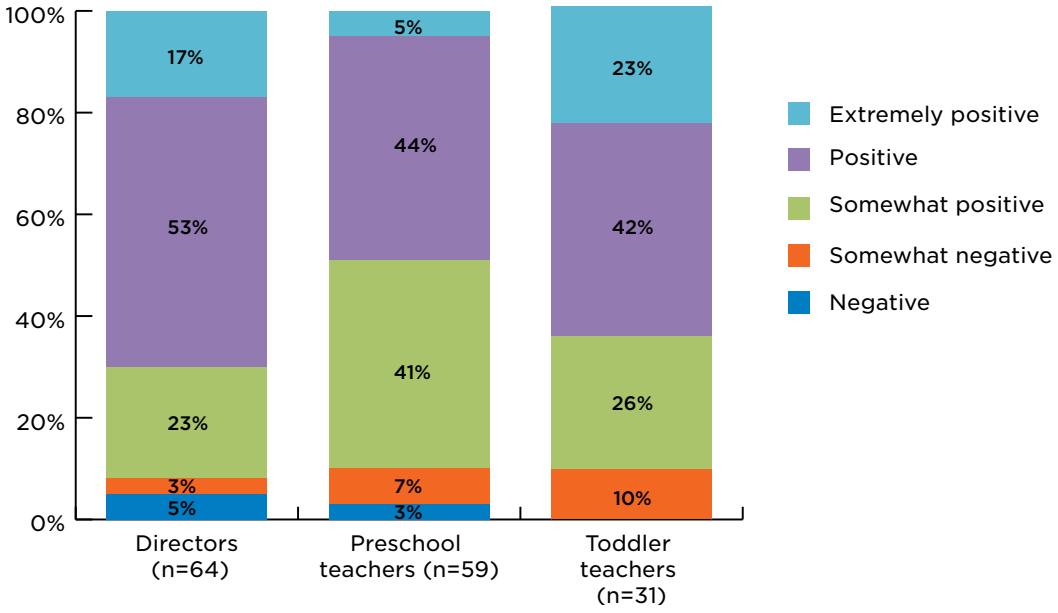
In my opinion…	Mean
We have made changes to our program as a result of joining BrightStars.	4.26
I am able to find the professional development trainings I need.	4.06
BrightStars has been beneficial to my program.	4.06
I would recommend that other programs join BrightStars.	4.04
Teachers are able to find the professional development trainings they need.	3.87
My experience with BrightStars has been what I expected.	3.81
I believe my program is of higher quality because we joined BrightStars.	3.78
The BrightStars application process was easy.	3.49
Average of the 8 items	*3.92*

Source: Child Trends director questionnaire (2016).

Directors, preschool teachers, and toddler teachers were asked: "What is your overall impression of BrightStars?" Response options were on a 5-point scale where 1 indicated *negative*, 2 indicated *somewhat negative*, 3 was *somewhat positive*, 4 was *positive* and 5 was *extremely positive*. Those who indicated no opinion are not included here.

Figure 7 shows the breakdown by response. Over half of directors and toddler teachers rated their impression of BrightStars as positive or extremely positive. Slightly fewer preschool teachers had positive or extremely positive impressions. Star rating was not related to either director or teacher impressions of BrightStars, which provides some assurance that their impressions are not simply driven by how pleased they were with their rating.

Figure 7. Director, preschool teacher, and toddler teacher impressions of BrightStars



Source: Child Trends director questionnaire, Child Trends preschool teacher questionnaire, and Child Trends toddler teacher questionnaire (2016).

After respondents selected their answers from the provided scale, they were also given an opportunity to explain their rating. The open-ended responses varied in length and specificity, detailing a wide range of impressions of the rating system. Below, we summarize the impressions for directors and teachers separately because we realize that the directors, rather than teachers, interact most directly with BrightStars.

Views from directors

Seventy-six percent of the directors who returned the questionnaire provided an explanation for their rating (53 of 70 directors). Of those, 42 percent detailed solely positive experiences and 19 percent detailed solely negative experiences with BrightStars in their open-ended responses. Another 32 percent of directors offered both positive and negative opinions about BrightStars, and 7 percent wrote comments that were neutral or not directly related to BrightStars (e.g., comments on the study questionnaire or expectations of young children). Thus, 74 percent of directors who explained their impressions of BrightStars had something positive to say.

Looking more closely at their comments, some themes emerged. Two of the more common positive comments related to 1) help in improving quality through grants or funds for materials (25 percent) and 2) access to professional development or technical assistance (21 percent). For instance, one director noted, "Being a member has helped us obtain access to resources and support. Made it possible for us to apply for grants to improve our program." Among the negative comments, 19 percent expressed some dissatisfaction with the Environment Rating Scales observation (e.g., is subjective, does not measure some important aspects of quality). One director explained, "We are moving away from quality care/education because teachers are forced to focus on the 'little things.'" Thirteen percent noted that the application process was tedious, time-consuming, confusing, or stressful. Nine percent expressed concern about the block rating structure (e.g., prefer rating to be an average, very rigid). One director wrote, "The scoring system of all or nothing is difficult. Having all 4s and 5s on 9 out of 10 domains and then getting one 3 makes you a 3-star program. Doesn't reflect the quality of the center." Eight percent noted challenges with BrightStars staff (e.g., turnover, unfriendliness).

Views from teachers

Seventy-nine percent of the preschool teachers who returned the questionnaire provided an explanation for their rating (53 of 67 preschool teachers). Seventy-four percent of the 31 toddler teachers who returned the questionnaire explained their overall rating (23 of 31 toddler teachers). Across all 76 teachers who offered their thoughts, 46 percent described solely positive impressions of BrightStars. Twenty-four percent described solely negative impressions, and 28 percent offered mixed impressions. Three percent of the teachers wrote comments that were not directly relevant to BrightStars. Thus, 74 percent of the teachers who explained their impressions of BrightStars had something positive to say.

The most common positive theme from the teachers' responses focused on the ways in which BrightStars participation improved the quality of their program (18 percent). One teacher said she "learned a lot from BrightStars." Another expressed mixed thoughts, saying, "It provides informative guidelines for child care and teachers, although some prove difficult to implement as a classroom teacher." The most common negative reactions to BrightStars focused on the standards themselves (17 percent of teachers) or the ERS observations (16 percent). Teachers described the standards as "unrealistic," "not age-appropriate," "strict," and "unnecessary." One teacher said, "Some of the BrightStars elements discourage or disregard well-trained professionals in the field." In terms of the ERS observations, one teacher said, "For example, teachers are pressured to be concerned with things in the classroom like perfection in their materials in their learning centers … Children are also pressured to be perfect in their cleaning and organizing materials in fear that their center will not acquire a star because the classroom could not meet a 5 or 7 under a certain standard." Other teachers mentioned the subjectivity of the ERS or expressed concern with judging the quality of their program based on a single observation.

In summary, about two thirds of directors and toddler teachers, and almost half of preschool teachers, had a positive or extremely positive impression of BrightStars. They appreciated the supports and focus on quality. Both directors and teachers expressed some concerns about the ERS observations, the standards, and the block rating structure.

Conclusions and Considerations

In this section, we highlight the major findings of the BrightStars evaluation and place the findings in the broader context of ELC validation studies. We also offer future considerations to Rhode Island leaders as they continue to strengthen BrightStars.

