Child TRENDS

Quality Rated Validation Study Report #4:

Quality Rated Star Ratings and Independent Measures of Quality, Children's Growth, and Work Climate

Diane M. Early, Kelly L. Maxwell, Amy Blasberg, Brenda Miranda, Nadia S. Orfali, Gary E. Bingham, Rihana S. Mason, Weilin Li, Erin Bultinck, & Tracy Gebhart

May 2019

Author Affiliations

Diane M. Early, Kelly L. Maxwell, Amy Blasberg, Brenda Miranda, Nadia S. Orfali, Weilin Li, Erin Bultinck, & Tracy Gebhart: Child Trends

Rihana S. Mason and Gary E. Bingham: The Urban Child Study Center at Georgia State University

Suggested Citation

Early, D. M., Maxwell, K. L., Blasberg, A., Miranda, B., Orfali, N. S., Bingham, G. E., ... , Gebhart, T. (2019). Quality Rated Validation Study Report #4: Quality Rated Star Ratings and Independent Measures of Quality, Children's Growth, and Work Climate. Bethesda, MD: Child Trends.

Acknowledgments

We would like to thank Commissioner Amy Jacobs, Kristin Bernhard, Catherine Broussard, Jessie Bruno, Randy Hudgins, Denise Jenson, Rob O'Callaghan, Bentley Ponder, Pam Stevens, Shayna Funke, Ani Whitmore, Alexandria Williams, and Sonja Steptoe from the Georgia Department of Early Care and Learning and Polina Aleshina and Nnenna Ogbu from the Governor's Office of Student Achievement for their support of this work. We would also like to thank Kyong-ah Kwon and Nicole Patton-Terry, colleagues at Georgia State University who contributed to the conceptualization of this study during the pilot phase. We sincerely appreciate the thoughtful feedback we received from our reviewers, Kevin Fortner of Georgia State University and Kathryn Tout at Child Trends. We are grateful for the commitment and enthusiasm of our data collection team, many of whom dedicated themselves to this study for two years: Rebecca Barria, Allyson Reyer Beaver, Patricia Cambron, Courtney Carter, Elizabeth Crofton, Nicole Venuto Gearing, Vergina Glymph, Kristyn Goodwin, Maria Guzman, Khadija Highsmith, Sharon Hudgins-Beck, Melissa Johns, Kimberly Leidinger, Chaehyun Lim, Joyanna Malutinok, Michelle Matthews, Nyeelah Matthews, Elizabeth Milling, Jacquelyn Parker, Page Carriveau Pattermann, Lidia Quinones, Muminah Rashid, Janice Royals, Melissa Schellenberg, Melissa Silva, Shanett Smiley, Moneesha Smith, Francheska Denise Starks, Alissa Sypsa, Sherlina Thomas. Lastly, this project would not have been possible without the directors, teachers, providers, and families who agreed to participate. We offer them our sincerest thanks.

This project was funded by the U.S. Department of Education and the U.S. Department of Health and Human Services through Georgia's Race to the Top—Early Learning Challenge grant. We would like to thank the Early Learning Challenge grant leaders for their guidance and support of this evaluation. We hope that this and future Quality Rated validation reports will be used to support the continued improvement of Quality Rated and the programs that serve young children in Georgia.

REPORT HIGHLIGHTS

Key Findings

1. **Center-based programs and family child care learning homes (FCCLHs) with the highest Quality Rated star rating (three stars) were generally of higher quality than lower-rated programs.** In particular, preschool and toddler classrooms in 3-star center-based programs had higher-quality teacher-child interactions than lower-rated programs. Three-star FCCLHs had higher-quality provider-child interactions than 2-star, but not 1-star, FCCLHs. Toddler teachers and FCCLH providers in 3-star programs also offered richer language environments than those in lower-rated programs.

2. **We did not find evidence of differences at every level of star rating or on every independent measure of quality.** Although 3-star (and sometimes 2-star) programs were generally of higher quality, some findings were unexpected, showing that there are inconsistent relationships between the ratings and other measures.

3. **Preschool children in higher-rated programs learned more than children in lower-rated programs in some, but not all, domains.** Preschoolers in 3-star programs had stronger math and social skills at the end of the school year than their peers in lower-rated programs, after accounting for their skills at the start of the school year and some other characteristics. The number of stars a program earned was not associated with preschoolers' early language, literacy, or executive function skills, nor with toddlers' development in language or social skills.

4. **In center-based programs with higher star ratings, the work climate was better in terms of turnover, wages, and employee benefits.** Fewer directors in 2- and 3-star programs, compared to 0-star programs, reported that two of every three teachers, or more, had left and had to be replaced. The entry-level hourly wage for teachers was more likely to be over $12.50 in higher-rated center-based programs compared to lower-rated programs, and staff members were more likely to receive benefits.

Background

Quality Rated is Georgia's systematic approach to assessing, improving, and communicating the level of quality in early care and education programs.

This report is the last in a series of four from the Quality Rated Validation Project, and the second that presents data collected by Child Trends and Georgia State University. The overarching goal of the Quality Rated Validation Project is to provide Georgia's early childhood leaders with high-quality data about the validity of Quality Rated that can be used to strengthen the system. This report examines the relationship between star ratings and independent observations of program quality, children's' development over the school year, and work climate.

This report uses child assessment, observation, and questionnaire data collected by the validation study team during the 2016-17 and 2017-18 school years in 158 FCCLHs, and during the 2017-18 school year in 181 center-based programs.

TABLE OF CONTENTS

Table of Figures

Table of Tables

Common Abbreviations

Throughout this report, some words are frequently abbreviated. A list of these abbreviations is below.

CAPS	Childcare and Parent Services
CCLCs	Child Care Learning Centers
CCR&R	Child care resource and referral
CDI	MacArthur-Bates Communicative Development Inventories
CLASS	Classroom Assessment Scoring System
COP/TOP	Child Observation in Preschool and Teacher Observation in Preschool
DECA	Devereux Early Childhood Assessment
DECAL	Department of Early Care and Learning
ERS	Environmental Rating Scale
FCCLHs	Family Child Care Learning Homes
HTKS	Head-Toes-Knees-Shoulders
LENA	Language Environment Analysis
NAEYC	National Association for the Education of Young Children
WJ-IV	Woodcock Johnson Test of Achievement, 4th edition
WM III	Woodcock Muñoz-III

Quality Rated Validation Study Report #4: Quality Rated Star Ratings and Independent Measures of Quality, Children's Growth, and Work Climate

Introduction

Quality Rated is Georgia's systematic approach to assessing, improving, and communicating the level of quality in early care and education programs. In Quality Rated, center-based programs[a] and family child care learning homes (FCCLHs) apply to receive a star rating based on a combination of an online portfolio[b] and classroom observations of global quality using standardized tools called the Environment Rating Scales (ERS).

This report is the fourth and final in a series presenting findings from the Quality Rated Validation Project (see the pull-out box on the next page for key findings from the first three reports). As part of Georgia's Race to the Top—Early Learning Challenge grant, Georgia's Department of Early Care and Learning (DECAL) invested in evaluating Quality Rated. One part of that evaluation is the Quality Rated Validation Project led by Child Trends in partnership with Georgia State University.

The objectives of the Quality Rated Validation Project were to support Quality Rated leaders in future implementation and revision by providing them with information about (1) their administrative data system and how the ratings are functioning, (2) the extent to which the ratings are accurate and meaningful indicators of quality, and (3) the extent to which the ratings are linked to children's development and learning.

Structure of this report

On the next page, we summarize the first three reports from the Quality Rated Validation Project. The remainder of the report starts with a description of our three research questions and a summary of the validation study design and procedures. That section is followed by the findings, structured around the research questions. The report ends with a discussion of the study's limitations, key findings, and recommendations. For ease of reading, specific details of the research design, methods, and statistical analyses appear in the appendices.

[a] As in Report #3, in this report, we use the term *center-based programs* to refer to both Child Care Learning Centers (CCLCs) and unlicensed programs that are subject to different government oversight, which were categorized as *Others* in Reports #1 and #2. Although Reports #1 and #2 presented information about CCLCs and Others separately, we combined the two groups into a single category in this report due to the small number of Other programs in the current study sample (n=10).

[b] As part of the rating process, programs submit evidence in an online portfolio to earn points based on increasingly difficult criteria aligned with five standards: director and teacher qualifications; child health, nutrition, and physical activity; family engagement; intentional teacher practices; teacher to student ratios and group size.

Key findings from previous reports

There are three previous reports from this project: Quality Rated Validation Study Reports #1, #2, #3 (Early, Maxwell, Orfali, & Li, 2017; Orfali, Early, & Maxwell, 2018; Early et al., 2018).

Report #1 was based on administrative data through May 2017. Key findings included:

1. Programs earned a higher proportion of the available Structural Quality points than Process Quality points.
2. Programs that were held to more rigorous standards, such as Georgia's Pre-K and Head Start, generally attained higher star ratings.
3. The star rating is driven almost entirely by the Process Quality component (i.e., Environment Rating Scale score).

Report #2 was based on administrative data through December 2018. Key findings included:

1. Programs with Childcare and Parent Services (CAPS) scholarships—that is, funding to serve children from low-income families—had lower ratings than those without them. In addition, child care learning centers (CCLCs) that served infants and/or toddlers had lower ratings than those that did not.
2. CCLCs that were accredited by the National Association for the Education of Young Children (NAEYC) had higher ratings than CCLCs that were not.
3. Programs took about a year to submit their portfolio after applying to Quality Rated. After the portfolio was submitted, it took about four months to receive a rating.
4. Most programs that were re-rated—because their rating was expiring or at their request—either maintained (44%) or increased (39%) their rating.

Report #3 was based on questionnaires completed by directors, teachers, and providers. Key findings included:

1. Over three-quarters of center directors and FCCLH providers reported that they joined Quality Rated to be recognized as a high-quality program.
2. A large majority of FCCLH providers, center directors, preschool teachers, and toddler teachers had positive impressions of Quality Rated.
3. Although FCCLH providers and center directors tended to agree that the Quality Rated application process was time consuming, they did not typically see the process as more work than it was worth.
4. The two most-used Quality Rated supports were the bonus package based on the star rating (an incentive package given to programs that earn a rating of 1-star or above) and technical assistance from the program's child care resource and referral (CCR&R) agency.

RESEARCH QUESTIONS

The first two reports in this series addressed the first objective. This report addresses the second and third objectives using data collected specifically for this project: (a) independent observations, including audio recordings of teachers and providers; (b) assessments of children's emerging academic and social skills; and (c) director, teacher, and provider reports of work climate. This report compares each star rating (0- through 3-star) to every other star rating to determine when different ratings are linked to differences in observations, skills, or climate. Quality Rated leaders can use the findings to understand how the rating system is working for programs that take part. This report does not include information about other important aspects of the Quality Rated system, such as supports for improving quality. Likewise, since this study included only programs taking part in Quality Rated, the report does not reflect the general quality of care in Georgia, and it cannot be used to compare children or programs in Quality Rated to those not in Quality Rated.

Specifically, we aimed to answer the following three questions:

1 Are Quality Rated star ratings related to independent measures of quality?

Program ratings are public and are often tied to resources (e.g., access to technical assistance or tiered child care subsidy reimbursement); therefore, policymakers and the public should have confidence that ratings reflect quality and are meaningful. Additionally, addressing this question is a requirement of the Race to the Top—Early Learning Challenge grant. In this study, we used two different observational systems, as well as audio recordings, to assess the extent to which the Quality Rated star ratings differentiate levels of program quality. We expected that higher-rated programs would receive higher scores on classroom observations of teacher-child interactions; we also expected that in higher-rated programs, teachers and providers would speak more to children and use a wider vocabulary.

2 Are Quality Rated star ratings related to children's development over the school year?

This question is meant to link program quality to children's development and learning. Some previous research has shown a small but significant relationship between the quality of early care and education and children's development (Burchinal, 2017; Yoshikawa et al., 2013), and the Race to the Top—Early Learning Challenge grant required that this question be examined. To do so, we conducted one-on-one assessments of preschoolers' early academic and executive function skills at the beginning and end of the year. Additionally, at the beginning and end of the year, we asked teachers and FCCLH providers to report on younger children's language acquisition and all children's social skills. We expected that children in higher-rated programs would score higher on the various measures of development, after controlling for pre-test scores and other demographic factors.

3 Are Quality Rated star ratings related to work climate?

This study also presented an opportunity to consider how the work climate might differ in settings with different star ratings. Work climate is a broad term used here to reflect how well the workplace supports staff members to succeed (Whitebook et al., 2018). According to the QRIS Compendium (2018), 32 out of 44 states include Quality Rating and Improvement System (QRIS) indicators related to teacher supports, such as paid planning time or salary

guidelines. To better understand work climate at Quality Rated programs, we gathered information about directors', teachers', and FCCLH providers' job stress and commitment to teaching. We also collected information on teacher turnover, teacher pay, and receipt of benefits such as health insurance and paid sick leave in center-based programs. We consider these factors, in addition to observed quality and child outcomes, important for understanding program-level quality.

We anticipated that higher star ratings would be linked to lower stress and higher commitment among center-based staff and FCCLH providers. Additionally, we expected that higher star ratings would be linked to lower turnover, higher pay, and greater receipt of benefits in center-based settings.

Study Design and Procedures

This section describes this study's design and procedures, including how programs and children were selected for participation in the study, response rates, data collection tools, and participant characteristics. See the referenced appendices for more detailed information on each of these topics.

Sampling and recruitment

This section describes the sampling and recruitment of center-based and FCCLH programs for the study. For more detailed information, see Appendix A.

Center-based programs

In center-based programs, data collection took place during a single school year, 2017-18, and recruitment took place from July to October 2017. We conducted a power analysis to decide how many programs to recruit so that we could compare findings from each level of star ratings. Based on that analysis, we aimed to recruit a sample of 50 center-based programs at each star rating (1-star, 2-star, 3-star), as well as programs that completed the rating process but did not meet the criteria for a star, which we refer to as *0-star* for the purposes of this report. We invited a subset of randomly selected programs to join the study from the 1,140 that were in Quality Rated at the time of recruitment.

In total, we contacted 411 center-based programs, and 181 (44%) agreed to participate. We did not meet the goal of recruiting 50 programs at each star rating because the total number of center-based programs in Quality Rated is small at some star ratings, and many programs declined to participate. See Table 1 for response rates by star rating. The overall response rate was in the mid-range of response rates seen in other QRIS validation studies: Tout et al. (2017) reviewed reports from nine states and found that response rates ranged from 25 to 73 percent, with a median of 44 percent.

Within each participating center, up to two classrooms (one serving preschoolers and one serving toddlers) took part in the study. Overall, the study included 180 classrooms serving preschoolers and 152 classrooms serving toddlers. Teachers received a $50 gift card for participating in each component of the study: fall child assessments, winter observations and surveys, and spring child assessments. Directors received a $50 gift card for completing the survey and supporting study activities.

To recruit children to take part in the study, we mailed packages of consent forms to the participating teachers and asked them to distribute the forms to the parents of each child in their

classroom. The study team gave teachers a $25 gift card for collecting consent forms from at least 75% of enrolled families, regardless of whether the parent agreed. Most children in the study started participating at the beginning of the school year (fall 2017), but some children joined the study during the spring (2018) to offset attrition from the fall sample.

Teachers distributed parent consent forms to 4,165 families, and parents returned 2,341 (56%) positive consent forms. From those children whose parent agreed, the assessor randomly selected up to six children per classroom to participate, resulting in 1,187 children (457 toddlers and 730 preschoolers) from 173 programs. This response rate was similar to that of Rhode Island (52%; Maxwell, Blasberg, Early, Li, & Orfali, 2016), the only state to report its parental consent rate out of nine states included in a recent synthesis of QRIS validation studies (Tout et al., 2017).

FCCLHs

For FCCLHs, data collection took place during two school years, 2016-17 and 2017-18, to maximize the number of programs that participated. Recruitment of FCCLH providers for the first year of data collection took place from July to November 2016, and recruitment for the second year took place from July to October 2017. We invited all FCCLHs in Quality Rated to participate, regardless of star rating, because the number of FCCLHs in Quality Rated was relatively small. Across the two years of data collection, we invited 407 FCCLHs to participate, and 158 (39%) agreed. See Table 1 for response rates by star rating. As mentioned in the previous section, this response rate is in the mid-range of those seen in other QRIS validation studies. The validation study team offered providers three $50 gift cards, one each for the fall, winter, and spring data collection components.

We mailed packages of consent forms to FCCLH providers and asked them to distribute the forms to the parents of each eligible child attending their program. To be eligible for the study, children had to be at least two months old (by May 31) and no older than six years and not attending school, including Georgia's Pre-K[c] or kindergarten, during the day. All eligible children whose parent returned a positive consent form took part in the study. In the second year, to improve response rates and be consistent with the center-based study, the validation study team offered providers a $25 gift card for returning consent forms from all or almost all enrolled families, regardless of whether the parent agreed. Most children began taking part in the study at the start of the school year (fall 2016 or 2017); however, as in center-based programs, some children joined the study during the spring (2017 or 2018) to offset attrition from the fall sample.

Providers distributed parent consent forms to 953 families, and parents returned 651 (68%) positive consent forms; however, 36 positive consents were from children who were ineligible due to attending school during the day, reducing the response rate to 65 percent. Overall, the analyses included 601 children (273 infants and toddlers and 328 preschoolers) from 147 programs. Seven programs do not have children represented in the sample because no parent in those programs returned a positive consent form, but the analyses of program observations do include those programs.

[c] Georgia's Pre-K is a state-funded pre-kindergarten program that is free for all eligible four-year-old children, regardless of family income. Georgia's Pre-K programs usually operate on the local public school calendar for 6.5 hours a day, 180 days a year, and can be offered at both public schools and private child care centers (Georgia Department of Early Care and Learning, n.d.a).

Table 1. Response rates by program type and star rating

The response rate tended to increase with the program's star rating.