This study provides evidence of BrightStars' validity in measuring quality. The quality of preschool classrooms, as measured by all three domains of the CLASS Pre-K (i.e., Instructional Support, Emotional Support, and Classroom Organization), was higher in programs with higher star ratings than those with lower star ratings. Six of the 10 BrightStars standards were also substantively, positively related to all three CLASS Pre-K domains, and the remaining four standards were positively related to Instructional Support. The quality of toddler classrooms, as measured by the CLASS Toddler Emotional and Behavioral Support and Engaged Support for Learning, was higher in programs with higher star ratings than those with lower star ratings. Eight of the 10 standards also were related to both CLASS Toddler scores, and the remaining two standards were positively related to Engaged Support for Learning.

Comparisons between findings in Rhode Island and in other states should be made cautiously because of the wide variability from state to state in QRIS. However, we think it is helpful to briefly review the other research addressing this question to help Rhode Island understand its findings in the larger national context of QRIS validity studies. The evidence that BrightStars is measuring quality is consistent with other studies regarding the validity of QRIS in other states. In a review of QRIS studies addressing a similar aspect of validity, Karoly (2014) reported that seven of seven studies found positive relationships between ratings and independent measures of quality. Of the three other ELC QRIS validation studies that have published findings to date about this question, all of them have reported statistically significant relationships between ratings and an independent measure of quality in centers (Karoly et al., 2016; Tout et al., 2016; Hawkinson et al., 2015).

This study provides evidence of a positive, meaningful relationship between several BrightStars standards and preschoolers' social competence. The curriculum standard was also related to children's math skills. Although there were no substantive relationships between the overall BrightStars rating and measures of children's school readiness skills (i.e., math, language, executive function, and social-emotional skills), 5 of the 10 BrightStars standards were substantively, positively related to children's social competence. The curriculum standard also was related to math skills. None of the standards were substantively related to language skills.

Findings from other studies addressing a similar question (i.e., "Are ratings related to children's skills?") have been mixed. As noted previously, comparisons between findings in Rhode Island and in other states should be made cautiously because of the wide variability from state to state in QRIS. Of the four pre- ELC QRIS studies that Karoly (2014) reviewed and considered to be well-designed, two reported relationships and two did not. Six ELC QRIS validation studies that have addressed this question have published or presented mixed findings to date. Two states, California and Wisconsin, reported no significant relationships between ratings and children's skills (Magnuson & Lin, 2016; Hawkinson & Quick, 2016). California's research team noted that they were able to compare programs only at the top rating levels because there were not enough programs at the lower levels to include in the analysis, which may have limited their ability to find relationships. Delaware reported significant, positive relationships between ratings and children's executive functioning (Karoly et al., 2016). Massachusetts reported some associations between rating levels and children's receptive language and attachment (Roberts & Gerena Melia, 2016). Minnesota reported significant, positive relationships with children's social competence and attention/persistence (Tout et al., 2016). Washington found relationships between ratings and children's receptive language skills (Soderberg et al., 2016). In summary, to date, two other states have found relationships with children's social-emotional development (i.e., MA and MN) and one other ELC state has reported a positive relationship between children's math skills and components of ratings (i.e., Delaware reported a relationship between the qualifications and professional development standard with children's math skills; Karoly et al., 2016).

Although these findings provide evidence of the validity of BrightStars, we recognize that some stakeholders may have expected stronger or more consistent findings. There are several possible reasons why we did not find more consistent relationships between ratings or standards and measures of children's development across the school readiness domains. First, BrightStars, like most QRIS, does not include criteria specifically related to the instructional practices to support a specific domain of development (e.g., language). The BrightStars framework includes some criteria that are related to instruction and classroom practices, such as training on the state's early learning and development standards, having a curriculum framework that is aligned with those standards, and scoring at a particular level on the ECERS-R (which includes a little about instructional practices along with several other more general aspects of quality). These criteria, however, may not be at the level of specificity needed to influence classroom practices that support children's development and learning.

Second, recent research suggests that quality may need to be at a fairly high level before it supports children's development and learning (Zaslow et al., 2016). Some past research using a precursor to the current CLASS Pre-K tool concluded that an Emotional Support score of 5.00 or more and an Instructional Support score of 3.25 or more is needed for pre-K programs to meaningfully contribute to children's social and academic outcomes (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). Using these cutoffs, 21 percent of the programs in this evaluation met that level of high quality, and they were distributed across the star rating levels.[8] It is possible that the quality of preschool classrooms in BrightStars is not at a high enough level to be related to children's development. Similarly, the small number of programs at the upper levels of the rating limited the statistical power needed to detect relationships; it's possible that relationships may be evident in the future if there are more programs at the top rating levels (i.e., at a high-quality level).

Finally, the way that ratings are created in QRIS makes it harder to find relationships with other constructs because the rating is a single variable with a very limited range of scores, and it is easier statistically to find relationships between variables that vary widely (e.g., 1 to 100). In Rhode Island, the BrightStars framework includes information about 23 criteria that is compressed into a value of 1 to 5 for each of 10 standards, and then the information from those 10 standards is used to determine a single program rating with limited variability (i.e., 1 to 5). Thus, the way ratings and standards are constructed, by default, diminishes the chances of finding relationships between ratings and measures of children's skills. In a way, this limitation in ratings highlights the importance of the significant findings between standards and social competence and math: we found a few relationships even when using a standard rating level with a small range.

The BrightStars rating was more strongly related to some children's math skills and social-emotional skills than others. For children from lower income families, there was a substantive, positive association between star rating and social competence and learning behaviors, but not for children from higher income families. This is consistent with some past research and may indicate that high-quality child care provides a buffer against the stress that can accompany poverty or provides positive experiences to children from lower income families (Peisner-Feinberg et al., 2001; Votruba-Drzal, Coley, & Chase-Lansdale, 2004). Surprisingly, we also found that for children from higher income families, math skills were substantively, positively associated with star rating, whereas the same was not true for children from lower income families. Additional research is needed to determine whether this finding can be replicated and, if so, what factors might explain it.

Directors and teachers expressed generally positive views about BrightStars, with a few offering ideas for changes. About 70 percent of the directors in this evaluation rated BrightStars positively or extremely positively. Directors and teachers appreciated the supports and focus on quality. Both directors and teachers also expressed some concerns about the ERS observations, the standards, and the block rating structure.

Study limitations

Like all research, this study had some limitations. The small sample size overall and the very small number of programs at the upper rating levels of 4 and 5 stars decreased the chance we would find meaningful associations and prevented us from comparing groups of programs at different star levels. Although we found evidence that programs with higher star ratings tended to have higher classroom quality, we cannot say with certainty that each star level is associated with higher quality. A related limitation was the modest response rate among programs (29 percent of those eligible for

[8] Programs with preschool quality at this high level were found across rating levels 1-5: two of the three 5-star programs, five of the eleven 4-star programs, two of the fifteen 3-star programs, four of the twenty-one 2-star programs, and two of the twenty-one 1-star programs

the study agreed to participate). Thus, we cannot be certain that this sample represents all programs in BrightStars. Programs that agreed to participate might be systematically different from those that declined, and it is possible that we would have obtained different results if a higher proportion of programs had agreed. Finally, we only focused on center-based programs for the program-level analyses and on preschool-aged children for the child-level analyses. To fully understand the value of BrightStars, additional research on family child care homes, school-aged programs, and the development of infants and toddlers would be necessary.

Considerations for the future

We offer below some suggestions for Rhode Island early childhood leaders to consider as they continue to implement and strengthen BrightStars.