Star rating	Center-based programs			FCCLHs		
	Attempted	In study	Response rate	Attempted	In study	Response rate
0-star	80	28	35%	25	7	28%
1-star	113	39	35%	108	29	27%
2-star	126	64	51%	169	78	46%
3-star	92	50	54%	105	44	42%
Overall	***411***	***181***	***44%***	***407***	***158***	***39%***

Note: The star ratings for the "Attempted" columns were as of the midpoint of the observation window (February 15) for the year in which the programs would have participated. The star ratings for the "In Study" columns were the rating at the time of the classroom or program observation.

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

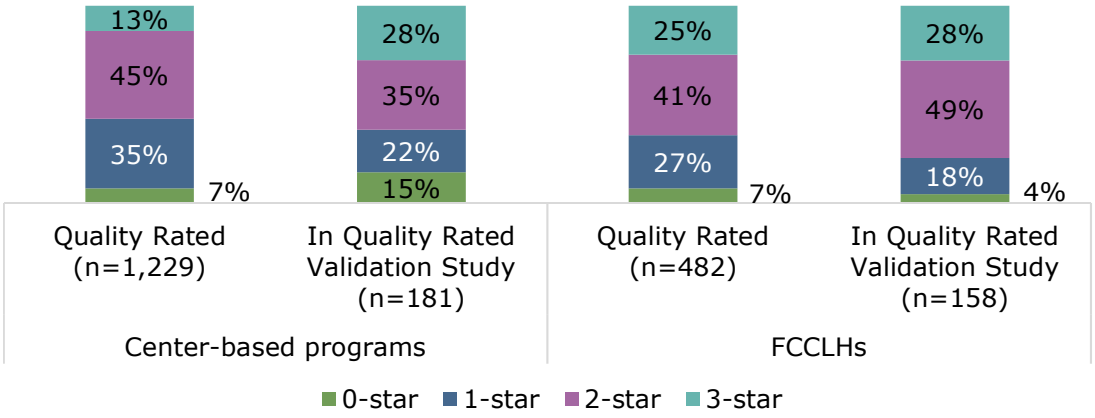
Star rating

The 181 center-based programs and 158 FCCLHs in the study represent 13 percent of rated center-based programs and 33 percent of FCCLHs in Quality Rated. To provide context for this report, Figure 1 presents the distribution of ratings for the programs in the study and for all Quality Rated programs. It is not surprising that the distribution of ratings for center-based programs was different in the study sample compared to Quality Rated programs overall because we intentionally sampled enough programs at each level to make comparisons.

As mentioned previously, for the purposes of this report, we refer to programs that completed the rating process but do not meet the criteria for one, two, or three stars, as 0-star. From a policy standpoint, DECAL considers these programs to be participating, but not rated, and does not use the term *0-star*. Because these 0-star programs sought a rating and took part in all aspects of the rating process, we thought it was important to include them when possible. However, very few FCCLH providers with 0-star ratings agreed to participate (n = 7), so the analyses separated by star rating do not include FCCLHs with 0-stars. They are, however, included in the overall FCCLH demographic information and in the child-level analyses when center-based programs and FCCLHs are combined.

Figure 1. Star ratings of all programs in Quality Rated and programs in the Quality Rated Validation Study sample

Programs in the study had a somewhat different star rating distribution than the overall Quality Rated population.



Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated. Programs that were rated after recruitment efforts for the study were not included.

Source: Quality Rated Administrative Data System, May 15, 2018

As described in Report #2 (Orfali, Early, & Maxwell, 2018), program ratings may have changed during the study period. This report uses the star rating that was current on the day of the Classroom Assessment Scoring System (CLASS) observation in the preschool classroom or FCCLH. Appendix A provides details about changes in star ratings for the study sample.

Representativeness of the sample

We intended for the programs in the study to represent the larger population of Quality Rated programs at each star rating. However, the response rate was lower for some star ratings than others. To ensure there were no systematic differences between the study sample and the larger population of Quality Rated programs, we used administrative data to compare the study sample by star rating to two groups: (1) programs that did not participate, and (2) the entire population of Quality Rated programs at the time of recruitment. We compared the groups on their average ERS scores, portfolio scores, and the percentage of center-based programs with Head Start funding or Georgia's Pre-K. The groups did not differ on most variables, with two exceptions: 3-star center-based programs in the study were more likely to receive Head Start funding than 3-star programs that did not participate, and 2-star center-based programs in the study had higher portfolio scores and were more likely to receive Head Start funding than 2-star programs that did not participate or 2-star programs overall. Due to the high level of similarity between the two groups, we concluded that the sample at each star rating adequately represents the population of Quality Rated programs. Details of the analyses appear in Appendix B.

Data collection process

Data collection took place during the fall, winter, and spring of each school year (2016-17 and 2017-18 for FCCLHs, and 2017-18 for center-based programs). Georgia State University and Child Trends worked in close collaboration to hire, train, and oversee the team of data collectors each year.

During the fall and spring visits, data collectors gave all participating preschool-aged children a brief set of assessments designed to measure their expressive vocabulary, early literacy, counting, early math, and executive function skills. Additionally, data collectors gave each teacher or FCCLH provider a questionnaire regarding the preschooler's social skills and behavior. We did not conduct

any individual assessments with infants or toddlers. Instead, data collectors gave each FCCLH provider or center-based teacher a questionnaire regarding the infant's or toddler's language acquisition and social skills.

FCCLHs and center-based classroom observations occurred in the winter. Classroom observations included measures of teacher-child interactions, as well as minute-by-minute coding of children's and teachers' activities. Additionally, the validation study team collected audio recordings of the language environment.

At the same time the winter classroom observations were taking place, we asked center directors, preschool and toddler teachers, and FCCLH providers to complete a questionnaire to gather information about their demographic characteristics, perceived stress, and job commitment. We also asked center directors about turnover and salary. For more information about data collection procedures, see Appendix C.

Description of measures

This study collected a wide array of data to provide a broad view of how program quality varies as a function of star rating, as well as the extent to which children's growth is linked to star rating. Appendix D provides additional details regarding each data collection instrument.

Classroom observations

During the winter of each data collection year, one or two observers visited each center-based classroom and FCCLH to measure teacher-child interactions and gather audio recordings of language.

Classroom Assessment Scoring System

Trained data collectors observed center-based preschool classrooms using the CLASS Pre-K (Pianta, La Paro, & Hamre, 2008), an observational tool that assesses the quality of the interactions between teachers and preschool-aged children (ages 3 to 5 years). The 10 CLASS Pre-K dimensions are organized into three domains: (1) Emotional Support, (2) Classroom Organization, and (3) Instructional Support.

Trained data collectors observed center-based toddler classrooms and FCCLHs using the CLASS Toddler (La Paro, Hamre, & Pianta, 2012), which assesses the quality of the interactions between teachers and toddlers (ages 15 to 36 months). CLASS Toddler includes eight dimensions organized into two domains: (1) Engaged Support for Learning, and (2) Emotional and Behavioral Support.

For both the CLASS Pre-K and the CLASS Toddler assessments, observers score each dimension on a 7-point scale, with scores of 1 and 2 considered low quality; 3, 4, and 5 considered mid-range quality; and 6 and 7 considered high quality.

Child Observation in Preschool and Teacher Observation in Preschool

We used the Child Observation in Preschool (COP; Farran, Plummer, Kang, Bilbrey, & Shufelt, 2006) and its companion, the Teacher Observation in Preschool (TOP; Bilbrey, Vorhous, Farran, & Shufelt, 2007), as additional observational measures of child and teacher behavior in a subset of center-based preschool classrooms.[d] To complete the COP/TOP, an observer conducts multiple rounds of coding, referred to as sweeps. During each sweep, the observer watches each teacher and each child in the classroom for approximately three seconds, in succession, starting with the lead teacher. Each three-second observation is coded on a series of dimensions. We later transformed these

[d] COP/TOP observations were conducted in only 138 of the 172 preschool classrooms because one data collector was not able to achieve reliability on the tool, and there was not enough time to train another individual. See Appendix D for more details on sampling for the COP/TOP and Appendix G for the sample sizes by star rating.

data into eight scores that previous research has indicated are linked to children's outcomes: (1) transition time (routines and wait time for children), (2) quality of instruction, (3) emotional climate, (4) teachers listening to children, (5) sequential activities, (6) social learning interactions, (7) child involvement, and (8) math opportunities (Farran, Meador, Christopher, Nesbitt, & Bilbrey, 2017). We also analyzed a ninth score, literacy opportunities, because it measures a construct of particular interest to DECAL and researchers.

Language Environment Analysis

The Language Environment Analysis digital language processor (LENA; Xu, Yapanel, & Gray, 2009) is a recording device intended to capture spoken language. To measure the richness of the language environment, we asked teachers and providers to wear the LENA device (rather than the children, as is traditionally done) to record their speech on the morning of their CLASS observation. Using these recordings, we created three variables for analysis. The first variable was *adult word count*, which measures the number of words spoken per minute. The second variable was *average length of utterances* in words. An utterance is a spoken sound, word, or statement (e.g., "mmhm," "hello," "that is a beautiful dress"), and can be any length from a single word to a phrase or sentence. This variable is a proxy for language quality, whereas adult word count is a measure of language quantity. The third variable was a proportion (ranging from zero to one) of the number of different words spoken in relation to the total number of words spoken (Kemper & Sumner, 2001). A higher proportion indicates that the speaker uses a more varied vocabulary, while a lower proportion indicates a more repetitive vocabulary.

Child assessments

Throughout this report, we refer to the fall assessment as *pre-test* and the spring assessment as *post-test*.

Preschool-aged children (at least 36 months old by the end of the spring assessment window) were directly assessed by trained data collectors in the fall and spring using the following assessments:

- Counting Bears (NCEDL, 2001) measures the child's ability to count to 40 objects using one-to-one correspondence.
- We used three subtests of the Woodcock Johnson Test of Achievement, 4th edition (WJ-IV; Schrank, McGrew, Mather, & Woodcock, 2014). The first, Picture Vocabulary, assesses the child's expressive vocabulary. The second, Letter-Word Identification, assesses the child's early literacy skills. The third, Applied Problems, assesses the child's early math skills.
- Head-Toes-Knees-Shoulders (HTKS; Ponitz, McClelland, Matthews, & Morrison, 2009) measures executive function skills of inhibitory control, working memory, and attention.

Teachers or FCCLH providers completed questionnaires to report on infants' emerging language skills, toddlers' early language and social skills, and preschool-aged children's social skills.

- The Devereux Early Childhood Assessment for Toddlers (DECA-T; Mackrain, LeBuffe, & Powell, 2007) and the DECA for Preschoolers, Second Edition (DECA-P2; LeBuffe & Naglieri, 2012) measure social skills. The DECA-P2 also measures behavioral concerns.
- The LENA Developmental Snapshot (Gilkerson, Richards, Greenwood, & Montgomery, 2016) measures young children's language acquisition skills.
- The MacArthur-Bates Communicative Development Inventories, short forms (CDI; Fenson, Pethick, Renda, & Cox, 2000) measure children's vocabulary production from a list of developmentally appropriate words.

See Table 2 for a summary of the constructs and tools used to assess each age group. Note that for the analyses, we combined the information about the language acquisition skills of infants and toddlers.

Children were assessed with both English and Spanish assessments when their parents reported that Spanish was spoken at home or was the child's dominant language. The battery included Counting Bears in Spanish and the same three subtests of the Woodcock Muñoz-III (WM III; Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005), in addition to Counting Bears in English and the three subtests of the WJ IV in English. Because HTKS is not a measure of language ability, children completed it once, in their dominant language.[e]

Table 2. Constructs and tools used to assess children's skills

The validation study team used a wide range of measures to gather data on infants', toddlers', and preschoolers' skills in the fall and spring.

Construct	Infants	Toddlers	Preschoolers
Language acquisition	LENA Developmental Snapshot	LENA Developmental Snapshot	-
Expressive vocabulary		CDI-Toddler	WJ-IV Picture-Vocabulary
Early literacy		-	WJ-IV Letter-Word
Counting		-	Counting Bears
Early math		-	WJ-IV Applied Problems
Executive functioning		-	HTKS
Social skills		DECA-Toddler	DECA-P2
Behavioral concerns		-	DECA-P2

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Work climate

Center directors, preschool teachers, toddler teachers, and FCCLH providers were given a questionnaire around the time of their classroom or program observation. The questionnaire included the following constructs related to work climate:

- The four-item version of the Perceived Stress Scale (Cohen, Karmarck, & Mermelstein, 1983) includes questions such as "In the last month, how often have you felt that things were going your way?" Participants responded using a Likert scale from never (0) to very often (4).
- The How Committed Am I? scale (Jorde-Bloom, 1988) includes 10 items such as "I often think of quitting," and "I put a lot of extra effort into my work." Participants rated each item on a scale from strongly disagree (1) to strongly agree (5).

[e] The child's FCCLH provider also filled out the CDI in Spanish if the provider regularly spoke Spanish with the child and had knowledge of his or her Spanish language abilities. However, this group was too small (n = 18 across all star ratings) to include in any analyses. No center-based teachers in our sample spoke Spanish with the children in their classrooms.

- Center directors reported (1) how many lead and assistant teachers they currently employed, and (2) how many lead and assistant teachers had left their program and had to be replaced in the past 12 months. We divided the number of teachers who needed to be replaced by the number currently employed to capture turnover for lead and assistant teachers. We then grouped the amount of turnover into four categories: none, 1 to 33 percent, 34 to 67 percent, and 68 percent or more.
- Directors reported the hourly wage for an entry-level preschool teacher and entry-level toddler teacher at their center. We grouped wages into four categories: $8.50 or below, $8.51 to $10.50, $10.51 to $12.50, and above $12.50. We also asked center directors and teachers what benefits they received, from a list of twelve benefits, such as health insurance and paid vacation.

We do not have turnover, salary, or benefit information about FCCLH providers because they are often small business owners with no employees, and their earnings from providing care can be difficult to obtain in a self-report format.

Analysis

Because all of our research questions are concerned with how programs with different star ratings differ from one another, we compared each star rating to every other star rating for each set of analyses. In this report, we describe all p-values of 0.05 or smaller as *statistically significant.* In addition, we calculated the effect size using Cohen's d for each statistically significant finding. Except where noted, all the effect sizes in this report met the What Works Clearinghouse (2014) definition of *substantively important,* meaning an effect size of 0.25 or higher.

For all independent measures of quality (e.g., CLASS, LENA, COP/TOP), we conducted ANOVAs followed by pairwise comparisons. We divided the sample into three groups: 1) preschool classrooms in center-based programs, 2) toddler classrooms in center-based programs, and 3) FCCLHs (1-, 2-, and 3-star only). We followed a similar strategy of conducting ANOVAs followed by pairwise comparisons for each star rating for teachers', directors', and providers' reported stress and commitment. For the measures of teacher turnover and pay, we categorized the responses because the data were highly skewed. We then used chi-squared tests to compare the groups by star rating. See Appendix E for more details.

To examine the extent to which children's early academic and social development varied by star rating, we conducted multilevel models. These models accounted for the fact that children attending the same program were more likely to be similar to one another than to children attending other programs. The multilevel models controlled for children's pre-test scores, as well as family poverty (below 100% of the poverty line, 100-185% of the poverty line, over 185% of the poverty line), children's race (black, white, other[f]), and children's dominant language (English, other[g]).

Details about data entry and validation, handling missing data, handling children with different assessment batteries in the fall and spring, and choosing control variables appear in Appendix E.

Program and child characteristics

This section provides a brief overview of the characteristics of study participants. More detailed information about teachers, providers, and programs in the study appears in Appendix F of this report and Appendix A of Report #3 (Early et al., 2018).

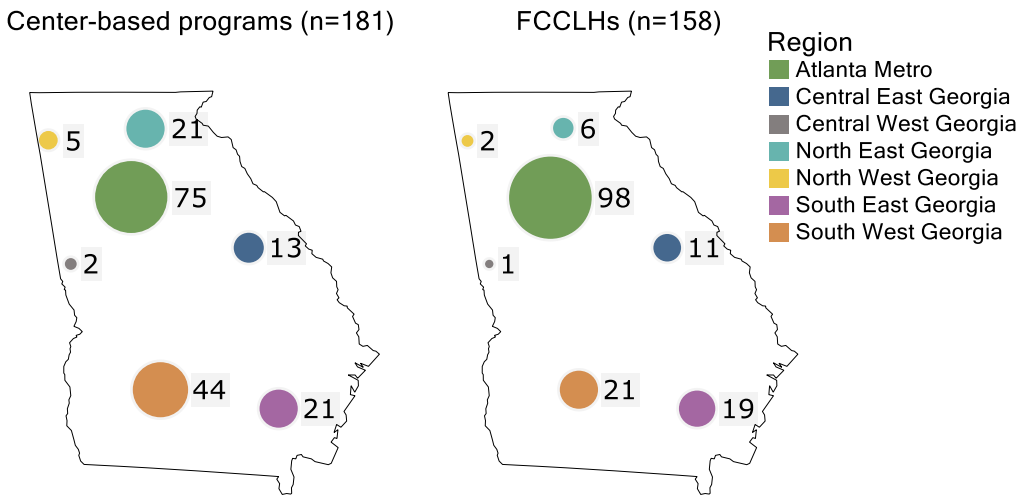
[f] Over half of the children in the "other" category were multi-racial, according to parent report. For more details about the breakdown of this category, see Table 4.
[g] The majority of children whose dominant language was something other than English spoke Spanish. See Table 4 for more details.

Characteristics of participating FCCLH and center-based programs

Figure 2 shows the number of programs in different regions of the state, with larger dots indicating more programs. The largest dot represents the large number of programs in the Atlanta area.[h] The remainder of the dots show the distribution of the programs in regions outside the Atlanta area.[i]

Figure 2. Distribution of programs participating in the study across the state

Most programs were in the Atlanta metropolitan area, but all regions of the state were represented in the study.



Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Participating center-based programs varied widely in the number of children enrolled, with a median of 88. Over three-quarters (78%) of participating center-based programs served at least one child receiving a Childcare and Parent Services (CAPS) scholarship—that is, child care subsidy funding to serve children from low-income families. The median ratio of adults to children, as reported by center directors, was 1:9 in preschool classrooms and 1:7 in toddler classrooms. FCCLH providers served a median of 6 children. Almost half (42%) of FCCLHs had at least one child enrolled who received a CAPS scholarship. Georgia's CAPS policies offer higher child care subsidy payment rates to programs that receive a star rating of 1, 2, or 3 in Quality Rated. Because center-based programs serve more children than FCCLHs, the increased subsidy payments may have been a greater motivator to centers than FCCLHs; this may, in part, explain the difference between CAPS participation in the two types of settings. The median ratio of adults to children, as reported by FCCLH providers, was 1:5. More details about participating programs appear in Appendix F.

Characteristics of participating children

Table 3 shows the demographic characteristics of the children in the study, as reported by their parent during the consent process. There were slightly more boys than girls, and roughly half were black/African American. Between one-fifth and one-third of children were from families with incomes below the poverty line for their family size. More details about participating children appear in Appendix F.

[h] The Atlanta area includes the entire Atlanta-Sandy Springs-Roswell Metropolitan Statistical Area: https://dch.georgia.gov/sites/dch.georgia.gov/files/Atlanta%20Service%20Area%20Map.pdf.

[i] To ensure the confidentiality of the participating programs, those outside the Atlanta area are grouped by Child Care Resource and Referral region (rather than exact location): http://decal.ga.gov/CCS/CCRRSystem.aspx.

Table 3. Demographic information about the children in the study

Between 19 and 40 percent of the families in the study had incomes that fell below the federal poverty level. The majority of children were either black or white and spoke English.

		Center-based Programs		FCCLHs	
		Toddler (n=374-457)	Preschool (n=604-730)	Infant/ Toddler (n=221-272)	Preschool (n=270-328)
		Percentage	*Percentage*	*Percentage*	*Percentage*
Gender	Boy	48%	51%	60%	53%
	Girl	52%	49%	40%	47%
Race/ Ethnicity	Black/African American	46%	46%	57%	56%
	White/Caucasian	37%	33%	26%	22%
	Hispanic/Latino	3%	8%	5%	8%
	Other	1%	2%	0%	1%
	Multi-racial	13%	11%	11%	13%
Family poverty level	Below 100%	31%	40%	19%	20%
	100-185%	24%	23%	20%	23%
	Above 185%	44%	38%	62%	57%
Language(s) spoken at home	English	97%	94%	98%	97%
	Spanish	5%	10%	7%	11%
	Other	4%	3%	2%	1%

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

FINDINGS

1 Are Quality Rated star ratings related to independent measures of quality?

There was evidence that programs with higher star ratings scored higher on the CLASS, an independent measure of quality. This was especially true for preschool and toddler classrooms in 3-star center-based programs and 3-star FCCLHs, which generally scored higher than those in lower-rated programs. However, we did not find differences between each star rating and did not find differences on some of our independent measures of quality.

The remainder of this section describes how the star ratings compared for the following measures:

- Teacher-child interactions in center-based preschool classrooms (as measured by CLASS Pre-K) and toddler classrooms (as measured by CLASS Toddler), and provider-child interactions in FCCLHs (as measured by CLASS Toddler)
- Richness of the language environment in center-based preschool classrooms, toddler classrooms, and FCCLHs (as measured by LENA)

- Child and teacher behavior in a subset of center-based preschool classrooms (as measured by COP/TOP)

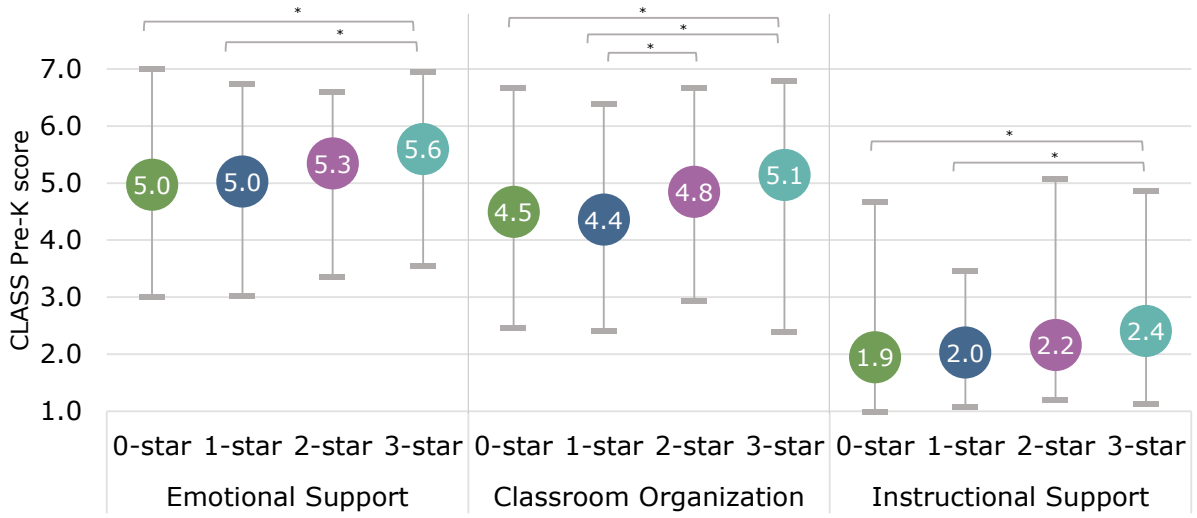
Teacher-child and provider-child interactions

Figure 3 illustrates the findings for the CLASS Pre-K. For all figures in this section (Figures 4 through 8), the values in the circles represent the average score for each star rating. Each circle is plotted on a vertical line that represents the range of scores (minimum to maximum) for the group. For example, in Figure 3, classrooms in 0-star programs scored 5.0 on Emotional Support, on average. The lowest scoring classroom in that group had a score of 3.0, and the highest scoring classroom in that group had a score of 7.0. The brackets with asterisks represent pairs of means that were statistically different from one another ($p < .05$). Pairs without a bracket were not significantly different from one another. For example, in Figure 3, the brackets indicate that preschool classrooms in 3-star programs scored significantly higher on Emotional Support than classrooms in 0- and 1-star programs, but not significantly higher than 2-star programs.

In each of the three domains, preschool classrooms in 3-star programs scored higher than those in both 0-star and 1-star programs. For Classroom Organization, preschool classrooms in 2-star programs also scored higher, on average, than those in 1-star programs. Otherwise, there were no differences among classrooms in 0-, 1-, or 2-star programs. See Appendix G for more descriptive information and effect sizes.

Figure 3. CLASS Pre-K averages and ranges for center-based preschool classrooms, by star rating

In each domain, classrooms in 3-star programs scored higher than classrooms in 0-star and 1-star programs. For Classroom Organization, classrooms in 2-star programs also scored higher than those in 1-star programs.

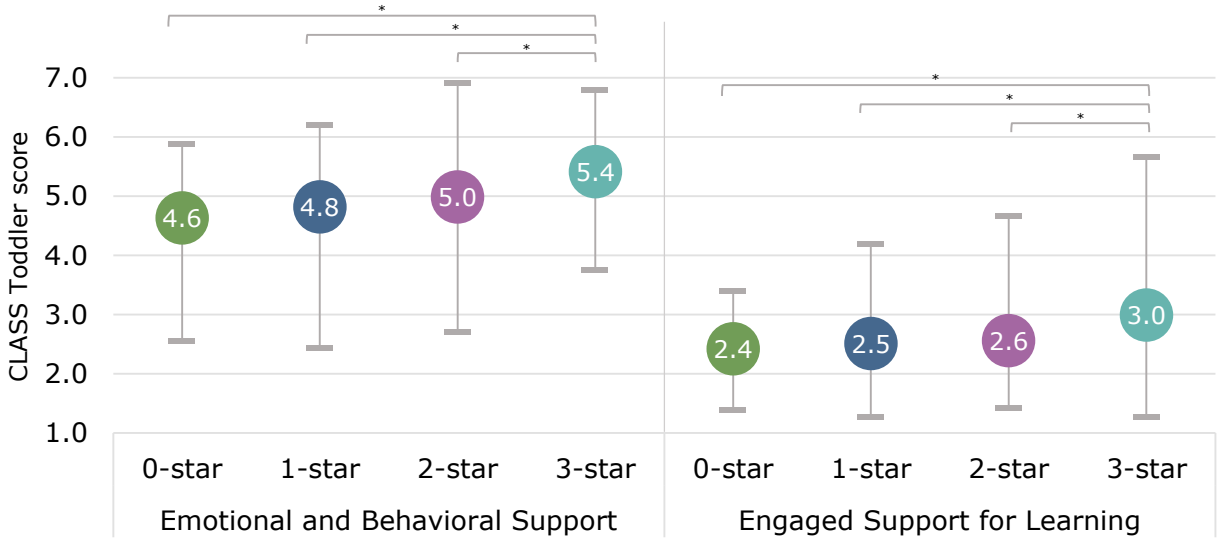


Source: Validation study team data collection in center-based programs, 2017-18 school year

As seen in Figure 4, toddler classrooms in 3-star center-based programs scored higher than any of the other groups (0-, 1-, and 2-star programs) on both Emotional and Behavioral Support and Engaged Support for Learning. No differences were found on CLASS Toddler domains among 0-, 1-, and 2-star programs.

Figure 4. CLASS Toddler averages and ranges for center-based toddler classrooms, by star rating

Both domain scores were higher in 3-star programs than any other star rating.

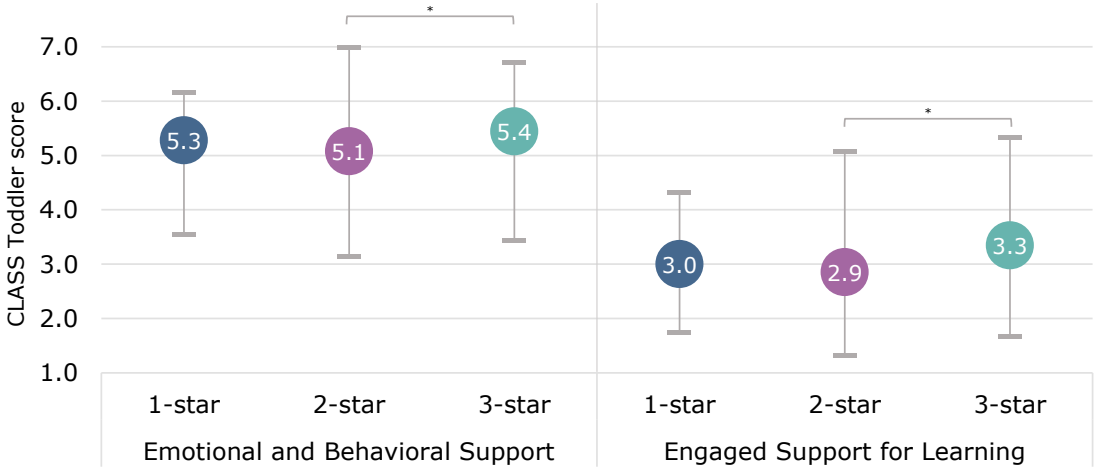


Source: Validation study team data collection in center-based programs, 2017-18 school year

Figure 5 shows the average CLASS Toddler domain scores for FCCLHs. In both domains, 3-star FCCLHs scored higher than 2-star FCCLHs, but not higher than 1-star FCCLHs. As noted earlier, there were too few 0-star programs to include them in these analyses.

Figure 5. CLASS Toddler averages and ranges for FCCLHs, by star rating

Across both domains, 3-star FCCLHs scores higher than 2-star FCCLHs.



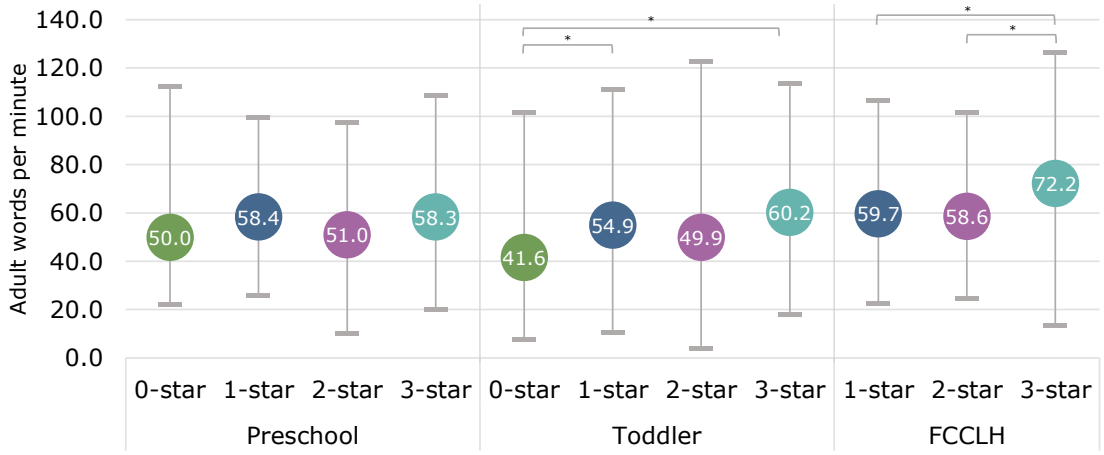
Source: Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Richness of the language environment

Figure 6 shows the number of words spoken per minute (i.e., adult word count) by teachers and FCCLH providers during their CLASS observation, by star rating. There were no significant differences by star rating on LENA adult word count per minute in preschool classrooms. Toddler teachers in 1- and 3-star programs spoke more words per minute than toddler teachers in 0-star programs. The difference between 1- and 3-star toddler teachers' words per minute was not significant, and the number of words spoken by toddler teachers in 2-star programs did not differ from any other star rating. Providers in 3-star FCCLHs spoke more words per minute than providers in 1- and 2-star FCCLHs. See Appendix G for more descriptive information and effect sizes.

Figure 6. LENA adult word count per minute averages and ranges, by setting and star rating

Toddler teachers in 1- and 3-star programs spoke more words per minute than those in 0-star programs. FCCLH providers in 3-star programs spoke more words per minute than other FCCLH providers.

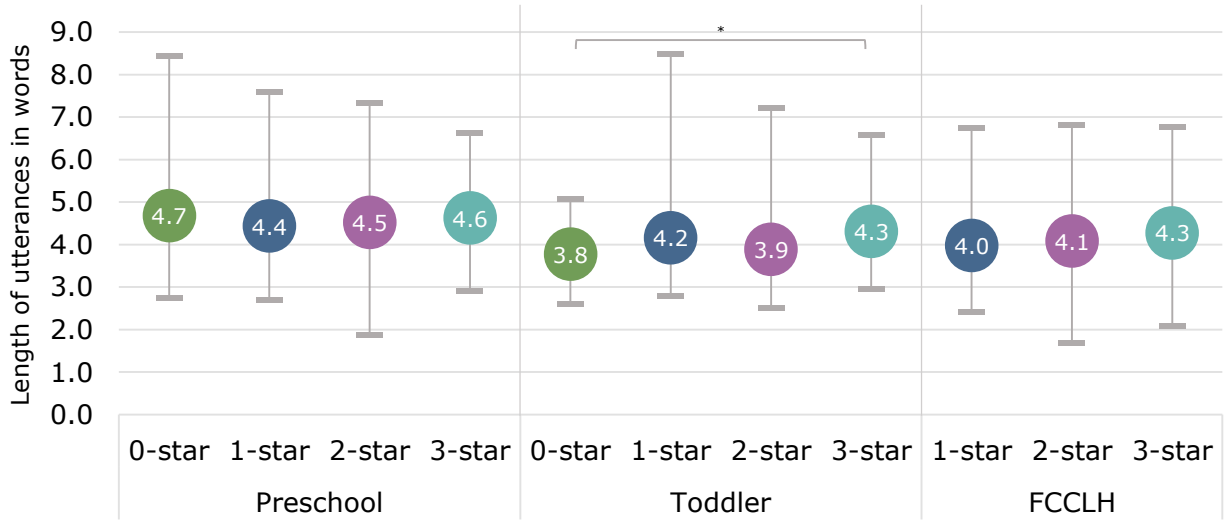


Source: Validation study team data collection in center-based programs, 2017-18 school year

As shown in Figure 7, toddler teachers in 3-star programs spoke longer utterances, on average, than toddler teachers in 0-star programs. There were no other significant differences by star rating, either in toddler classrooms or preschool classrooms. FCCLH providers also did not differ in the average length of their utterances by star rating.

Figure 7. Length of utterances averages and ranges, by setting and star rating

Toddler teachers in 3-star programs spoke more words per utterance than those in 0-star programs.

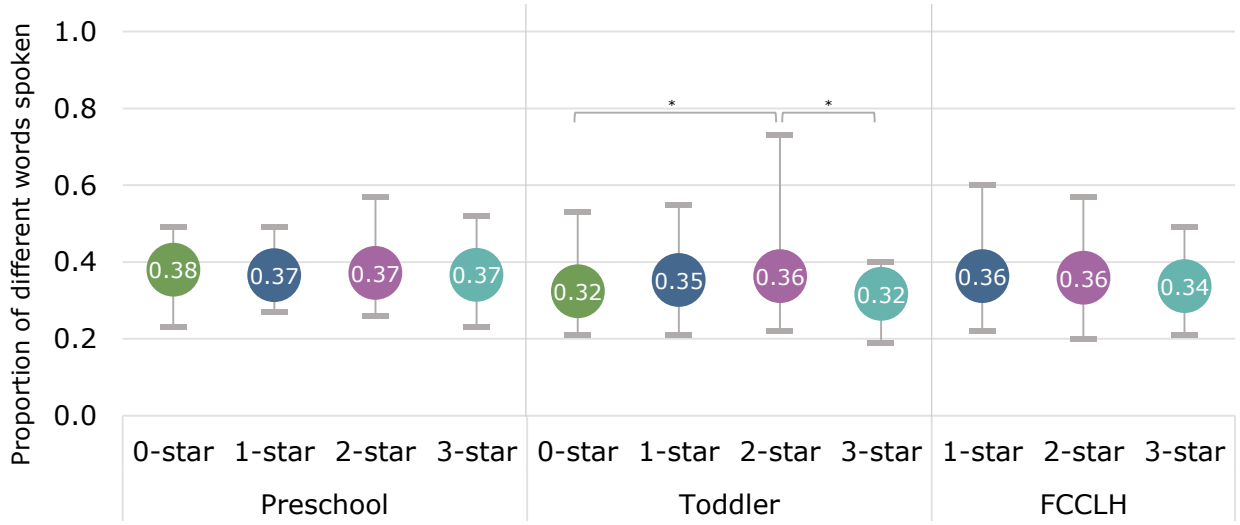


Source: Validation study team data collection in center-based programs, 2017-18 school year

In terms of language sophistication, toddler teachers in 2-star programs used a wider variety of words than those in 0- or 3-star programs. Toddler teachers in 1-star programs were not different from any other group on word variety. As shown in Figure 8, there were not any differences in word variety between preschool teachers or FCCLH providers by star rating.