We encourage Rhode Island leaders to use findings from multiple state QRIS evaluations to inform future revisions of the QRIS. Validation studies do not produce a "Yes, QRIS is working" or "No, QRIS is not working" answer but instead provide information to ground discussions about what is working and what could be working better. Because the ELC grant required states to validate their QRIS, we can expect reports from each of the 20 ELC states in the next few years. Although reports from other states should be reviewed carefully to consider which findings are most applicable to Rhode Island's circumstances, Rhode Island leaders should not only use the findings from this evaluation, but also from other QRIS evaluations, to inform future BrightStars revisions. We believe that looking across multiple QRIS evaluations will allow stakeholders across the country to identify general patterns that may collectively inform future QRIS revisions. For example, if multiple evaluations point to a link between QRIS and social competence, but not with early academic skills, that might mean that states interested in boosting early academic skills may need to focus more on supporting specific classroom practices that promote language and math skills. Authors of some recently released QRIS evaluation reports are working to develop a report that synthesizes cross-study findings and their implications as part of the Quality Initiatives Research and Evaluation Consortium (known as INQUIRE). This synthesis, as well as the detailed reports, can inform future BrightStars revisions.

Gathering more program-level data on the BrightStars standards would help leaders better understand the extent to which the framework is working as intended. We recommend that, to the extent possible, data be collected for all BrightStars standards at all rating levels. Currently data are only collected for the level for which the program is applying, and that limits the conclusions that can be drawn. Even if it is not possible to do this for all programs, it might be possible to gather more detailed information from a sample of programs. If data from all criteria were available at all star rating levels, leaders could determine which standards were easier or harder to meet and which standards tend to prevent programs from earning higher star ratings. Furthermore, the additional information could be used to better understand which components of the rating are meaningfully differentiating higher quality programs from lower quality programs.

Rhode Island leaders would benefit from data to ensure that the BrightStars framework is working as intended for family child care homes and school-age programs. This evaluation included only those programs rated with the BrightStars Preschool and Child Care Center Quality Framework. Although we have no reason to believe that the findings would be different, we cannot be certain that the findings from this study generalize to family child care homes or programs serving only school-age children. Additional research is needed to examine how well the framework works in these other settings.

Rhode Island leaders may find it useful to regularly answer particular questions of interest using BrightStars administrative data. BrightStars staff can analyze data that are routinely collected as part of the rating process to help address key questions. This will be easier to do when they have transitioned to using an electronic data system rather than recording rating information on paper. Some possible questions of interest are listed below. We encourage BrightStars leaders to work with stakeholders to refine this list and develop a timeline for periodically analyzing each question. Although many questions may be addressed using solely BrightStars data, other important questions may require additional data (e.g., data from the child care subsidy system).

- What is the distribution of ratings for each program type (centers and preschool programs, family child care, school-age programs) and does it change over time?

- Does the requirement that CCAP programs participate in BrightStars change the percentage of programs participating in CCAP? What percentage of all eligible programs participates in CCAP and how does participation change over time?
- Of the programs that are re-rated using the 2013 framework, what percentage maintains the same rating? Earn a lower rating? Earn a higher rating?
- What percentage of children receiving a child care subsidy is enrolled in a 4- or 5-star program?

We encourage Rhode Island leaders to continue to support improvements in the quality of early care and education for all children. This evaluation focused on the R (rating) in QRIS and provided information about centers and preschool programs and preschoolers' development. We recognize, though, that the I (improvement) in QRIS is equally, or even more, important in ensuring that all children and families who want or need child care can enroll in high-quality programs. As noted above, only one out of every five centers in this study were demonstrating teacher-child interactions, as measured by the CLASS, that are considered high enough quality to influence children's development. This suggests that more work is needed to ensure that children have access to the quality of experiences needed to support their success in school. We encourage Rhode Island leaders to continue to invest its resources to support quality improvement. Directors of programs participating in BrightStars may be able to make some improvements on their own, but they rely on state supports to make many of the needed improvements. They need access to effective trainings and coursework, for example, to meet higher BrightStars standards and improve practice. As one director noted, "Teachers are having a hard time getting into the RIELDS classes." Programs also rely on health care, early intervention, and other settings to offer developmental screenings and services for enrolled children. Rhode Island's early childhood system has benefited from state and federal investments, including most recently the Race to the Top-Early Learning Challenge grant. Sustaining key investments is critical to maintaining and improving the strength of Rhode Island's early care and education system.

References

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647–663. http://doi.org/10.1111/j.1467-8624.2007.01019.x

Burchinal, M., Tarullo, L. & Zaslow, M. (2016). *Best practices in creating and adapting Quality Rating and Improvement System (QRIS) rating scales.* OPRE Research Brief #2016-25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25*, 166–176. doi: 10.1016/j.ecresq.2009.10.004

Hawkinson, L. & Quick, H. (2016, July). *Independent Evaluation of California's RTT-ELC QRIS: Validation Study Results*. Presentation at the QRIS National Meeting, New Orleans, LA.

Hawkinson, L. E., Quick, H. E., Muenchow, S., Anthony, J., Weinberg, E., Holod, A., ... & Karoly, L. A. (2015). *Independent Evaluation of California's Race to the Top – Early Learning Challenge Quality Rating and Improvement System: Half-Term Report.* San Mateo, CA: American Institutes for Research.

Karoly, L. (2014). *Validation Studies for Early Learning and Care Quality Rating and Improvement Systems*. RAND Education and RAND Labor and Population.

Karoly, L. A., Schwartz, H. L., Setodji, C. M., & Haas, A. C. (2016). *Evaluation of Delaware Stars for Early Success: Final Report*. Retrieved from http://www.rand.org/pubs/research_reports/RR1426.html

La Paro, K.M., Hamre, B.K., & Pianta, R.C., (2012). *Classroom Assessment Scoring System - Toddler*. Baltimore, MD: Paul H. Brookes Publishing Co., Inc.

LaFreniere, P. J., & Dumas, J. E. (1996). Social competence and behavior evaluation in children ages 3 to 6 year: the short form (SCBE-30). *Psychological Assessment, 8*(4), 369-377.

Magnuson, K., & Lin, Y. (2016). *Validation of the QRIS YoungStar Rating Scale [Report 2]*. Madison, WI: UW-Madison School of Social Work and Institute for Research on Poverty.

Maxwell, K. (2008). *Pilot test of the draft Rhode Island BrightStars child care center and preschool framework*. Chapel Hill, NC: The University of North Carolina, FPG Child Development Institute.

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the Preschool Learning Behaviors Scale. *Psychology in the Schools, 39,* 353–365.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*(5), 1534–1553.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment Scoring System – Pre-K.* Baltimore, MD: Paul H. Brookes Publishing Co., Inc.

QRIS Compendium. (2016). Data retrieved November 28, 2016 from qriscompendium.org.

Rhode Island Association for the Education of Young Children (RIAEYC). (2016). *Our Mission*. Retrieved from http://www.brightstars.org/about-us/our-mission

Rhode Island Department of Children, Youth & Families, Rhode Island Department of Human Services, & Rhode Island Department of Education (2014). *Quality for every child: Developing a continuum of quality standards for Rhode Island's early care and education programs*. Retrieved from https://exceed.ri.gov/Docs/RI_PSA_TechPaper.pdf

Roberts, J. & Gerena Melia, M. (2016, July). *Massachusetts QRIS Validation Study*. Presentation at the QRIS National Meeting, New Orleans, LA.

Smith-Donald, R., Raver, C.C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly 22*(2): 173-187.

Soderberg, J., Joseph, G. E., Stull, S., & Hassairi, N. (2016). *Early Achievers Standards Validation Study.* Seattle, WA: Childcare Quality and Early Learning Center for Research & Processional Development (CQEL).