Figure 8. Vocabulary sophistication averages and ranges, by setting and star rating

Toddler teachers in 2-star programs used a wider vocabulary than those in 0- or 3-star programs.



Source: Validation study team data collection in center-based programs, 2017-18 school year

Child and teacher behavior in preschool classrooms

Eight of the nine scores derived from the COP/TOP data did not differ significantly by star rating. Across all star ratings, preschool children spent from 27 to 31 percent of their time in transitions and from 8 to 10 percent of their time in literacy activities. Preschool classrooms in 2-star programs scored significantly higher than those in 0- or 1-star programs on one of the constructs, the emotional climate of the classroom. See Appendix H for more detailed information about scores on the COP/TOP by star rating.

Summary of independent measures of quality findings

The table below summarizes the findings from the independent measures of quality and indicates some relationship between star ratings and each of the measures. We found fewer differences by star rating for the language environment and child and teacher behavior than for teacher-child interactions in preschool classrooms. Results indicated that classrooms in 3-star programs were of higher quality than those in lower-rated programs. Few differences were seen between 0-, 1-, and 2-star programs.

			1-star	2-star	3-star
Preschool classrooms	Teacher-child interactions (CLASS)	Emotional support	n.s.	n.s.	3>0, 3>1
		Classroom organization	n.s.	2>1	3>0, 3>1
		Instructional support	n.s.	n.s.	3>0, 3>1
	Language environment (LENA)	Adult word count	n.s.	n.s.	n.s.
		Average length of adult utterances	n.s.	n.s.	n.s.
		Vocabulary sophistication	n.s.	n.s.	n.s.
	Child and teacher behavior (COP/TOP)	Transition time	n.s.	n.s.	n.s.
		Quality of instruction	n.s.	n.s.	n.s.
		Emotional climate	n.s.	2>0, 2>1	n.s.
		Teachers listening to children	n.s.	n.s.	n.s.
		Sequential activities	n.s.	n.s.	n.s.
		Social learning interactions	n.s.	n.s.	n.s.
		Child involvement	n.s.	n.s.	n.s.
		Math	n.s.	n.s.	n.s.
		Literacy opportunities	n.s.	n.s.	n.s.
Toddler classrooms	Teacher-child interactions (CLASS)	Engaged support for learning	n.s.	n.s.	3>2, 3>1, 3>0
		Emotional and behavioral support	n.s.	n.s.	3>2, 3>1, 3>0
	Language environment (LENA)	Adult word count	1>0	n.s.	3>0
		Average length of adult utterances	n.s.	n.s.	3>0
		Vocabulary sophistication	n.s.	2>3, 2>0	n.s.
FCCLHs	Teacher-child interactions (CLASS)	Engaged support for learning		n.s.	3>2
		Emotional and behavioral support		n.s.	3>2
	Language environment (LENA)	Adult word count		n.s.	3>2, 3>1
		Average length of adult utterances		n.s.	n.s.
		Vocabulary sophistication		n.s.	n.s.

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

2 Are Quality Rated star ratings related to children's academic and social development?

There was evidence that star ratings were related to preschoolers' emerging math and social skills. This was especially true for 2- and 3-star programs, where preschoolers' math and social development were significantly higher than those of preschoolers in lower-rated programs. However, the number of stars a program earned was not related to preschoolers' or toddlers' early language development or to preschoolers' early literacy or executive function skills.

Details for each outcome appear below. First, we compared all children's post-test scores by star rating. Next, we conducted these same analyses to examine children attending center-based programs and FCCLHs separately.[j] Each analysis controls for pre-test scores as well as demographic characteristics. For the full regression tables and effect sizes, see Appendix I.

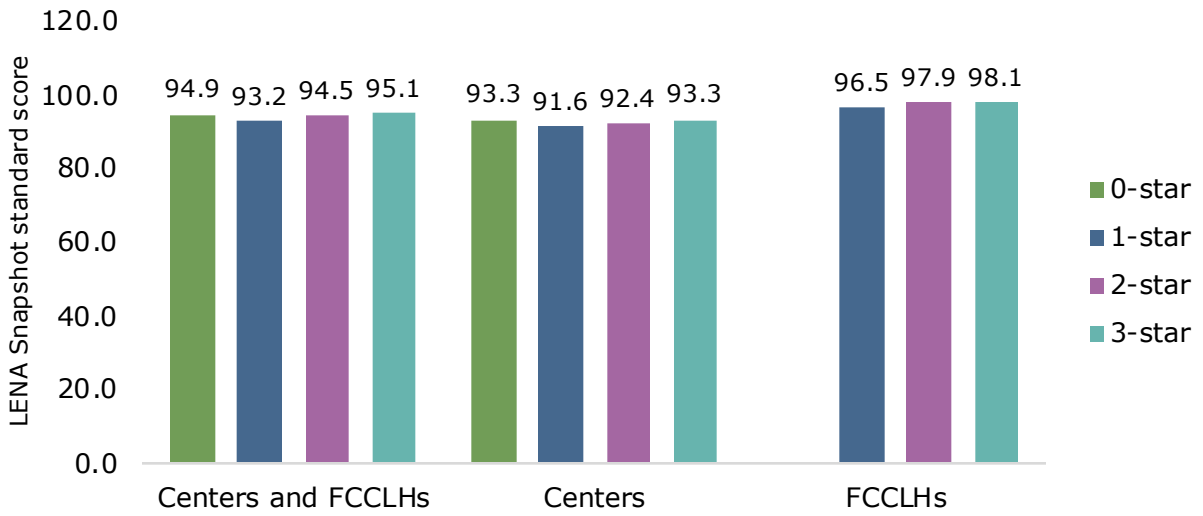
Figures in this section present adjusted post-test means. An adjusted mean is the estimated average post-test score for a child who had average values on all the other variables in the model, including pre-test and demographic characteristics.

Language and literacy

Infants' and toddlers' language acquisition (LENA Snapshot) did not vary by program star rating. This finding was also true when center-based programs and FCCLHs were analyzed separately. See Figure 9 for the adjusted means by group. The adjusted mean indicates that a child who had an average language acquisition score at pre-test, for example, had a score of 94.9 at post-test if they were in a 0-star program and 95.1 at post-test if they were in a 3-star program.

Figure 9. Adjusted means for infants' and toddlers' language acquisition, by setting and star rating

The number of stars a program earned was not related to infants' and toddlers' language acquisition.



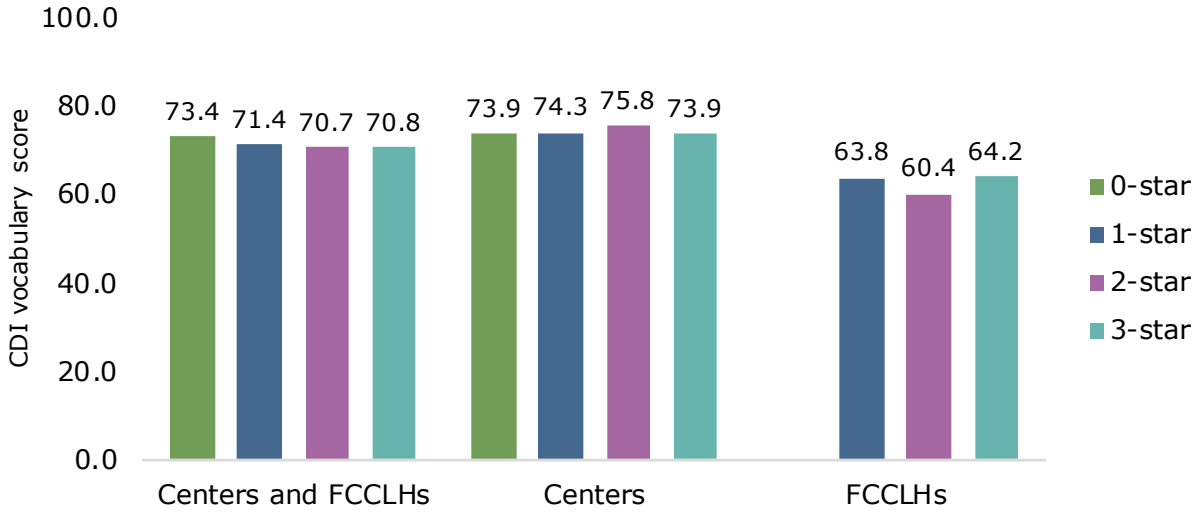
Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

The same pattern held for toddlers' expressive vocabulary, as measured by teachers' reports on the CDI (see Figure 10). Expressive vocabulary did not vary by star rating when center-based programs and FCCLHs were combined nor when they were examined separately.

[j] As elsewhere in this report, due to the small number of 0-star FCCLHs, the sub-analysis compared only 1-, 2-, and 3-star programs.

Figure 10. Adjusted means for toddlers' expressive vocabulary, by setting and star rating

The number of stars a program earned was not related to toddlers' expressive vocabulary.

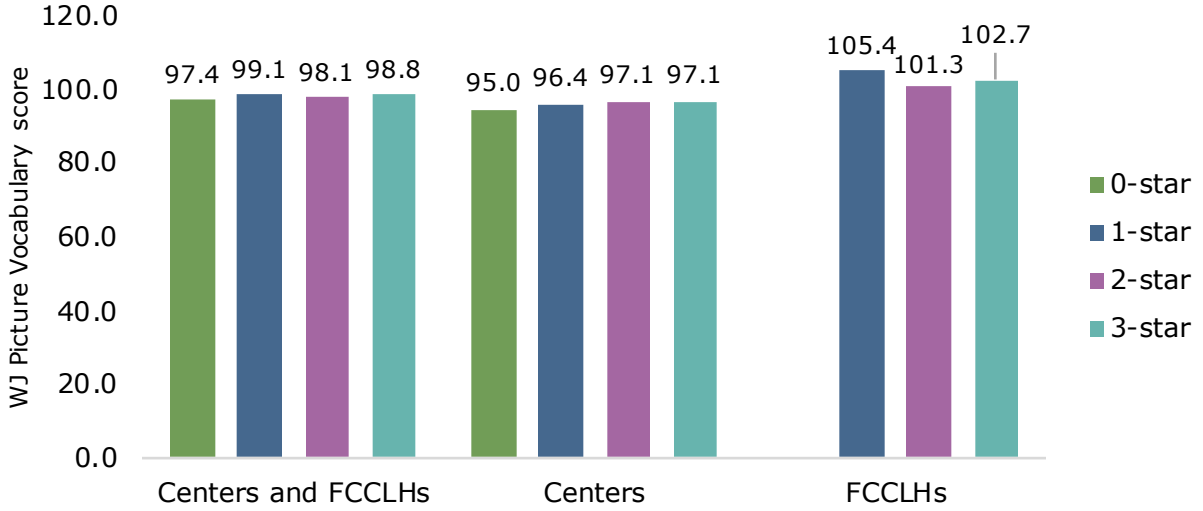


Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school year.

Preschoolers' expressive vocabulary (Woodcock-Johnson Picture Vocabulary) did not vary by program star rating (see Figure 11). This was also true when examining preschoolers attending center-based programs and FCCLHs separately.

Figure 11. Adjusted means for preschoolers' expressive vocabulary skills, by setting and star rating

The number of stars a program earned was not related to preschoolers' expressive vocabulary skills.

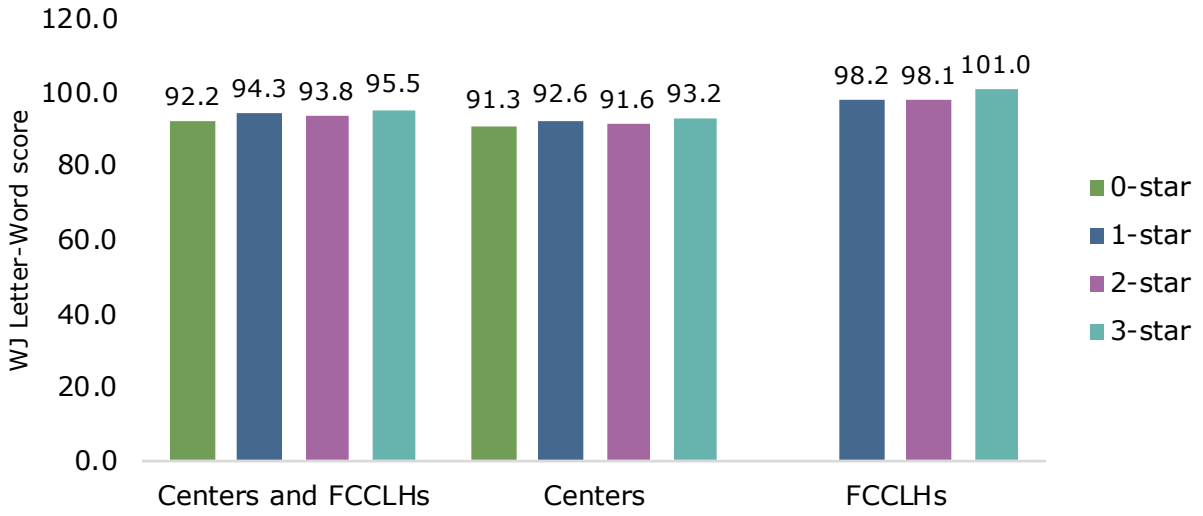


Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Preschoolers' early literacy skills (Woodcock-Johnson Letter-Word Identification) did not vary by program star rating (see Figure 12). This was also true when examining preschoolers attending center-based programs and FCCLHs separately.

Figure 12. Adjusted means for preschoolers' early literacy skills, by setting and star rating

The number of stars a program earned was not related to preschoolers' early literacy skills.



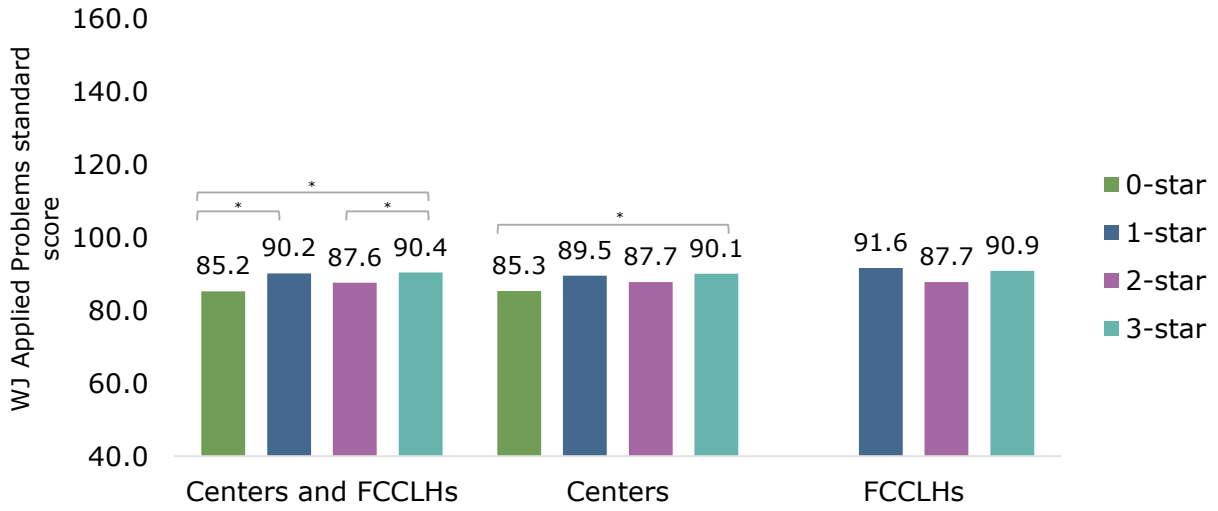
Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Math

Across center-based programs and FCCLHs, preschoolers attending 3-star programs had higher early math skills (Woodcock-Johnson Applied Problems), compared to those attending 2- and 0-star programs (see Figure 13), although the difference between 3- and 2-star programs was too small (*d* = 0.18) to be defined as substantively important. Preschoolers attending 1-star programs also had higher early math skills than those attending 0-star programs. The difference between 2-star and 0- or 1-star programs were not significant. In center-based programs, preschoolers attending 3-star programs had higher early math skills compared to those attending 0-star programs. The differences between 0-, 1-, and 2-star center-based programs were not significant. For preschoolers attending FCCLHs, early math skills did not differ significantly by star rating.

Figure 13. Adjusted means for preschoolers' early math skills, by setting and star rating

Preschoolers attending 3-star programs had higher early math skills compared to those attending 2- and 0-star programs. Preschoolers in 1-star programs had higher math skills than those in 0-star programs.

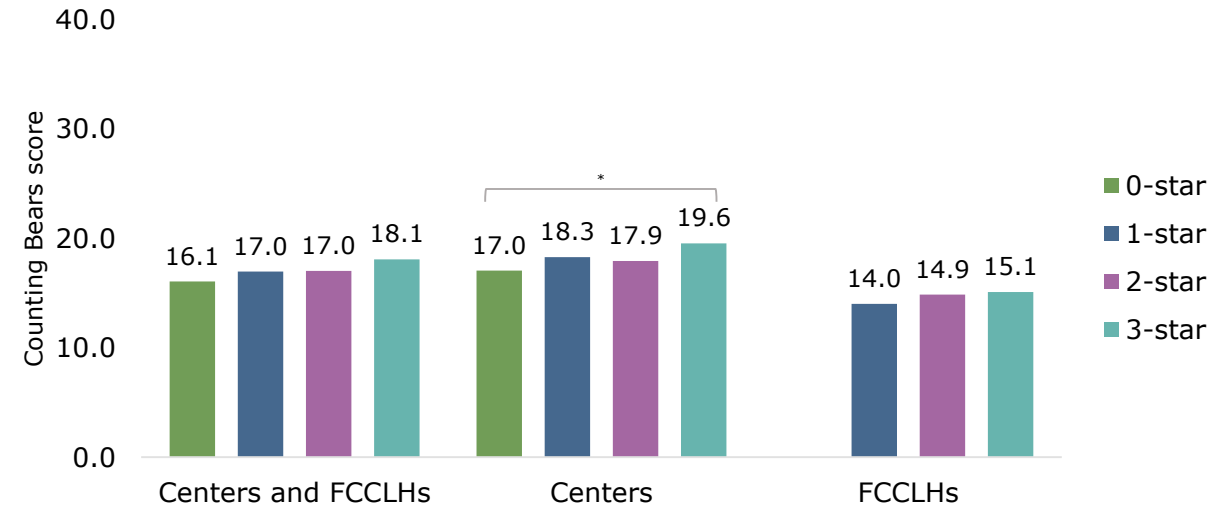


Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Preschoolers' counting abilities, as measured by Counting Bears, did not vary by star rating when children attending center-based programs and FCCLHs were combined. This finding held for preschoolers who attended FCCLHs. In center-based programs, preschoolers attending 3-star programs scored significantly higher on Counting Bears, compared to those attending 0-star programs (see Figure 14); however, this difference was slightly smaller (d = 0.22) than the What Works Clearinghouse (2014) definition of substantively important.

Figure 14. Adjusted means for preschoolers' counting abilities, by setting and star rating

Preschoolers attending 3-star center-based programs counted higher compared to those attending 0-star center-based programs.



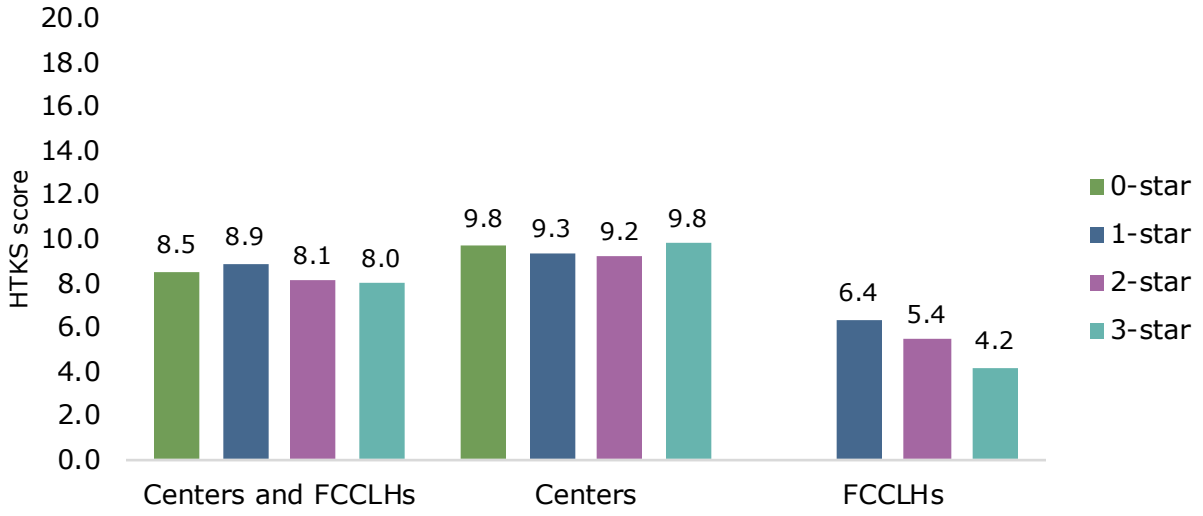
Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Executive function

Preschoolers' executive functioning (Head-Toes-Knees-Shoulders) did not vary by program star rating (see Figure 15).[k] This finding was also true when we analyzed center-based programs and FCCLHs separately.

Figure 15. Adjusted means for preschoolers' executive functioning, by setting and star rating

The number of stars a program earned was not related to preschoolers' executive functioning.



Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

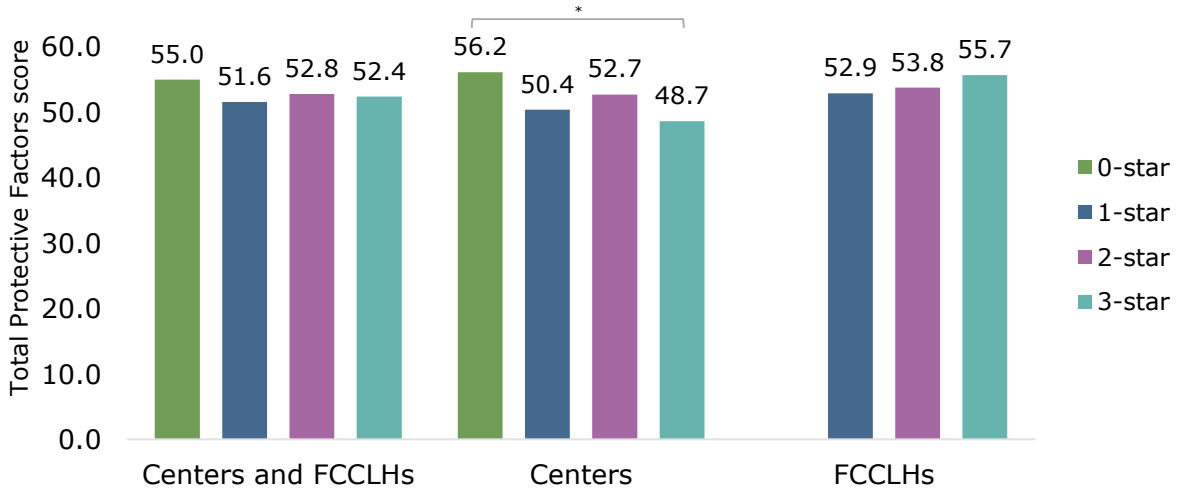
Social and emotional

Toddlers' social skills (DECA Total Protective Factors) did not vary by star rating (see Figure 16) when we combined center-based programs and FCCLHs. This finding was also true for toddlers attending FCCLHs when centers and FCCLHs were analyzed separately. However, toddlers attending 0-star center-based programs had higher social skills than those attending 3-star center-based programs. This finding was surprising because we expected higher-rated programs would be linked to stronger skills. However, this unexpected finding was not part of a larger pattern of higher ratings being related to lower skills; for this reason, we do not believe it represents a deeper issue with the rating system.

[k] Scores on the HTKS assessment tended to be positively skewed, as a large number of children scored 0 on the measure. To address this issue, we conducted a sensitivity analysis by analyzing a 4-level categorical version of the HTKS variable (0, 1-10, 11-30, 31-60) using an ordinal logistic regression. This analysis also yielded non-significant results. We conducted a second sensitivity analysis by limiting the sample to preschoolers who were at least 4 years old at post-test (in the spring). This sub-analysis also yielded non-significant results.

Figure 16. Adjusted means for toddlers' social skills, by setting and star rating

For the most part, there were no differences in toddlers' social skills across star rating.

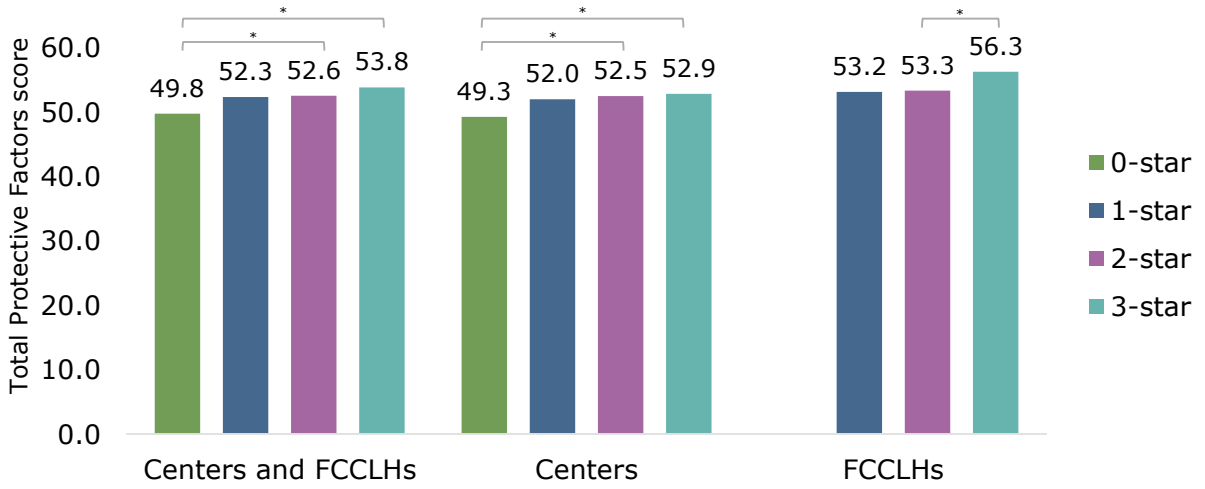


Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Preschoolers attending 3- and 2-star center-based programs and FCCLHs (combined) had stronger social skills (DECA Total Protective Factors) compared to those attending 0-star programs (see Figure 17). Differences between children in 1-star programs and those in any other group were not significant. This finding was consistent in center-based programs. In FCCLHs, preschoolers attending 3-star programs had stronger social skills than those attending 2-star programs. Unlike the findings for toddlers, these differences were all in the expected direction, with higher star ratings associated with stronger skills.

Figure 17. Adjusted means for preschoolers' social skills, by setting and star rating

Preschoolers attending 3- and 2-star programs had stronger social skills than those attending 0-star programs.

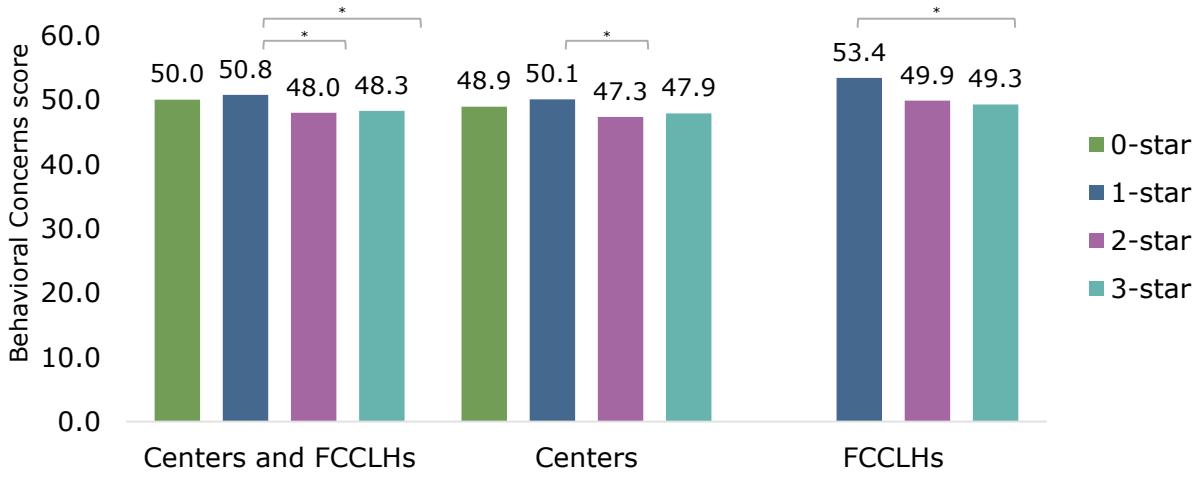


Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Teachers and providers reported fewer behavioral concerns (DECA Behavioral Concerns) for preschoolers attending 3- and 2-star programs compared to those attending 1-star programs (see Figure 18), when centers and FCCLHs were combined. This is in the expected direction because higher scores on this measure indicate more behavioral concerns. In center-based programs, teachers in 2-star programs reported fewer behavioral concerns for preschoolers than those in 1-star programs. In 3-star FCCLHs, providers reported fewer behavioral concerns for preschoolers than those in 1-star FCCLHs.

Figure 18. Adjusted means for preschoolers' behavioral concerns, by setting and star rating

Teachers reported fewer behavioral concerns for preschoolers attending 3- and 2-star programs than those attending 1-star programs.



Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

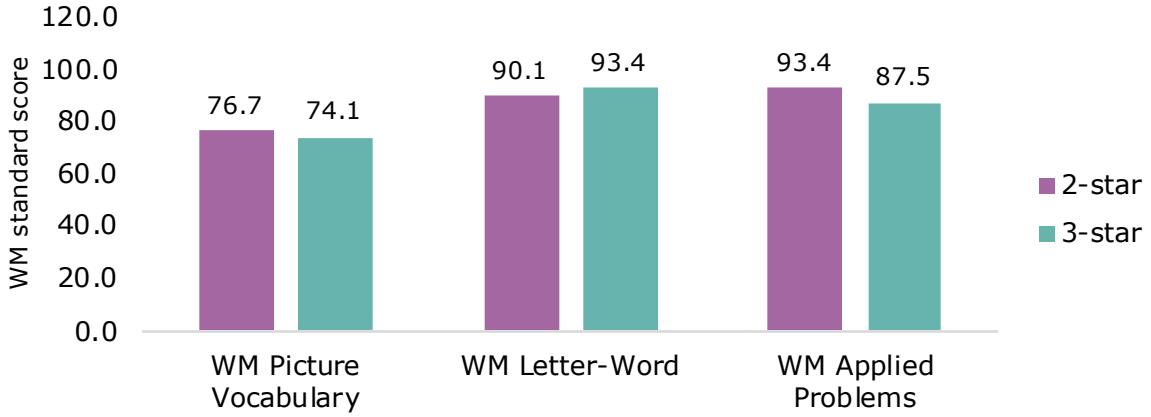
Spanish-speaking children

Children who spoke Spanish at home took an additional set of Spanish assessments including the Picture Vocabulary, Letter-Word Identification, and Applied Problems subtests of the WM-III, and the Spanish version of the Counting Bears task. Because the sample of Spanish-speaking children in lower-rated programs was very small, we were only able to compare Spanish-speaking children in 2- versus 3-star programs. The sample of Spanish-speaking children was also not large enough to examine children attending center-based programs and FCCLHs separately.

When comparing children in 2- and 3- star programs, there were no significant differences between Spanish-speaking preschoolers' expressive vocabulary (WM-III Picture Vocabulary), early literacy (WM-III Letter-Word Identification), or emerging math skills (WM-III Applied Problems; see Figure 19). Likewise, there were no differences in children's Spanish counting abilities (Counting Bears; see Figure 20).

Figure 19. Adjusted means for Spanish-speaking preschoolers' expressive vocabulary, early literacy, and early math skills, by star rating

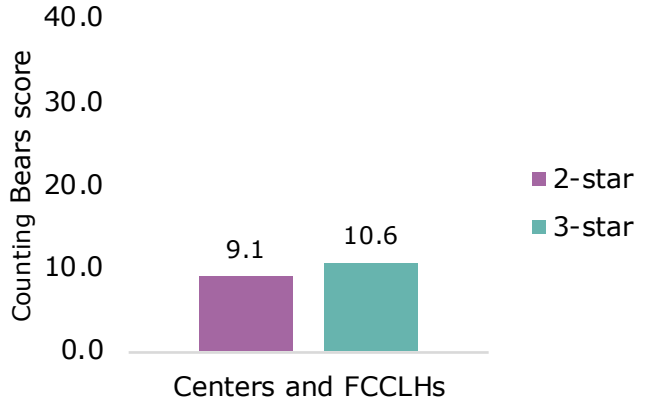
There were no differences in Spanish-speaking preschoolers' expressive vocabulary, early literacy, or early math skills in 2- versus 3-star programs.



Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Figure 20. Spanish-speaking preschoolers' counting abilities, by star rating

There were no differences in Spanish counting abilities across star 2- and 3-star programs.



Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Summary of children's development findings

For the most part, the number of stars a program earned was not significantly associated with children's academic and social development, with some exceptions. In early math, preschoolers in 3-star programs had higher scores at the end of the year than children in some of the lower-rated programs. For social skills and behavioral concerns, teachers reported stronger skills for children in both 2- and 3-star programs than those in lower-rated programs.

		1-star	2-star	3-star
Preschoolers	Early literacy	n.s.	n.s.	n.s.
	Expressive vocabulary	n.s.	n.s.	n.s.
	Early math	1>0	n.s.	3>0; 3>2
	Counting	n.s.	n.s.	n.s.
	Executive function	n.s.	n.s.	n.s.
	Social skills	n.s.	2>0	3>0
	Behavioral concerns	n.s.	2<1	3<1
Toddlers	Language acquisition	n.s.	n.s.	n.s.
	Expressive vocabulary	n.s.	n.s.	n.s.
	Social skills	n.s.	n.s.	n.s.