Torquati, J. C., Raikes, H. H., Huddleston-Casas, C. A., Bovaird, J. A., & Harris, B. A. (2011). Family income, parent education, and perceived constraints as predictors of observed program quality and parent rated program quality. *Early Childhood Research Quarterly, 26*(4), 453–464.

Tout, K., Cleveland, J., Li, W., Starr, R., Soli, M. & Bultinck, E. (2016). *The Parent Aware Evaluation: Initial Validation Report.* Minneapolis, MN: Child Trends.

United States Census Bureau. (2015). *Poverty thresholds* [Data file]. Retrieved from http://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html

United States Department of Agriculture, Food and Nutrition Service. (2013, September). *National School Lunch Program – Fact Sheet.* Retrieved from http://www.fns.usda.gov/sites/default/files/NSLPFactSheet.pdf

United States Department of Agriculture, Food and Nutrition Service. (2016, April 5). *Women, Infants and Children (WIC): WIC Income Eligibility Guidelines.* Retrieved from http://www.fns.usda.gov/wic/wic-income-eligibility-guidelines

Votruba-drzal, E., Coley, R. L., & Chase-lansdale, P. L. (2004). Child care and low-income children's development: Direct and moderated effects. *Child Development, 75*(1), 296–312.

What Works Clearinghouse. (2014). *Procedures and standards handbook: version 3.0.* US Department of Education. Retrieved from ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement (Third Edition).* Rolling Meadows, IL: Riverside Publishing.

Zaslow, M., Burchinal, M., Tarullo, L., & Martinez-Beck, I. (2016, July). *Quality Thresholds, Features, and Dosage in Early Care and Education: Secondary Data Analyses of Child Outcomes.* In Burchinal, M., Zaslow, M., & Tarullo, L. (Eds.), *Monographs of the Society for Research in Children Development, 81*(2).

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

Zill, N., & Resnick, G. (1998). *Color Names. Head Start Family and Child Experiences Survey*. Rockville, MD: Westat.

Appendix A. Detailed Description of the Study Design and Procedures

Recruitment of programs

As noted in the introduction of this report, Rhode Island leaders made two major changes to BrightStars as part of the state's Early Learning Challenge grant: 1) the BrightStars framework was revised in 2013 and 2) BrightStars moved from a voluntary to mandatory system as of April 2014. These two changes had important implications for the timing and design of the evaluation. Because BrightStars re-rates programs every 3 years, when this study began in the fall of 2015, not all programs were rated using the 2013 revised framework; some still had ratings based on the previous version. To appropriately evaluate the 2013 revision, we limited the study sample to only those programs that had a rating based on the 2013 framework. We also used a second criterion to limit the sample. In the transition from a voluntary to mandatory system, many programs were automatically assigned to a 1-star. Those programs were likely a mixture of lower quality and higher quality programs that, for various reasons, chose not to apply for a higher star rating during the transition process. Because we were concerned about the possible variability in quality among 1-star programs, we limited study participation to only those programs that had completed a full rating process and earned a 1-star rating. We postponed the evaluation a year to give BrightStars time to rate as many programs as possible on the 2013 framework and, in general, to give BrightStars administrators and programs time to *settle* into the new system.

The person we contacted for recruitment efforts was the same person who submitted a BrightStars application on behalf of that program. In most cases, this was the director or owner for child care centers or an administrator if the program was school-based.

The original list obtained from the Rhode Island Department of Human Services included 231 programs; however five programs were ineligible because they had closed, were suspended from BrightStars, served only children with special needs, or served only infants/toddlers. Thus, the sample we attempted to recruit for the full study included 226 programs.

Recruitment for the full study took place over a four-month period in the summer and fall of 2015. Child Trends recruiters attempted to contact all 226 programs (167 one- and two- star, and 59 three-, four -, and five- star) with recruitment targets of 50 one- and two- star programs and 50 three-, four-, and five-star programs. The goal of 50 one- and two-star programs was met in early October 2015, and recruitment of those star levels stopped later that month. At that point, initial efforts had been made to contact all one- and two-programs; those that are not in the sample are a mix of programs that actively declined and programs that were never reached. The target of 50 three-, four-, and five-star programs was not met due to the small number of programs in the recruitment pool. Therefore, we continued to recruit additional 3-, 4-, and 5-star programs through January 2016 using the original contact list, plus 16 additional programs that were newly rated since the summer, bringing the total number of programs we attempted to recruit over the course of the study to 242. The programs recruited in the winter took part in the classroom observation portion of the study only.

Child Trends initiated contact with programs by mailing and emailing them a letter that provided information about the evaluation and attached a letter from state administrators encouraging the programs to participate. A team of recruiters followed up with programs by phone and by email. During phone calls with center directors, the recruitment team explained the purpose and activities of the study and obtained verbal consent. If there was more than one 4-year-old classroom at the center, the recruiter randomly selected a 4-year-old classroom to participate during the recruitment call. In programs where all of the preschool-aged classrooms served mixed ages, preference was given to the classroom that served the most children who would be entering kindergarten the next year.

Out of 167 total one- and two-star eligible programs, 42 participated (25 percent). Out of 75 three-, four-, and five-star programs, 29 participated (39 percent). In total, 71 programs (29 percent) took part in the study; 65 for the full study and 6 for the winter only. Figure 1 shows the distribution of programs in the final sample by star rating: 21 one-star, 21 two-star, 15 three-star, 11 four-star, and 3 five-star.

Recruitment of children and families

The selected preschool classroom's teacher and the center director were asked to distribute and collect consent packets to parents of all children in the selected classroom. The packet included a permission form and a brief demographic questionnaire. All children with parent permission who were at least 3 years of age and attended the program at least 9 hours per week were eligible.

Up to eight children in each classroom were selected to participate from those with permission. If more than eight had permission, eight were selected at random, giving preference to 4-year-olds and children who had complete information on the brief demographic questionnaire.

All assessments were administered in English. In the four instances when a parent indicated on the demographic questionnaire that a child's English language skills were *poor*, the data collector spoke with the teacher. In all cases the teacher confirmed that the child could be assessed in English, and the child was assessed.

If at least four forms were not returned when the data collector called to schedule the assessment, she would remind the director or teacher to collect the forms from parents. If fewer than four consent forms were obtained from a particular classroom, all of the children with permission were assessed towards the end of the fielding window.

In the 65 preschool classrooms, teachers reported that there were a total of 959 children enrolled. Overall, 497 (52 percent) parents gave permission for their child to participate in the study (an average of 8 per classroom, with a range of 3 to 17). The final sample includes 332 children with both fall and spring data. Within classrooms, the average number of participating children was five (range two to eight).

Information collected

Child assessments

All participating children were given a brief set of assessments in the fall and spring, designed to test their expressive language, pre-literacy, math, and executive function skills. The assessments are described below, in the same order in which they were given.

The Color Bears Task measures children's ability to identify 10 basic colors. In this task, children are first asked to name the colors of 10 colored bears. If the child cannot name the color, he or she is later asked to find the bear of that color. This task is drawn from the Head Start Family and Child Experiences Survey (Zill & Resnick, 1998) and was used as an ice breaker to familiarize children with the assessment process.

The Woodcock Johnson Tests of Achievement, 3rd edition (WJ-III; Woodcock, McGrew, & Mather, 2001) is an individually-administered nationally-normed measure of academic achievement validated for ages 2 to 90 years. Three of the twelve subtests recommended for preschoolers were used in this study: Picture Vocabulary, Letter-Word Identification, and Applied Problems. Each subscale produces nationally-normed standard scores that are adjusted for children's age. These subtests demonstrate high internal reliability and acceptable validity and are widely used in national studies.

The Picture Vocabulary subtest measures children's expressive language by asking them to name a series of increasingly complex images such as a bird, a clock, and a fire extinguisher. The Letter-Word Identification subtest assesses children's pre-literacy skills by asking children to identify first single letters, then progressively more complex words. On the Applied Problems subtest, children are asked to complete math-related tasks such as showing two fingers, counting objects, and adding or subtracting small numbers.

The Pencil Tap Task (Smith-Donald, Raver, Hayes, & Richardson, 2007) measures executive function, or the child's attentional skills that help him/her focus, remember, and plan. Executive function is an important component of school readiness that has been linked to later academic skills (Blair & Razza, 2007). In the Pencil Tap Task, children are instructed to do the opposite of what the assessor does. When the assessor taps her pencil one time, the child is told to tap his/her pencil two times, and vice versa. After a series of practice trials, 16 scored trials are administered. The percent of correct responses is used as an index of executive function and has good concurrent and construct validity (Smith-Donald et al., 2007).

Child assessment visits primarily occurred in the morning. Data collectors typically conducted assessments in a private space outside of the main classroom to minimize distraction. The full assessment battery lasted about 20 minutes. Children were given a book as a thank you for their participation.

Assessments were administered to 387 children in the fall between the months of September and November. Spring assessments were administered to 332 of the original 387 children between the months of April and June. Children were assessed in the spring an average of 6.6 months (range 5.8 to 7.5 months) after their fall assessment. The 55 children who took part in the fall assessment but not the spring were no longer enrolled in that program.

Teacher report of social-emotional and behavior skills

In the fall and spring, at roughly the same time as the child assessments, teachers were asked to report on children's social-emotional and behavior skills, using the standardized tools, described below.

The Social Competence and Behavior Evaluation short form (SCBE-30; LaFreniere & Dumas, 1996) is designed for children between the ages of 2.5 and 6 years old. It includes 30 questions and measures three components of children's social skills: Social Competence (emotionally mature, pro-social behaviors), Anger Aggression (oppositional behaviors, poor frustration tolerance), and Anxiety Withdrawal (anxious, depressed). Each subscale consists of 10 items rated on a 6-point scale indicating how often the child engages in the behavior, ranging from *never* to *always*. Scores are created by summing the 10 items, creating subscales with 60 possible points. Higher scores indicate higher levels of the behavior, so higher scores are desirable on the Social Competence subscale; whereas lower scores are desirable on Anger Aggression and Anxiety Withdrawal. The SCBE-30 has high test-retest reliability, internal consistency, and construct validity (LaFreniere & Dumas, 1996).

The Preschool Learning Behavior Scale (PLBS; McDermott, Leigh, & Perry, 2002) is a 29-item nationally standardized, teacher-completed rating scale of readily observable learning behaviors, designed for children between the ages of 3 and 5.5 years old. Teachers rate children's behavior on a Likert scale, with response options of *most often applies, sometimes applies,* or *doesn't apply*. The total score is the sum of all items and reflects the child's motivation, persistence, and attitude toward learning. Higher scores on this scale are more desirable.

For this study, data collectors prepared a one-page questionnaire for each child in the study and handed these out to teachers during the assessment visit. The questionnaires included the full SCBE-30 and PLBS. Teachers were asked to complete the questionnaires and return them via mail. Teacher reports were completed for 361 (93 percent of the 387) children in the fall and 320 (96 percent of the 332) children in the spring.

Classroom observation measures

The Classroom Assessment Scoring System Pre-K (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008) is an observational tool that assesses the quality of the interactions between children and teachers in preschool classrooms (ages 3 to 5 years). The CLASS Pre-K has 10 subscales that fall under 3 domains: Emotional Support, Classroom Organization, and Instructional Support. Each subscale is scored on a 7-point scale with scores of 1 and 2 considered low quality, 3 to 5 considered mid-range quality, and 6 and 7 considered high quality. Observers rate the classrooms and teachers on each of the dimensions every 30 minutes throughout the observation morning.

The Classroom Assessment Scoring System Toddler (CLASS Toddler; La Paro, Hamre, & Pianta, 2012) is an observational tool that assesses the quality of the interactions between children and teachers in toddler classrooms (ages 15 to 36 months). The CLASS Toddler has eight dimensions that fall under two domains: Emotional and Behavioral Support and Engaged Support for Learning. The scoring for the CLASS Toddler matches that of CLASS Pre-K.

For this study, data collectors worked with directors or classroom teachers to schedule their observations for a typical day (i.e., one without any unusual activities or circumstances). Observations generally began at 8:30 a.m.; if the center opened after 8:30 a.m. then the data collector would arrive 15 minutes after the center opened. Data collectors observed the classroom for 20 minutes at a time and each 20-minute period was followed by a 10-minute coding session. Depending on the classroom's daily schedule, either four or five 30-minute observation cycles were conducted for each classroom that was observed. Our team completed a total of 71 CLASS Pre-K observations as part of this study. The majority of these observations (n = 67) were conducted in

January, February, and March. Four additional observations were conducted in April or May due to lead teacher absences. CLASS Toddler observations (n = 32) were conducted in May, June, and July. For both the preschool and toddler versions, scores are derived by first averaging together the values on each dimension, across the cycles. Next, the dimensions are averaged together to create the domain scores.

Questionnaires

We developed questionnaires for the following groups: parents, teachers, and directors. The parent questionnaire included 11 items focusing on child and family demographics including: child gender, date of birth, parental education, and household income and size. Teachers distributed these brief questionnaires to parents in the consent packets. Parent questionnaires were returned for 332 children who participated in the fall and spring assessment and are in the current analyses.

The teacher questionnaire had demographic items such as highest level of education, major area of study, and years of experience. Teachers were also asked about their opinion of the BrightStars rating system. Data collectors delivered paper copies of the questionnaire to teachers at the time of the classroom observation and teachers were instructed to return the questionnaires by mail. Additional follow-up with teachers was conducted via phone and e-mail to ensure completion of the questionnaires. Sixty-eight (96 percent of 71) preschool teachers and 31 (97 percent of 32) toddler teachers completed the questionnaire.

The director questionnaire collected the same demographic information as the teacher questionnaire. Directors were also asked a series of questions about their program including total capacity, enrollment numbers by age group, teacher turnover, and whether or not the center accepts children receiving child care subsidies. The last set of questions focused on participation in BrightStars, including why the director decided to participate, what supports the center used, and the director's overall impressions of the BrightStars rating system. Director questionnaires were delivered in-person at the time of the classroom observation. Seventy (99 percent of 71) directors completed the questionnaire; one opted out.

Administrative data collection

The research team obtained information about each programs' star rating and their score on each of the 10 standard from the Rhode Island Association for the Education of Young Children. When programs enroll in BrightStars, they decide which star level they are seeking and provide evidence that their program meets the criteria for that star rating. The criteria are organized into 10 standards, and the star rating is equal to the standard on which they score the lowest. This is often referred to as a *block* QRIS.

One goal of this evaluation is to understand how the various standards do or do not work together and if there are components of the star rating that are more or less strongly related to classroom quality and children's outcomes. With a block rating structure, though, programs do not necessarily submit evidence on each criterion at every level. They instead may submit evidence only for the rating level that they expect to be able to reach. For instance, if a program knows that its teachers' credentials only meet the 3-star level, they may choose not to submit evidence that their curriculum meets the 4-star level because they know that they cannot attain an overall 4-star rating. Although we have analyzed the available administrative data regarding participating programs' ratings for each standard, it is possible that programs met higher levels for particular standards but did not submit evidence because they were not required to do so.