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

3 Are Quality Rated star ratings related to work climate?

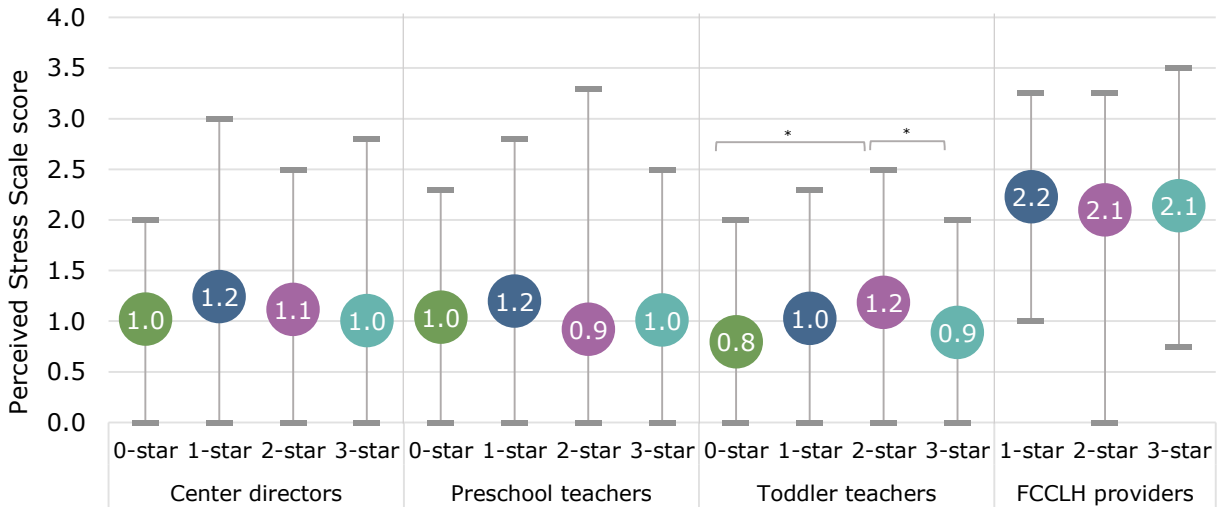
There was evidence that star rating was associated with work climate. Center-based programs with 2 or 3 stars generally had lower turnover, higher wages, and provided more benefits. There were no differences in level of perceived stress or job commitment by star rating among FCCLH providers. Details for each aspect of work climate appear below. For more details about the analytic methods used, see Appendix F.

Perceived stress

There were no significant differences among center directors', preschool teachers', or FCCLH providers' reported stress in programs with different star ratings (see Figure 21). The average stress of toddler teachers in 2-star center-based programs was higher than the average stress of toddler teachers in 0- or 3-star center-based programs, but none of the other groups were significantly different from one another.

Figure 21. Reported stress averages and ranges for center directors, teachers, and FCCLH providers by star rating

In general, stress levels did not vary by star rating.



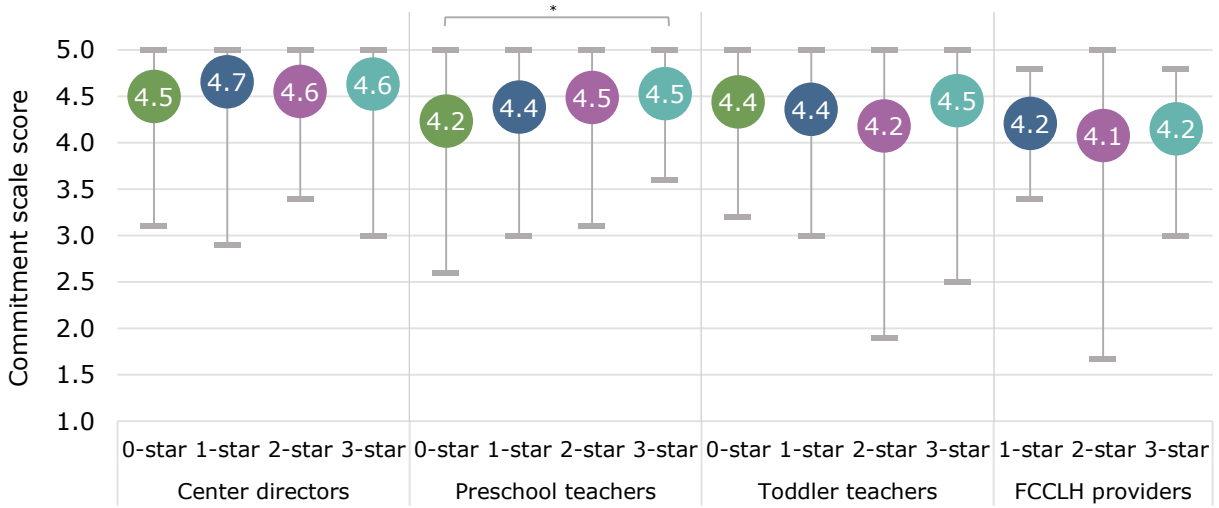
Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018; Child Trends' provider questionnaire, winter 2016–2017 and winter 2017–2018

Job commitment and teacher turnover

Center directors, preschool teachers, toddler teachers, and FCCLH providers were highly committed to their jobs, with averages over 4.0 out of 5.0 on the commitment scale (see Figure 22). Among center directors, toddler teachers, and FCCLH providers, there were no significant differences in the level of commitment in programs with different star ratings. Preschool teachers in 3-star programs were more committed to their jobs than preschool teachers in 0-star programs, but there were no other significant differences between groups.

Figure 22. Job commitment averages and ranges for center directors, teachers, and FCCLH providers by star rating

Job commitment was very high on average for all survey participants, and generally did not vary by star rating.

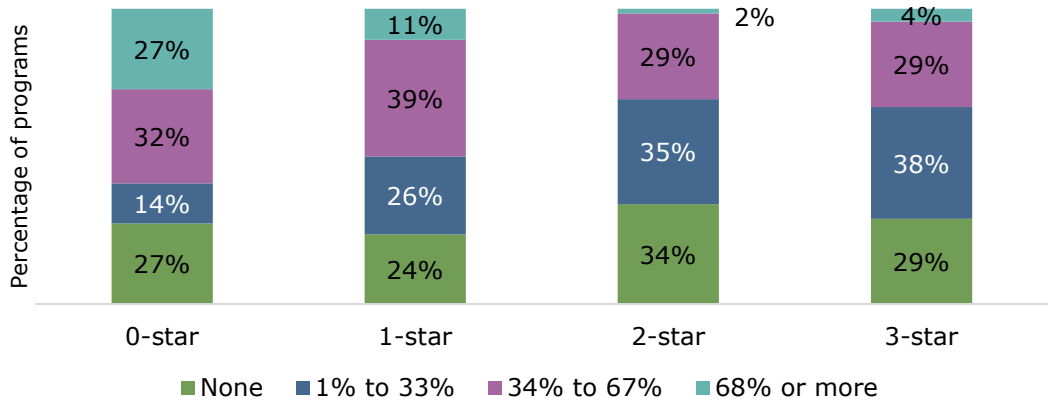


Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018; Child Trends' provider questionnaire, winter 2016–2017 and winter 2017–2018

Figure 23 shows the percentage of programs in each turnover category for lead teachers by star rating. Significantly more directors in 0-star programs reported that over two-thirds of their lead teachers had left in the past 12 months, and significantly fewer reported a rate from 1 to 33 percent, than directors in 2- or 3-star programs. There were no other differences between star ratings. As described in Appendix K, programs had an average of five to seven lead teachers across star ratings.

Figure 23. Percentage of programs with each percent turnover for lead teachers as reported by the center director, by star rating

Significantly more 0-star center-based programs fell into the highest category of lead teacher turnover than 2- or 3-star programs.

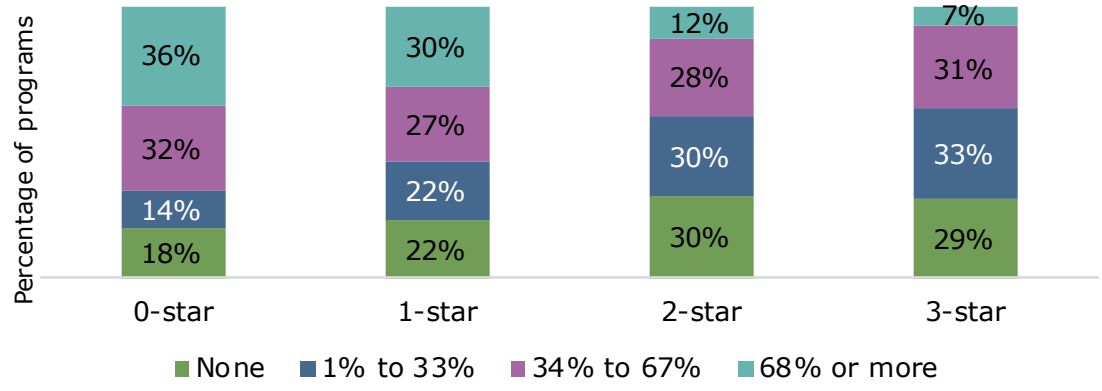


Source: Child Trends' director questionnaire winter 2017–2018

Figure 24 shows the percentage in each turnover category for assistant teachers by star rating. Significantly more directors in 0- and 1-star center-based programs reported that over two-thirds of their assistant teachers had left and needed to be replaced in the past 12 months than directors in 2- or 3-star programs. There were no differences within 0- and 1-star programs or 2- and 3-star programs. As described in Appendix K, programs had an average of four to six assistant teachers across star ratings.

Figure 24. Percentage of programs with each percent turnover for assistant teachers as reported by the center director, by star rating

Significantly more 0- and 1-star center-based programs reported the highest level of assistant teacher turnover than 2- or 3-star programs.



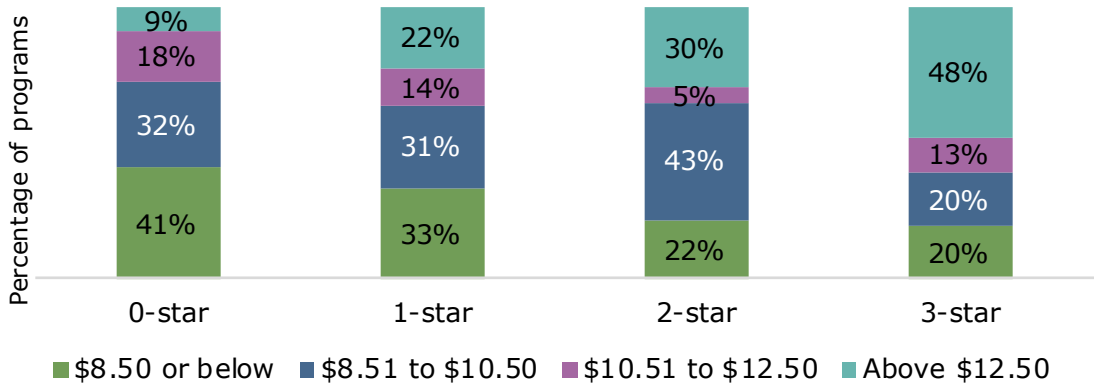
Source: Child Trends' director questionnaire winter 2017–2018

Entry-level hourly wages and benefits

Significantly more 3-star programs reported an entry-level hourly wage for preschool teachers over $12.50 than did 0- and 1-star programs (see Figure 25). In addition, a greater percentage of 2-star programs reported paying over $12.50 per hour than did 0-star programs. Significantly more 2-star programs than 3-star programs reported paying their entry-level preschool teachers an hourly wage from $8.51 to $10.50. There were no other differences between star ratings. See Appendix K for more detailed information on teacher pay.

Figure 25. Percentage of programs reporting ranges of hourly wages for an entry-level preschool teacher as reported by the center director, by star rating

Significantly more 3-star center-based programs paid their entry-level preschool teachers an hourly wage above $12.50 than did 0- or 1-star programs.

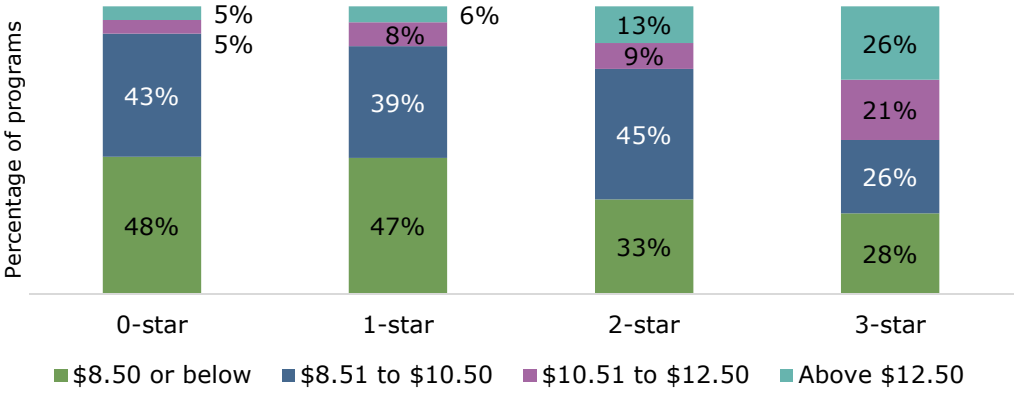


Source: Child Trends' director questionnaire winter 2017–2018

Significantly more 3-star center-based programs reported paying entry-level toddler teachers above $12.50 per hour than did 0- and 1-star programs (see Figure 26). More 3-star programs paid entry-level toddler teachers an hourly wage from $10.51 to $12.50 than did 0-star programs, and fewer 3-star programs paid entry-level toddler teachers an hourly wage from $8.51 to $10.50 than did 2-star programs. There were no other differences between star ratings.

Figure 26. Percentage of programs reporting ranges of hourly wages for an entry-level toddler teacher as reported by the center director, by star rating

Significantly more 3-star center-based programs paid their entry-level toddler teachers an hourly wage above $12.50 than did 0- or 1-star programs.



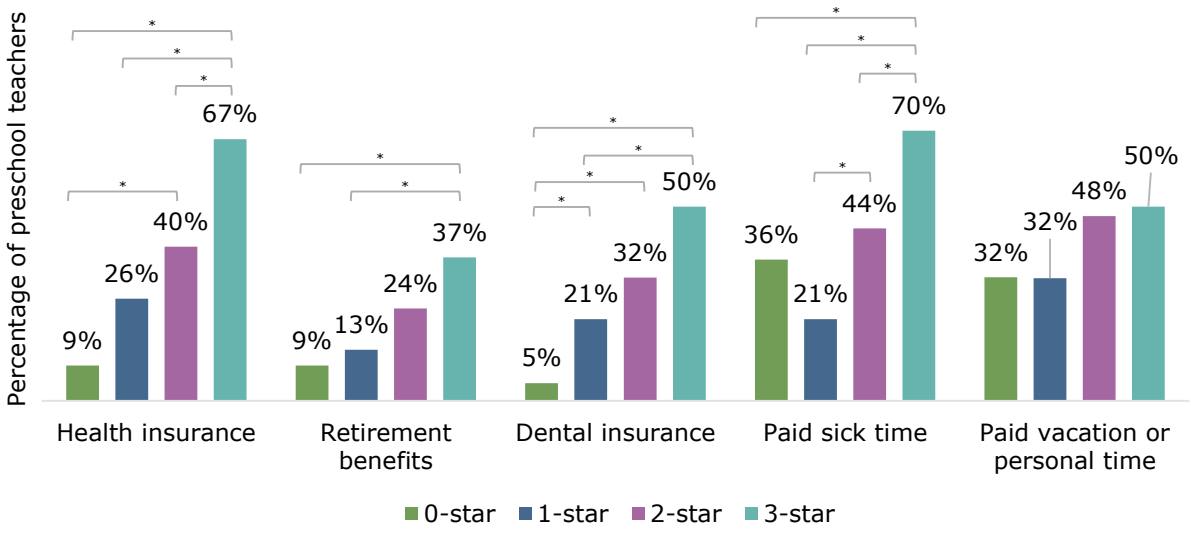
Source: Child Trends' director questionnaire winter 2017–2018

A majority of directors (range of 73% to 96%), preschool teachers (range of 55% to 96%), and toddler teachers (range of 62% to 81%) reported having at least one employment benefit. To examine differences across star ratings, we focused on the most commonly reported benefits for preschool teachers: paid sick leave, paid vacation or personal leave, health insurance, dental insurance, and retirement benefits. See Appendix L for more details about benefits for center directors and toddler teachers by star rating.

As seen in Figure 27, preschool teachers in higher-rated programs were generally more likely to have health insurance, retirement benefits, dental insurance, and paid sick leave, compared to those in lower-rated programs. The percentage of preschool teachers who had paid vacation or personal leave did not vary by star rating.

Figure 27. Percentage of preschool teachers who had health insurance, retirement benefits, and dental insurance by star rating

Preschool teachers in higher-rated programs were more likely to have benefits than those in lower-rated programs.



Source: Child Trends' teacher questionnaires, winter 2017–2018

Summary of work climate findings

Results indicated that center-based staff in 2- and 3-star programs tended to have a more positive work climate than staff in lower-rated programs This was particularly true for wages and employee benefits such as health insurance and paid sick leave. There were no significant differences in level of stress or job commitment by star rating among 1-, 2-, and 3-star FCCLH providers.

		1-star	2-star	3-star
Center-based directors	Perceived stress	n.s.	n.s.	n.s.
	Commitment	n.s.	n.s.	n.s.
	Lead teacher turnover of 68% or more	n.s.	2<0	3<0
	Assistant teacher turnover of 68% or more	n.s.	2<0, 2<1	3<0, 3<1
	Hourly preschool teacher wages of $12.50 or above	n.s.	2>0	3>1, 3>0
	Hourly toddler teacher wages of $12.50 or above	n.s.	n.s.	3>0, 3>1
Preschool teachers	Perceived stress	n.s.	n.s.	n.s.
	Commitment	n.s.	n.s.	3>0
	Health insurance	n.s.	2>0	3>2, 3>1, 3>0
	Retirement benefits	n.s.	n.s.	3>1, 3>0
	Dental insurance	1>0	2>0	3>1, 3>0
	Paid sick leave	n.s.	2>1	3>2, 3>1, 3>0
	Paid vacation or personal leave	n.s.	n.s.	n.s.
Toddler teachers	Perceived stress	n.s.	2>0; 2>3	n.s.
	Commitment	n.s.	n.s.	n.s.
FCCLH providers	Perceived stress		n.s.	n.s.
	Commitment		n.s.	n.s.

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Study Limitations

This study and its results have some limitations. Our power analysis indicated that we needed 50 programs at each star rating to ensure that we would detect differences if they existed. We met this goal for 1-, 2-, and 3- star levels when we combined center-based programs and FCCLHs; however, we did not meet it for 0-star programs or for some of the star ratings when center-based programs and FCCLHs were analyzed separately. Smaller than expected sample sizes may explain some of the non-significant findings, particularly between 0- and 1-star center-based programs. Additionally, because there were not enough 0-star FCCLHs to include in the analysis of quality or work climate, we could not test any differences between 0- and 1-star FCCLHs on those measures.