Staff hiring and training

In preparation for a year-long fielding effort, we hired three part-time data collectors in August 2015. We selected data collectors based on their experience conducting similar research or following standardized protocols, their experience working with children ages 3 to 5, and their communication skills. Data collectors were trained on the protection of human subjects, the child abuse and neglect protocol, the child assessment battery, and other data collection procedures such as scheduling and data entry during a two-day training conducted in Rhode Island by an experienced member of the project team.

Following the training, data collectors practiced the assessment battery with other adults and young children. Prior to entering the field for the first time, each data collector videotaped herself conducting the child assessment battery with a child between the ages of 3 to 5. These videos were

reviewed by the trainer and data collectors were given feedback prior to entering the field. During the first week of child assessment visits, the Child Trends trainer traveled to Rhode Island and observed data collectors in the field to ensure they were conducting the child assessments correctly and following all procedures. Each week, data collectors mailed assessment data to the main Child Trends office and research assistants reviewed each assessment to ensure they were administered accurately.

In December 2015, the data collection team was trained to conduct classroom observations by the same team member who facilitated the fall training. This team member is a certified Teachstone[9] affiliate trainer and the training adhered to Teachstone's guidelines. During the weeks following the training, all three data collectors achieved reliability on the CLASS Pre-K in alignment with Teachstone's requirements. Specifically, each data collector watched and coded five videos online, scoring at least 80 percent within one point of the master codes across all codes on the five videos and coding each dimension within one point of the master code in at least two of the five videos. One week prior to entering the field, the entire team coded a calibration video accessed through the Teachstone website. Each data collector's scores were within one point of the master codes 90 percent of the time. Data collectors conducted their first two observations in pairs, either with another member of the team or with the Child Trends trainer. Following each observation, the observers discussed their scores on each dimension and arrived at consensus for all of the scores. To ensure ongoing interrater reliability on the CLASS, a second calibration video was coded by the team in mid-February; each data collector's scores were within one point of the master codes 85 percent of the time, on average. Observations were also coded in pairs throughout the fielding window. During nine double-coded observations, the team's scores were within one point of one another 91 percent of the time, on average.

In April 2016, we hired additional data collectors specifically for classroom observations in toddler classrooms. Because of the short time frame for this data collection effort, we exclusively hired data collectors who were already reliable on the CLASS Pre-K; one of the three had extensive experience conducting CLASS Toddler observations. The other two data collectors attended a two-day training in April 2016 conducted by the same team member who led the previous trainings. The two-day training followed Teachstone's guidelines for CLASS Toddler training and the trainer was certified by Teachstone. Following the training, the two data collectors achieved reliability on the CLASS Toddler following Teachstone's process. The process for maintaining interrater reliability on CLASS Toddler was similar to that for CLASS Pre-K. The team conducted observations in pairs for their first few visits. The two team members who conducted the majority of the observations (94 percent) were 96 percent reliable with each other, on average, across three double-coded observations.

Data entry and verification

All data were doubled entered to ensure correctness. When both rounds of data entry were complete and the files were in equivalent formats, a Child Trends staff member used computer software to find inconsistencies between the two files. All inconsistencies were then checked against the hard copies and corrected to create a final data file for each source. The last step of the process was to spot-check 10 percent of each data file to ensure accuracy with the hard copies.

Data available for analysis

This report includes all director, teacher and parent questionnaire data, CLASS observations, and child assessment data. Child assessment and parent questionnaire data are included for children who participated in both the fall and spring assessments. Table 3 shows the sample sizes for each type of data.

[9] Teachstone is the company that oversees training and reliability testing on the CLASS tools.

Appendix B. Fall and Spring Child Assessments

Table B-1. Fall and spring child assessments

	Fall					Spring					Fall vs. spring	
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	Gain	p-value
WJ: Applied Problems	384	107.42	12.94	68	139	329	107.56	10.76	62	134	0.14	n.s.
WJ: Letter Word	387	101.71	12.43	67	186	331	102.45	11.54	74	176	0.74	n.s.
WJ: Picture Vocabulary	387	104.47	10.09	71	149	331	104.26	9.79	77	145	-0.21	n.s.
Pencil Tapping Test	374	9.71	5.02	0	16	329	12.41	4.24	0	16	2.70	<0.001
SCBE: Social Competence	361	39.00	10.16	14	60	314	41.00	9.81	17	60	2.00	<0.01
SCBE: Anxiety-Withdrawal	361	16.32	6.72	10	48	315	15.71	6.22	9	46	-0.61	n.s.
SCBE: Anger-Aggression	360	18.24	9.60	10	58	313	17.76	9.09	10	59	-0.48	n.s
PLBS: Learning Behaviors	362	50.37	9.21	7	61.67	319	52.29	8.33	22.33	61.67	1.92	<0.01

Source: Child Trends child assessment data collection (2015-2016).

Appendix C. Descriptive Information and Associations among Star Ratings and Standards

As seen on Table C-1, the average star rating is 2.35. As noted earlier, the star rating is based on a block rating structure. Programs apply for a specific star rating and are required to submit evidence that they meet all 10 standards at the level for which they are applying. If a program exceeds the star rating for which they are applying on some standards, they may or may not submit evidence supporting that higher rating on those standards. For instance, if a program knows that its teachers' credentials only meet the 3-star level, they may choose not to submit evidence that their curriculum meets the 4-star level because they know that they cannot attain an overall 4-star rating.

In general, programs in the study scored highest on staff-child ratios and maximum group sizes. This is not surprising because the ratio and group size standards are rated as either a 1 or 5 due to the way BrightStars is constructed. The learning environment, which is mostly measured by the ERS scores, was the standard on which they typically scored lowest. Table C-1 also provides descriptive information for the ECERS-R and ITERS-R scores, when available. On average, these programs score in the mid-range, between minimal and good quality as defined by the scale authors.

Table C-1. Descriptive information for star rating and standards

Variable	N	Mean	SD	Min	Max
Overall star rating	71	2.35	1.18	1	5
S1: Learning environment	71	2.59	1.14	1	5
S2: Minimum staff-child ratio	71	3.87	1.81	1	5
S3: Maximum group size	70	3.80	1.85	1	5
S4: Teacher qualifications	71	2.77	1.49	1	5
S5: Program leadership	71	3.23	1.57	1	5
S6: Continuous quality improvement	71	3.21	1.57	1	5
S7: Curriculum	71	3.13	1.33	1	5
S8: Child assessment	71	3.00	1.41	1	5
S9: Inclusive classroom practices	71	3.52	1.57	1	5
S10: Family communication and involvement	71	3.08	1.57	1	5
Average ECERS-R Score	31	3.94	0.71	2.81	5.33
Average ITERS-R Score	16	3.52	0.98	2.07	4.94

Source: BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).

Table C-2 presents the correlations among the standards and the star rating. They are all significantly correlated with one another, and all the correlations are medium or large. Coupled with the high Cronbach's alpha (.97), these correlations tend to indicate that a single underlying construct is being measured.