The smaller-than-desired sample sizes, especially at the 0- and 1-star ratings, resulted from two issues: lower response rates for 0- and 1-star programs as compared to 2- and 3-star programs, and the small numbers of programs from which to recruit. The small pool of programs was especially problematic for FCCLHs, of which there were only 28 0-star programs in Quality Rated. Our main motivation for collecting data from a second cohort of FCCLHs was to bolster our sample size. We also were unable to collect COP/TOP data from all the preschool classrooms in our study; the small sample size on those measures may have contributed to the lack of significant findings by star rating.

Another limitation is that programs that agreed to participate may have been qualitatively different from those not in the study. We compared our sample to Quality Rated programs overall using variables in the Quality Rated administrative data system and found few differences. However, it is possible that the study programs were different on qualities that are not in the data system, such as satisfaction with Quality Rated or commitment to quality improvement. Although the response rate in this study was similar to that seen in other QRIS validation studies, it is not possible to know the extent to which the sample at each star rating was representative of all programs at that level, or whether nonparticipants were systematically different from participants.

Our decision to use LENA to capture the language environment also had some limitations. First of all, LENA was originally designed to record speech during one-on-one interactions between caregivers and children. Our study was one of the first in which LENA was worn by adults to capture speech in child care settings. We conducted a series of reliability checks to ensure the device was adequately capturing adult speech. Although the checks indicated that the data were suitable for analysis, they appear to contain more error than we would like. Further, there is a lack of research about which variables derived from the recordings are most important for measuring children's language development. Finally, some teachers and providers declined to wear the LENA device, and there were some technical difficulties in using it; therefore, we were not able to collect and analyze LENA data from all the participants in our sample.

Discussion of Key Findings

This study investigated the extent to which Georgia's Quality Rated star ratings were associated with independent measures of classroom quality, children's growth, and work climate. We found mixed evidence for these associations. This section summarizes the key findings and discusses them in the context of other studies.

Key Finding 1: Center-based programs and FCCLHs with the highest Quality Rated star rating (three stars) were generally of higher quality than lower-rated programs. In particular, preschool and toddler classrooms in 3-star center-based programs had higher-quality teacher-child interactions than lower-rated programs, as measured by the CLASS. In FCCLHs, 3-star programs had higher-quality provider-child interactions than 2-star, but not 1-star, programs. Toddler teachers and FCCLH providers in 3-star programs also offered richer language environments, as measured by LENA, than those in lower-rated

programs. Each of these differences was large enough to be considered substantively important by What Works Clearinghouse (2014).

The general association between the CLASS and Quality Rated is consistent with other validation studies. Of the seven states that used the CLASS Pre-K as an independent measure of quality, five found relationships between program ratings and preschool classroom quality (Tout et al., 2017). Of the three states that used the CLASS Toddler, two reported relationships between program ratings and toddler classroom quality. Due to limitations in sample size, some other states' validation studies were not able to test differences between each star rating. Instead they either compared groups of ratings (e.g., levels 1 and 2 vs. levels 3, 4, and 5) or looked at general trends across ratings. The Georgia validation study design allowed us to address specific questions about differences between each rating, and the findings suggest that programs at the 3-star rating are of higher quality, as measured by the CLASS, than programs at lower star ratings. Generally, we did not find evidence that the rating differentiates 0-, 1-, and 2-star programs from one another on the CLASS.

Key Finding 2: We did not find evidence of differences at every level of star rating or on every independent measure of quality. Although there seemed to be a general pattern of 3-star (and sometimes 2-star) programs' being of higher quality, some of the findings were unexpected and showed that there are inconsistent relationships between the ratings and other measures. For instance, although 3-star FCCLHs had higher CLASS scores than 2-star FCCLHs, there were no differences between 3- and 1-star FCCLHS on the CLASS. Toddler teachers in 2-star centers used more sophisticated vocabularies than those in 0- or 3-star centers.

The evidence for links between star rating and independent measures of quality other than CLASS in preschool classrooms was limited. The richness of the language environment did not vary by star rating in preschool classrooms. The only other QRIS validation study that used LENA to capture the language environment was in the state of Washington, and they did not report on LENA's relationship with star rating (Soderberg, Joseph, Stull, & Hassairi, 2016). As a point of reference, in this sample, adult word count (from 41.6 to 60.2 for all groups except 3-star FCCLHs, with a value of 72.2) and average length of utterances (from 3.8 to 4.7) were generally lower than those seen in a sample of Head Start teachers (69.1 and 6.5, respectively; Dickinson, Hofer, Barnes, & Grifenhagen, 2014), but adult word count was relatively consistent with that seen in Washington (48.9; Soderberg et al., 2016).

There was little evidence that child and teacher behaviors in preschool classrooms as measured by the COP/TOP varied by star rating. We chose to use this relatively new tool because past research showed a strong relationship between the behaviors it measures and gains across a range of developmental outcomes in three "model" public pre-K programs in Tennessee (Farran et al., 2017). Additional research is needed to understand the extent to which this measure is consistently related to children's development.

Key Finding 3: Preschool children in higher-rated programs learned more than children in lower-rated programs in some, but not all, domains. Preschoolers in 3-star programs had stronger math and social skills at the end of the school year than their peers in lower-rated programs, after accounting for their skills at the start of the school year; for the most part, these differences were large enough to be considered substantively important by What Works Clearinghouse (2014). The number of stars a program earned was not associated with preschoolers' expressive vocabulary, early literacy, or executive function skills, nor with toddlers' development in language or social skills. These results are consistent with those in the QRIS validation synthesis (Tout et al., 2017; see text box on page 35 for more details).

Collection of data about infants and toddlers was a strength of this study; of the seven states in the QRIS synthesis, three included toddlers, and only one included infants. However, gathering data about the developing skills of such young children poses challenges. To study these young children, the research team chose to rely on teacher or provider report of skills rather than conducting one-on-one

assessments as was done with preschoolers. Teachers and providers likely vary in their knowledge of social-emotional or language development; consequently, they may have different levels of understanding about what each question on the assessment was asking. Additionally, teachers with a more nuanced understanding of social-emotional or language development may be more realistic about children's skills, and these more knowledgeable teachers may rate children as less advanced. Teachers or providers who have had more training in observing and assessing children's development may also have completed the forms differently. These confounding variables could partially explain the lack of findings on toddler outcomes.

The limited associations between QRIS ratings and children's outcomes are not surprising. As noted above, QRIS validation studies in other states reported mixed findings regarding the associations between QRIS ratings and children's outcomes. The first report in this series of Quality Rated Validation reports indicated that the star rating was almost entirely determined by the program's average ERS score (Early, et al., 2017). Past research has found somewhat inconsistent and small associations between measures of classroom quality, such as ERS, and measures of children's development (Burchinal, 2017). Thus, the limited associations between Quality Rated and children's development are likely due to the fact that ERS is inconsistently related to children's outcomes.

Comparison of findings to other state QRIS validation studies

In 2017, the Office of Planning, Research, and Evaluation partnered with Child Trends to produce a synthesis of QRIS validation studies from 10 states (Tout et al., 2017). We present this information to help contextualize the current findings. It should be noted, however, that this table simplifies complex findings from other states, and that each state had different methods, definitions, and ways of combining groups of children and types of programs.

Table 4. Comparison of the validation synthesis findings to the current report

	QRIS Validation Synthesis	Were results found in this study?
Quality outcomes	5 out of 7 studies found a relationship between ratings and CLASS Pre-K	Yes
	2 out of 3 studies reported relationships between ratings and CLASS Toddler	Yes
Preschool outcomes	4 out of 6 studies found an association between ratings and children's social and emotional development	Yes
	1 out of 6 studies found an association between ratings and math skills	Yes
	2 out of 7 studies found an association between ratings and children's language and literacy outcomes	No
Toddler outcomes	0 out of 3 studies found an association between ratings and children's social and emotional development	No
	0 out of 1 study found an association between ratings and math skills	Did not measure
	1 out of 2 studies found an association between ratings and children's language and literacy outcomes	No

Source: Tout et al., 2017 and Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Key Finding 4: In center-based programs with higher star ratings, the work climate was better in terms of turnover, wages, and employee benefits. Work climate is an important but often overlooked aspect of program quality, so it is encouraging that star rating meaningfully differentiated programs in this way. For example, directors in 2- and 3-star programs reported lower rates of teacher turnover than those in 0-star programs. In 2012, half of a nationally representative sample of center directors reported zero turnover of lead teachers over the past 12 months (Whitebook, Philips, & Howes, 2014), compared to 24 to 34 percent in this study. It is promising, however, that turnover is lower in higher-rated programs.

Likewise, in higher-rated center-based programs, the entry-level hourly wage was more likely to be over $12.50, and staff were more likely to receive benefits than in lower-rated programs. Ensuring that employees are retained and treated fairly is important for the well-being of staff (Whitebook & Sakai, 2003), the quality of the program (U.S. Department of Education, 2016), and the development of the children they teach (Whitebook, Philips, & Howes, 2014).

The findings also indicate that significant challenges remain with respect to teachers' wages. Even in 3-star programs, fewer than half of starting preschool teachers and just over one-quarter of starting toddler teachers made more than $12.50 per hour. Researchers at the Massachusetts Institute of Technology (Glasmeier, 2017) found that a living wage in Georgia is $11.93 per hour for an adult with no children and $24.00 per hour for an adult with one child. Additionally, FCCLH providers at all star ratings reported higher stress than center-based staff. Although we did not ask FCCLH providers about their hourly salaries, research shows that low wages contribute to stress for FCCLH providers (Porter et al., 2010). DECAL has started to address this important aspect of quality by developing initiatives that encourage career advancement through salary bonuses, such as INCENTIVES,[l] but this study indicates that work is still needed in this area.

Linking star rating to work climate was a strength of this study. Although a few QRIS validation studies in other states gathered information on aspects of work climate, such as teacher turnover (e.g., California, Rhode Island), they did not examine the relationship between ratings and work climate. The Georgia validation study makes an important contribution to the field's understanding of the relationship between star ratings and work climate.

Future Considerations

This section offers considerations, based on the research findings, to help Georgia leaders further strengthen the Quality Rated system.

Continue current revisions to the rating system. The findings from this fourth validation report suggest that differences among rated programs in observed quality and children's development are most evident when comparing 3-star programs to those with lower ratings. Findings from the first validation report suggest that the rating is driven almost entirely by the ERS observation, and that the portfolio makes a minimal contribution to the rating. DECAL has already started working with stakeholders to revise the portfolio portion of the rating. DECAL can use the work climate findings from this validation report to identify key standards to consider in the portfolio and possible thresholds at each star rating. Further, DECAL is moving toward use of the Infant/Toddler Environment Rating Scale-Third Edition (ITERS-3) and the Family Child Care Environment Rating Scale-Third Edition (FCCERS-3), in place of their predecessors, ITERS-Revised and FCCERS-Revised. These third editions place increased emphasis on interactions. We hope that the findings from this fourth validation report will support DECAL's revisions to ensure that the ratings better differentiate quality between 1- and 2-star and 2- and 3-star programs for preschoolers as well as infants and toddlers.

[l] For more details about the INCENTIVES program, see: https://www.decalscholars.com/pages/inc_landing.cfm.

Continue to invest in quality improvement. Validation studies, by definition, focus on the rating. Quality Rated, however, is more than just a rating. A QRIS has multiple components: standards that define quality across levels, a process for monitoring and rating programs, quality improvement supports, financial incentives to improve and maintain quality, and consumer education to inform families about quality and ratings (National Center on Early Childhood Quality Assurance, n.d.). Although this report indicates that revisions to the rating system may be needed to better differentiate levels of quality, it is important to remember that making the rating more accurate will not, by itself, improve quality. If DECAL's leaders are interested in supporting children's literacy and language development, for instance, they will need to offer evidence-based professional development in that area. Changing the rating system by adding language and literacy standards might increase awareness, but it is unlikely to significantly change practice in and of itself.

An ongoing consideration for DECAL will be how to improve program quality once programs have joined Quality Rated. Given the high level of job commitment expressed by study participants across settings and star ratings, Quality Rated staff may be especially interested in professional development opportunities to strengthen teachers' skills. Report #3 (Early et al., 2018) of this series described the training and technical assistance that is available to Quality Rated programs. A large percentage of center directors (85%) and FCCLH providers (78%) reported using technical assistance from their child care resource and referral (CCR&R) agency. DECAL leaders may find it useful to review the technical assistance offered through CCR&Rs to ensure that the technical assistance is evidence based and focused on improving quality at the classroom as well as program levels. They can also build on their previous experiences with quality improvement initiatives as they develop or refine the technical assistance provided to Quality Rated programs. For example, DECAL has experience providing CLASS-specific technical assistance to improve teacher-child interactions in Georgia's Pre-K Program (Early et al., 2014). If DECAL wants to strengthen teacher-child interactions across Quality Rated programs, it might be possible to offer similar supports.

We encourage Georgia to continue its focus on quality improvement for infants and toddlers. There is growing evidence regarding the importance of supporting infant and toddler development, and the findings from this study indicated that the quality of center-based care for toddlers was in the low- to mid-range. That finding is consistent with past research indicating that quality of infant and toddler care tends to be low (Maxwell et al., 2009; Mulligan & Flanagan, 2006). Georgia already has several initiatives to expand high-quality early learning for infants and toddlers. For instance, Georgia's Early Head Start—Child Care Partnership grant aligns the strengths of Early Head Start with those of Quality Rated to improve programs for infants and toddlers, including those in FCCLHs. DECAL's Lifting Infants and Toddlers Through Language Rich Environments (LITTLE) Grants are designed to support language and literacy instruction in infant and toddler classrooms throughout Georgia by providing on-site coaching, professional learning opportunities, and materials (Georgia Department of Early Care and Learning, n.d.b). LITTLE Grants began in center-based programs in 2017 and expanded to FCCLHs in 2019. Quality Rated Subsidy Grants are supporting improved quality of infant and toddler care by offering higher subsidies for children receiving CAPS scholarships in programs meeting higher quality standards.

Continue efforts to improve compensation of the early care and education workforce. Compensation is related to quality, and it is one of five essential early childhood workforce policies delineated by the Center for Study of Child Care Employment (Whitebook, McLean, Austin, & Edwards, 2018). In the current study, teachers in higher-rated programs were paid more than teachers in lower-rated programs. Compensation was low, however, across all star ratings. With such low levels of compensation, teachers may leave for higher-paying positions. High turnover among teaching staff may make it difficult to sustain quality improvement efforts aimed at teachers because teachers who participate in quality improvement may leave to take a higher-paying position before new quality practices are implemented. In part for this reason, some more-recent quality

improvement strategies aim to foster a continuous learning environment at the program level rather than focusing solely on a single teacher within a program (e.g., Daily et al., 2018; Pacchiano, Klein, & Hawley, 2016; Young, 2017).

Compensation and retention strategies may also be useful in supporting quality. DECAL currently supports the INCENTIVES program, which provides bonuses to teachers who meet certain education and tenure criteria. A few studies have evaluated these types of bonus programs in other states, and the findings generally indicate an association with lower turnover (Shaw et al., 2019). DECAL has made strides in improving compensation for Georgia's Pre-K teachers who work in private child care so that they are paid comparably to pre-K teachers in public schools (Suggs, 2017). Lessons learned from that effort might also be useful in refining compensation strategies for non-pre-K teachers.

Compensation should be part of any long-term initiative to support high-quality early care and education programs. Research on the importance of brain development (e.g., National Scientific Council on the Developing Child, 2007) and the complexity of providing developmentally appropriate instruction for young children (e.g., Board on Children, Youth, and Families, 2015) underscores the need for a qualified workforce. Addressing compensation, therefore, is key in attracting and maintaining a workforce of teachers who understand child development and can provide instructional activities and supports within the context of positive relationships.

Continue to focus on all areas of children's development. The findings from this fourth validation report demonstrated that quality ratings were associated with some, but not all, areas of children's development. Domains of development are interconnected, especially in young children, so we encourage Georgia leaders to consider strategies for supporting all areas of development rather than focusing heavily on a single area, like language and literacy. Indeed, there is some evidence that early math skills are better predictors of later reading and math achievement than early reading skills, although early reading is important as well (Duncan et al., 2007). This type of cross-domain learning reminds us that children need high-quality experiences in all areas, and that focusing narrowly on a single domain is unlikely to be the best strategy for supporting overall learning.

References

Bilbrey, C., Vorhaus, E., Farran, D., & Shufelt, S. (2007). *Teacher Observation in Preschool* [Measurement Instrument]. Unpublished instrument. Nashville, TN: Peabody Research Institute, Vanderbilt University.

Board on Children, Youth, and Families (2015). *Child development and early learning: A foundation for professional knowledge and competencies*. Washington, DC: The National Academies of Sciences, Engineering, and Medicine.

Burchinal, M. (2017). Measuring early care and education quality. *Child Development Perspectives, 12*(1), 3-9. doi: 10.1111/cdep.12260

Cohen, S., & Janicki-Deverts, D. (2012) Who's stressed? Distributions of psychological stress in the United States in probability samples from 1983, 2006 and 2009. *Journal of Applied Social Psychology, 42*(6), 1320-1334. doi: 10.1111/j.1559-1816.2012.00900.x

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385-396.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*, 386-396.

Daily, S., Tout, K., Douglass, A., Miranda, B., Halle, T., Agosti, J., … Doyle, S. (2018). *Culture of Continuous Learning Project: A literature review of the Breakthrough Series Collaborative (BSC)* (OPRE Report #2018-28). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Dickinson, D. K., Hofer, K. G., Barnes, E. M., & Grifenhagen, J. F. (2014). Examining teachers' language in Head Start classrooms from a Systemic Linguistics Approach. *Early Childhood Research Quarterly*, *29*(3), 231-244.