Table C-2. Correlations among standards and star ratings

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1. Learning environment	1.00									
S2. Minimum staff-child ratio	0.58	1.00								
S3. Maximum group size	0.58	0.97	1.00							
S4. Teacher qualifications	0.79	0.75	0.75	1.00						
S5. Program leadership	0.83	0.67	0.63	0.84	1.00					
S6. Continuous quality improvement	0.74	0.87	0.86	0.84	0.80	1.00				
S7. Curriculum	0.78	0.75	0.71	0.86	0.85	0.82	1.00			
S8. Child assessment	0.78	0.80	0.79	0.86	0.79	0.85	0.82	1.00		
S9. Inclusive classroom practices	0.70	0.79	0.77	0.83	0.83	0.86	0.88	0.79	1.00	
S10. Family communication and involvement	0.74	0.82	0.80	0.88	0.86	0.90	0.88	0.89	0.93	1.00
Star Rating	0.92	0.69	0.72	0.90	0.82	0.81	0.81	0.86	0.75	0.84

Source: BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).
Note: n = 71 for all variables except S3. n for S3 is 70; all correlation coefficients are at significant level of $p < .001$

Appendix D. Descriptive Information About CLASS and Full Classroom-Level Regression Tables

Table D-1 provides descriptive information about the CLASS Pre-K and CLASS Toddler. On average, preschool classrooms scored at the high end of the mid-range on Emotional Support and Classroom Organization and in the low range on Instructional Support. This pattern of Instructional Support being much lower than Emotional Support and Classroom Organization is typical for this tool and was reported in other QRIS evaluations (Tout et al., 2016; Hawkinson et al., 2015).

Table D-1. Descriptive information for the CLASS Pre-K and CLASS Toddler

Variable	N	Mean	SD	Min	Max
CLASS Pre-K Emotional Support	71	5.54	0.76	3.20	6.75
CLASS Pre-K Classroom Organization	71	5.30	0.71	3.47	6.60
CLASS Pre-K Instructional Support	71	2.73	0.85	1.40	5.27
CLASS Toddler Emotional and Behavioral Support	32	5.18	0.79	3.72	6.52
CLASS Toddler Engaged Support for Learning	32	3.22	1.04	1.67	5.00

Source: Child Trends classroom observation data collection (2016).

Table D-2 provides coefficients, standard errors and effect sizes for all CLASS Pre-K models. For this study, effect size (ES) indicates the difference in outcome (CLASS, in this case) associated with a one star difference in rating, in standard deviations. So, the ES of 0.33 for Instructional Support indicates that each star is associated with a 0.33 standard deviation difference in Instructional Support. For example, Instructional Support in a 2-star center is 0.33 standard deviations higher than in a 1-star center, and Instructional support in a 3-star center is 0.33 standard deviations higher than in a 2-star and 0.66 standard deviations higher than in a 1-star center.

We are using a cutoff of $p < .10$ to determine statistical significance because the sample size is relatively small. To determine which effect sizes were large enough to warrant interpretation, we consulted the What Works Clearinghouse (WWC, 2014), an initiative of the U.S. Department of Education that reviews and assesses educational research evidence. WWC refers to an effect size of 0.25 or higher when comparing two groups as *substantively important*. The effect sizes we calculated in this evaluation represent the amount of difference (in standard deviations) between programs that differ by one star (e.g., 1- vs 2-star or 3- vs 4-star). We realize, though, that the lowest and highest (1- vs. 5-star) groups are most likely to produce meaningfully different outcomes in children. The highest and lowest are four star levels apart. Thus, we decided to interpret our effect sizes of .07 or greater as *substantively important* because when multiplied by four, those effect sizes exceed .25. An effect size of .06 falls below .25 when multiplied by four.

The first row on table D-2 shows the association between the star rating and the three CLASS domains. The remaining rows represent the associations between each of the BrightStars standards and the three CLASS domains. Each row represents three separate regression analyses, one for Emotional Support, one for Classroom Organization and one for Instructional Support. All of the significant findings have effect sizes of .07 or above. Findings indicate that learning environment, teacher qualifications, program leadership, curriculum, inclusive classroom practices, and family communication and involvement are associated with all three CLASS domains. Additionally, ratios, groups size, continuous quality improvement, and child assessment are related to Instructional Support.

Table D-2. Star ratings and standards as predictors of CLASS Pre-K (n = 71)

	Emotional Support			Classroom Organization			Instructional Support		
	Coeff.	SE	ES	Coeff.	SE	ES	Coeff.	SE	ES
Star rating	0.15+	0.07	0.19	0.14+	0.07	0.19	0.28***	0.08	0.33
S1. Learning environment	0.18*	0.08	0.23	0.16*	0.07	0.23	0.27**	0.08	0.32
S2. Minimum staff-child ratio	0.04	0.05	0.05	0.05	0.05	0.08	0.11*	0.05	0.13
S3. Maximum group size	0.03	0.05	0.04	0.05	0.05	0.06	0.12*	0.05	0.14
S4. Teacher qualifications	0.14*	0.06	0.19	0.12*	0.05	0.17	0.22***	0.06	0.26
S5. Program leadership	0.15*	0.06	0.19	0.12*	0.05	0.17	0.19**	0.06	0.23
S6. Continuous quality improvement	0.08	0.06	0.10	0.08	0.05	0.11	0.17**	0.06	0.20
S7. Curriculum	0.14*	0.07	0.18	0.13*	0.06	0.18	0.21**	0.07	0.25
S8. Child assessment	0.09	0.06	0.12	0.08	0.06	0.11	0.20**	0.07	0.23
S9. Inclusive classroom practices	0.13*	0.06	0.17	0.10+	0.05	0.14	0.20**	0.06	0.24
S10. Family communication and involvement	0.10+	0.06	0.14	0.09+	0.05	0.13	0.20**	0.06	0.24

Source: Child Trends classroom observation data and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).
+ $p < .10$; * $p < .05$; ** , $p < .01$; *** $p < .001$. Notes: Each row represents three regression analyses—one per outcome. Effect sizes represent the number of standard deviations of change on the outcome associated with a one-unit (star or point) change on the predictor variable.

Table D-3 presents similar information as Table D-2, but for the CLASS Toddler. Programs with higher star ratings had higher scores on Emotional and Behavioral Support as well as on Engaged Support

for Learning. Further, all 10 standards are associated with Engaged Support for Learning and all except ratios and group size are significantly associated with Emotional and Behavioral Support.

Table D-3. Star ratings and standards as predictors of CLASS Toddler (n = 32)

	Emotional and Behavioral Support			Engaged Support for Learning		
	Coeff.	SE	ES	Coeff.	SE	ES
Star Rating	0.41**	0.12	0.52	0.56**	0.16	0.54
S1. Learning environment	0.42***	0.11	0.53	0.57***	0.15	0.55
S2. Minimum staff-child ratio	0.11	0.07	0.14	0.20*	0.09	0.19
S3. Maximum group size	0.11	0.07	0.13	0.19+	0.09	0.18
S4. Teacher qualifications	0.29*	0.09	0.37	0.39**	0.12	0.37
S5. Program leadership	0.23**	0.08	0.29	0.35**	0.10	0.33
S6. Continuous quality improvement	0.18*	0.08	0.22	0.25*	0.10	0.24
S7. Curriculum	0.25*	0.10	0.32	0.34*	0.13	0.32
S8. Child assessment	0.29**	0.09	0.37	0.39**	0.11	0.37
S9. Inclusive classroom practices	0.21*	0.08	0.26	0.27*	0.11	0.26
S10. Family communication and involvement	0.21*	0.08	0.26	0.30**	0.11	0.29

Source: Child Trends classroom observation data and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2016).
+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$. Notes: Each row represents two regression analyses—one per outcome. Effect sizes represent the number of standard deviations of change on the outcome associated with a one-unit (star or point) change on the predictor variable.

Appendix E. Star Rating and Standards as Predictors of Children's Skills

Table E-1 shows the results for the hierarchical linear models using star rating and standards to predict children's early academic skills and executive function in the spring. Table E-2 shows the results for the HLMs using star rating and standards to predict teachers' reports of children's social-emotional skills in the spring. These models control the fall score on the same measure, as well as family income (above or below 185 percent of the federal poverty line for family size). Each row represents a separate set of models (four models per row for table E-1 and four models per row for table E-2).

As with the CLASS models, we focus our interpretation on effects with a p value less than .10 and an effect size of .07 or higher. See Appendix D for the rationale for these cutoffs. Using these cutoffs, there are few substantive associations. Star rating is not associated with any of the child assessment outcomes. Several of the individual standards (teacher qualifications, program leadership, curriculum, inclusive classroom practices, and family communication and involvement) are associated with social competence. The curriculum standard is related to math skills.

As noted in table E-1, there are several significant, negative associations with expressive language. These indicated that children in higher rated programs scored lower in the spring on the WJ Picture Vocabulary subtest; however the effect sizes are very small so we are choosing not focusing on them in our discussion.

Table E-1. Star ratings and standards as predictors of children's early academic and executive function skills

	Math (WJ: Applied Problems; n = 329)			Pre-Literacy (WJ: Letter-Word; n = 331)			Expressive Language (WJ: Picture Vocabulary; n = 331)			Executive Function (Pencil Tap; n = 329)		
	Coeff.	SE	ES	Coeff.	SE	ES	Coeff.	SE	ES	Coeff.	SE	ES
Star rating	0.67	0.42	0.05	-0.12	0.37	-0.01	-0.54+	0.30	-0.05	0.28	0.19	0.06
S1. Learning environment	0.69	0.43	0.05	0.05	0.39	0.00	-0.44	0.31	-0.04	0.07	0.20	0.02
S2. Minimum staff-child ratio	0.36	0.28	0.03	-0.33	0.24	-0.03	-0.47*	0.20	-0.05	0.06	0.13	0.01
S3. Maximum group size	0.46	0.28	0.04	-0.34	0.24	-0.03	-0.39+	0.20	-0.04	0.06	0.13	0.01
S4. Teacher qualifications	0.64+	0.35	0.05	-0.21	0.31	-0.02	-0.52*	0.25	-0.05	0.20	0.15	0.05
S5. Program leadership	0.48	0.33	0.04	-0.02	0.29	0.00	-0.40+	0.24	-0.04	0.06	0.15	0.01
S6. Continuous quality improvement	0.63+	0.32	0.05	-0.20	0.29	-0.02	-0.44+	0.24	-0.04	0.13	0.15	0.03
S7. Curriculum	1.03*	0.39	0.08	-0.28	0.35	-0.02	-0.62*	0.29	-0.06	0.18	0.18	0.04
S8. Child assessment	0.48	0.38	0.04	-0.27	0.33	-0.02	-0.59*	0.27	-0.06	0.13	0.17	0.03
S9. Inclusive classroom practices	0.68*	0.33	0.05	-0.38	0.29	-0.03	-0.50*	0.24	-0.05	0.15	0.15	0.03
S10. Family communication and involvement	0.59+	0.33	0.05	-0.36	0.29	-0.03	-0.50*	0.24	-0.05	0.25+	0.15	0.06

Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).

+ $p < .10$; * $p < .05$; ** , $p < .01$; *** $p < .001$. Notes: Each row represents four HLM analyses—one per outcome. These analyses control for fall score on the same measure and family poverty (above or below 185 percent of poverty of the family size). Effect sizes represent the number of standard deviations of change on the outcome associated with a one-unit (star or point) change on the predictor variable.

Table E-2. Star ratings and standards as predictors of children's early social-emotional skills

	Social Competence (SCBE; n = 314)			Anxiety (SCBE; n = 315)			Aggressiveness (SCBE; n = 313)			Learning Behaviors (PLBS; n = 319)		
	Coeff.	SE	ES	Coeff.	SE	ES	Coeff.	SE	ES	Coeff.	SE	ES
Star rating	0.58	0.52	0.06	-0.03	0.31	-0.01	0.12	0.34	0.01	-0.07	0.35	-0.01
S1. Learning environment	0.76	0.53	0.08	-0.08	0.32	-0.01	0.09	0.36	0.01	0.16	0.36	0.02
S2. Minimum staff-child ratio	0.42	0.34	0.04	-0.26	0.20	-0.04	-0.36	0.23	-0.04	0.03	0.23	0.00
S3. Maximum group size	0.33	0.34	0.03	-0.27	0.19	-0.04	-0.31	0.23	-0.03	0.02	0.23	0.00
S4. Teacher qualifications	0.84*	0.42	0.09	-0.09	0.26	-0.01	0.04	0.29	0.00	0.04	0.29	0.00
S5. Program leadership	0.87*	0.40	0.09	-0.13	0.24	-0.02	-0.06	0.28	-0.01	0.14	0.28	0.02
S6. Continuous quality improvement	0.56	0.40	0.06	-0.07	0.24	-0.01	-0.21	0.27	-0.02	0.10	0.28	0.01
S7. Curriculum	1.33**	0.47	0.14	-0.30	0.29	-0.05	-0.35	0.33	-0.04	0.47	0.33	0.06
S8. Child assessment	0.73	0.45	0.07	-0.01	0.27	0.00	-0.17	0.31	-0.02	0.07	0.31	0.01
S9. Inclusive classroom practices	0.84*	0.40	0.09	-0.31	0.24	-0.05	-0.36	0.27	-0.04	0.39	0.28	0.05
S10. Family communication and involvement	0.73+	0.41	0.07	-0.16	0.25	-0.03	-0.39	0.28	-0.04	0.19	0.28	0.02

Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).

+ $p < .10$; * $p < .05$; ** , $p < .01$; *** $p < .001$. Notes: Each row represents four HLM analyses– one per outcome. These analyses control for fall score on the same measure and family poverty (above or below 185% of poverty of the family size). Effect sizes represent the number of standard deviations of change on the outcome associated with a one-unit (star or point) change on the predictor variable.

Appendix F. Interactions

Table F-1. Interaction of star sating and family income as predictor of spring early academic, executive function and social-emotional skills

	Star rating	Family income	Rating X family income interaction	ES for lower income	ES for higher income
Math (WJ Applied Problems)	-0.27	3.65	-1.43+	-0.02	0.11
Pre-Literacy (WJ: Letter-Word)	0.21	-3.02	0.49		
Expressive Language (WJ: Picture Vocabulary)	-0.32	-2.84+	0.34		
Executive Function (Pencil Tap)	0.11	-0.19	-0.26		
Social Competence (SCBE)	1.72*	-6.04**	1.80*	0.18	-0.01
Anxiety (SCBE)	0.07	-0.02	0.16		
Aggressiveness (SCBE)	-0.51	2.64	-0.98		
Learning Behaviors (PLBS)	0.63	-4.29*	1.10+	0.08	-0.05

Source: Child Trends parent questionnaire, Child Trends child assessment data, and BrightStars administrative data from the Rhode Island Association for the Education of Young Children (2015-2016).
+ $p < .10$; * $p < .05$; ** , $p < .01$; *** $p < .001$. Notes: Each row represents a single hierarchical linear model. These analyses control for fall score on the same. Effect sizes represent the number of standard deviations of change on the outcome associated with a one star change on the predictor variable. For family income, 0 = higher income (greater than 185 percent of poverty for family size), 1 = lower income (less than or equal to 185 percent of poverty for family size).