Duncan, G. J. Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement., *Developmental Psychology, 43*(6), 1428-1446.

Early Childhood Learning & Knowledge Center (2013). *2013 Head Start grantee-level data from the Classroom Assessment Scoring System (CLASS).* Retrieved from https://eclkc.ohs.acf.hhs.gov/publication/2013-head-start-grantee-level-data-classroom-assessment-scoring-system-classr

Early, D. M., Maxwell, K. L., Orfali, N. S., & Li, W. (2017). *Quality Rated validation study report #1: What makes up a Quality Rated star rating? An in-depth look at the criteria, standards, and components.* Chapel Hill, NC: Child Trends.

Early, D. M., Orfali, N. S., Maxwell, K. L., Bultinck, E., Nugent, C., Mason, R., … Bingham, G. (2018). *Quality Rated validation study report #3: Director, teacher, and provider perceptions of Quality Rated.* Chapel Hill, NC: Child Trends.

Early, D. M., Maxwell, K. L., Skinner, D., Kraus, S., Hume, K., & Pan, Y. (2014). *Georgia's pre-k professional development evaluation: Final report.* Chapel Hill, NC: University of North Carolina at Chapel Hill, FPG Child Development Institute.

Epstein, D., Hegseth, D., Friese, S., Miranda, B., Gebhart, T., Partika, A., & Tout, K. (2017). *Quality First: Arizona's early learning Quality Improvement and Rating System implementation and validation study.* Chapel Hill, NC: Child Trends.

Farran, D. C., Meador, D., Christopher, C., Nesbitt, K. T., & Bilbrey, L. E. (2017). Data-driven improvement in prekindergarten classrooms: Report from a partnership in an urban district. *Child Development, 88*(5), 1466-1479.

Farran, D. C., Plummer, C., Kang, S., Bilbrey, C., & Shufelt, S. (2006). *Child Observation in Preschool* [Measurement Instrument]. Unpublished instrument. Nashville, TN: Peabody Research Institute, Vanderbilt University.

Fenson, L., Pethick, S., Renda, C., & Cox, J. L. (2000). Short form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics, 21,* 95–116.

Forry, N., Iruka, I., Tout, K., Torquati, J., Susman-Stillman, A., Bryant, D., & Daneri, M.P. (2013). Predictors of quality and child outcomes in family child care settings. *Early Childhood Research Quarterly, 28,* 893-904.

Georgia Department of Early Care and Learning (n.d.a). About Georgia's Pre-K Program. Retrieved from http://decal.ga.gov/Prek/About.aspx

Georgia Department of Early Care and Learning (n.d.b). Lifting Infants and Toddlers Through Language-Rich Environments (LITTLE) grants. Retrieved from http://decal.ga.gov/InstructionalSupports/EarlyLanguageandLiteracy.aspx

Gilkerson, J., Richards, J. A., Greenwood, C. R., & Montgomery, J. K. (2016). Language assessment in a snap: Monitoring progress up to 36 months. *Child Language Teaching and Therapy, 33*(2), 99-115.

Glasmeier, A. (2017). Living wage calculation for Georgia. Retrieved from http://livingwage.mit.edu/states/13

Halle, T.G., Hair, E.C., Burchinal, M., Anderson, R., & Zaslow, M. (2012). *In the running for successful outcomes: Exploring the evidence for thresholds of school readiness*. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services.

Heilmann, J., Nockerts, A., & Miller, J. (2010). Language sampling: Does the length of the transcript matter? Language, Speech and Hearing Services in Schools, 41(4), 393-404.

Henley, J.R., & Adams, G. (2018). *Increasing access to quality care for four priority populations: Challenges and opportunities with CCDBG reauthorization.* Washington, DC: Urban Institute.

Isner, T. K., Tout, K., Zaslow, M., Soli, M., Quinn, K., Rothenberg, L., & Burkhauser, M. (2011). *Coaching in early care and education programs and quality rating and improvement systems (QRIS): Identifying promising features.* Washington, DC: Child Trends.

Jorde-Bloom, P. (1988). Factors influencing overall job satisfaction and organizational commitment in early childhood work environments. *Journal of Research in Childhood Education, 3*(2), 107-122. doi: 10.1080/02568548809594933

Kemper, S., & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging, 16*(2), 312–322.

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

LeBuffe, P.A., & Naglieri, J.A. (2012). *The Devereux Early Childhood Assessment for Preschoolers, second edition (DECA-P2) assessment, technical manual, and user's guide* [Measurement Instrument]. Lewisville, NC: Kaplan.

Lipscomb, S. T., Weber, R. B., Green, B. L., & Patterson, L. B. (2017). *Oregon's Quality Rating Improvement System (QRIS) validation study one: Associations with observed program quality.* Oregon: Oregon State University, Portland State University.

Mackrain, M., LeBuffe, P.A., & Powell, G. (2007). *The Devereux Early Childhood Assessment for Toddlers (DECA-T) assessment, technical manual, and user's guide* [Measurement Instrument]. Lewisville, NC: Kaplan.

Maxwell, K. L., Blasberg, A., Early, D. M., Li, W., & Orfali, N. (2016). *Evaluation of Rhode Island's BrightStars child care center and preschool quality framework.* Chapel Hill, NC: Child Trends.

Maxwell, K. L., Early, D. M., Bryant, D., Kraus, S., Hume, K., & Crawford, G. (2009). *Georgia study of early care and education: Child care center findings.* Chapel Hill, NC: The University of North Carolina at Chapel Hill, FPG Child Development Institute.

Miller, J., Andriacchi, K., & Nockerts, A. (2011*). Assessing language production using SALT software: A clinician's guide to language sample analysis.* Middleton, WI: SALT Software LLC.

Mulligan, G.M., & Flanagan, K.D. (2006). *Age 2: Findings from the 2-year-old follow-up of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)* (NCES Report No. 2006-043). Washington, DC: National Center for Education Statistics.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005). *Bateria III Woodcock-Muñoz* [Measurement Instrument]. Rolling Meadows: IL: Riverside Publishing

National Center for Early Development and Learning (2001). *Identifying Letters, Identifying Numbers, and Counting.* Unpublished.

National Center on Early Childhood Quality Assurance (n.d.). About QRIS. Retrieved from https://qrisguide.acf.hhs.gov/about-qris.

National Scientific Council on the Developing Child (2007). *The timing and quality of early experiences combine to shape brain architecture: Working paper #5.* Boston, MA: Center on the Developing Child, Harvard University.

Orfali, N. S., Early, D. M., & Maxwell, K. L. (2018). *Quality Rated validation study report #2: a further look at the programs in Quality Rated.* Chapel Hill, NC: Child Trends.

Pacchiano, D., Klein, R., & Hawley, M.S. (2016). *Reimagining instructional leadership and organizational conditions for improvement: Applied research transforming early education.* Chicago, IL: Ounce of Prevention Fund.

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.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139), 1236-1238. doi: 10.1136/bmj.316.7139.1236

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral regulation and its contributions to kindergarten outcomes. *Developmental Psychology, 45*, 605-619.

Porter, T., Paulsell, D., Del Grosso, P., Avellar, S., Hass, R., & Vuong, L. (2010). *A review of the literature on home-based child care: Implications for future directions.* Princeton, NJ: Mathematica Policy Research.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*(1), 19-35.

Schrank, F.A., McGrew, K.S., Mather, N., & Woodcock, R.W. (2014). *Woodcock-Johnson IV* [Measurement Instrument]. Rolling Meadows, IL: Riverside Publishing.

Shaw, S., Hilty, R., Lloyd, C., Nagle, K., Paschall, K., Warner-Richter, M., … Tout, K. (2019). *Evaluation of R.E.E.T.A.I.N.: Minnesota's child care workforce retention program – Final report* (DHS Report No. 7809A 1-19). Minneapolis, MN: Child Trends for the Minnesota Department of Human Services.

Soderberg, J. (2014). *Differential Benefit: Preschool Children, Quality of Early Childhood Education Environment and Developmental Gains Important for School Readiness* (Unpublished doctoral dissertation). University of Washington, Washington.

Soderberg, J., Joseph, G. E., Stull, S., & Hassairi, N. (2016). *Early Achievers standards validation study: Final report.* Seattle, WA: Childcare Quality and Early Learning Center for Research & Professional Development, College of Education, University of Washington.

Suggs, C. (2017). *Overview: 2018 fiscal year budget for lottery-funded programs.* Atlanta, GA: Georgia Budget and Policy Institute.

The Build Initiative & Child Trends. (2017). *A Catalog and Comparison of Quality Initiatives* [Data System]. Retrieved from http://qualitycompendium.org/ on December 10, 2018.

Todd, C. M., & Deery-Schmitt, D. M. (1996). Factors affecting turnover among family child care providers: A longitudinal study. *Early Childhood Research Quarterly*, *11*(3), 351-376.

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

Tout, K., Magnuson, K. Lipscomb, S., Karoly, L, Starr, R., Quick H., ... & Wenner, J. (2017). *Validation of the Quality Ratings used in Quality Rating and Improvement Systems (QRIS): A synthesis of state studies* (OPRE Report No. 2017-92). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

U.S. Department of Education. (2016). *High-quality early learning settings depend on a high-quality workforce: Low compensation undermines quality.* Washington, DC.

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

Whitebook, M., McLean, C., Austin, L.J.E., & Edwards, B. (2018). *Early childhood workforce index – 2018.* Berkeley, CA: Center for the Study of Child Care Employment, University of California, Berkeley.

Whitebook, M., Phillips, D., & Howes, C. (2014). *Worthy work, STILL unlivable wages: The early childhood workforce 25 years after the National Child Care Staffing Study.* Berkeley, CA: Center for the Study of Child Care Employment, University of California, Berkeley.

Whitebook, M., & Sakai, L. (2003). Turnover begets turnover: An examination of job and occupational instability among child care center staff. *Early Childhood Research Quarterly*, *18*(3), 273-293.

Xu, D., Yapanel, U., & Gray, S. (2009*). Reliability of the LENA language environment analysis system in young children's natural home environment technical report* (Technical Report No. 05-02). Boulder, CO: LENA Foundation.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T. ... Zaslow, M. J. (2013). *Investing in our future: the evidence base on preschool education*. Washington, DC: Society for Research in Child Development & Foundation for Child Development.

Young, B. (2017). *Continuous quality improvement in early childhood and school age programs: An update from the field.* Boston, MA: BUILD Initiative.

APPENDICES

Appendix A: Detailed program sampling and recruitment

Center-based programs

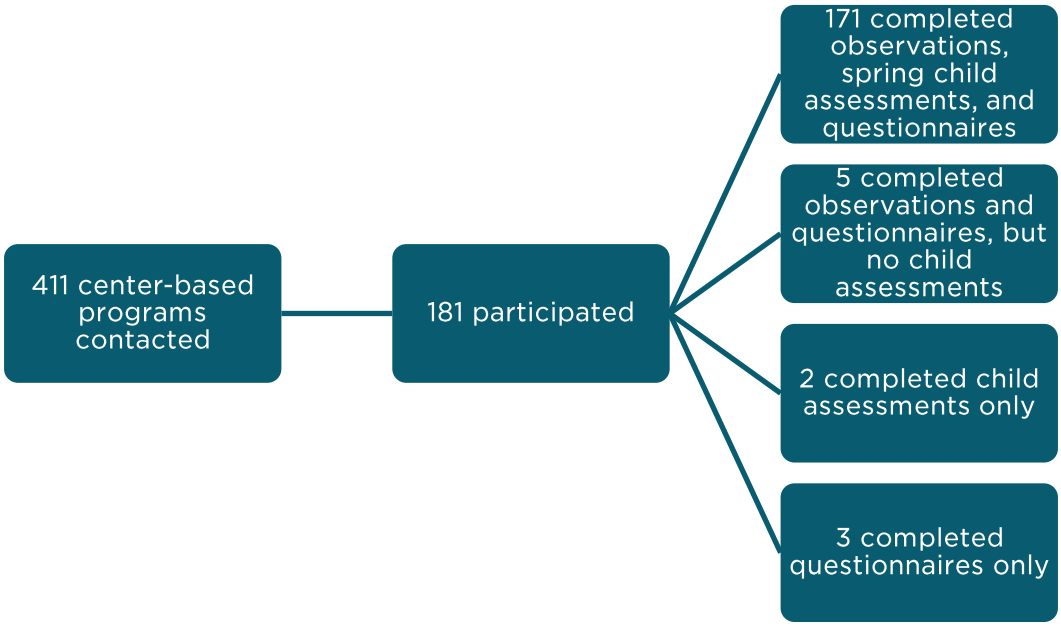
In center-based programs, data collection took place during a single school year (2017-18), and recruitment took place from July to October 2017. The total number of center-based programs in Quality Rated was large (*n* = 1,140 at the time of recruitment), so a randomly selected sample was invited to join the study. Based on a power analysis,[1] we aimed to recruit a random sample of 50 center-based programs at each star rating, including 1-, 2-, and 3-star, and programs that completed the rating process but did not meet the criteria for a star, which we refer to as *0-star*.

We created four lists of programs, one for each star rating, ordered randomly. We contacted programs on each list, starting at the top and continuing until 50 programs had agreed to participate or all programs had been contacted. When we exhausted the lists at the lower ratings, we contacted additional programs at the higher ratings. In total, we contacted 411 center-based programs, and 181 (44%) agreed to participate. We did not meet the goal of recruiting 50 programs at each star rating because the total number of center-based programs in Quality Rated is small at some star ratings, and many programs declined to participate. See Table A1 for response rates by star rating. The overall response rate was in the mid-range of response rates seen in other QRIS validation studies. Tout et al. (2017) reviewed reports from nine states and found that response rates ranged from 25 to 73 percent, with a median of 44 percent. See Figure A1 for details about which programs completed which portions of the data collection effort.

[1] The power analysis was based on assumptions of (1) power of 0.8; (2) significance level of 0.05; and (3) CLASS standard deviations of 0.34 for Emotional Support, 0.43 for Classroom Organization and 0.50 for Instructional Climate. These standard deviations were selected based on a study of Head Start programs across the country (Early Childhood Learning & Knowledge Center, 2013). This power analysis indicated that with 50 programs at each star rating, we could detect CLASS point differences of 0.23 for Emotional Support, 0.29 for Classroom Organization, and 0.34 for Instructional Support. In other words, if programs at different star ratings had CLASS scores that differed by at least 0.23 to 0.34 points, a sample of 50 programs would be sufficient to detect that difference.

Figure A1. Participation of center-based programs in the study

Center-based programs participated at a rate typical for QRIS validation studies.



Source: Validation study team data collection in center-based programs, 2017-18 school year

Within each participating center, up to two classrooms (one serving preschoolers and one serving toddlers) were recruited to participate. If there was only one preschool or toddler classroom at the center, that was the classroom we selected. If there were multiple, one was selected at random. In programs where the classrooms served mixed ages, preference was given to the classroom that served the most children in our target age ranges (3- and 4-year-old children in preschool classrooms and children ages 18-36 months in toddler classrooms). We did not distinguish between classrooms that were and were not part of Georgia's Pre-K. Overall, we recruited 180 classrooms serving preschoolers and 152 classrooms serving toddlers. Teachers received a $50 gift card for participating in each component of the study: fall child assessments, winter observations and surveys, and spring child assessments. Directors received a $50 gift card for completing the survey and supporting study activities.

To recruit children to take part in the study, packages of consent forms with instructions were mailed to the teachers in the selected classrooms, and the teachers were asked to distribute the forms to the parents of each child in their classroom. Up to six children per classroom were selected at random to participate from those whose parent returned a positive consent form. Teachers were given a $25 gift card for collecting consent forms from at least 75 percent of enrolled families, regardless of whether the parent agreed. Most children were recruited at the start of the school year (fall 2017), but some were added during spring 2018 to offset attrition from the fall sample.

Parent consent forms were distributed to 4,165 families and 2,341 positive consent forms were returned (56%). Overall, 1,187 children (457 toddlers and 730 preschoolers) from 173 programs completed enough data collection activities to be included in the analysis. This response rate was similar to that of Rhode Island (52%; Maxwell et al., 2016), the one state to report their parental consent rate out of nine states included in a recent synthesis (Tout et al., 2017).

FCCLHs

For FCCLHs, data collection was split into two school years (2016-17 and 2017-18). Recruitment of FCCLH providers for the first year of data collection took place from July to November 2016, and

recruitment for the second year took place from July to October 2017. We invited all FCCLHs in Quality Rated to participate, regardless of star rating, because the number of FCCLHs in Quality Rated was relatively small. Across the two years of data collection, we invited 407 FCCLHs to participate and 158 (39%) agreed. See Table A1 for response rates by star rating. As mentioned in the previous section, this response rate is in the mid-range of that seen in other QRIS validation studies. Providers were offered three $50 gift cards, one each for fall, winter, and spring data collection components.

Packages of consent forms with instructions were mailed to FCCLH providers, and the providers were asked to distribute them to the parents of each eligible child attending their program. All eligible children whose parent retuned a positive consent form were included in the study. To be eligible for the study, children had to be at least 2 months old (by May 31) and no older than six years old and not attending school, including Georgia's Pre-K or kindergarten, during the day. To improve response rates and be consistent with the center-based study, providers were also offered a $25 gift card in the second year for returning consent forms from all or almost all enrolled families, regardless of whether the parent agreed. Most children were recruited at the start of each school year (fall 2016 and 2017), but as in center-based programs, some were added during spring of each school year (2017 and 2018) to offset attrition from the fall sample.

Parent consent forms were distributed to 953 families, and 651 positive consent forms were returned (68%); however, 36 positive consents were from children who were ineligible due to attending school during the day, reducing the response rate to 65 percent. Overall, 601 (273 infants and toddlers and 328 preschoolers) children from 147 programs completed enough data collection activities for inclusion in the analyses. Seven programs do not have children represented in the sample because no positive consent forms were received from parents, but those programs are included in analyses of program observation.

Table A1. Response rates by program type and star rating

The response rate tended to increase with the star rating of the program.

Star rating	Center-based programs			FCCLHs		
	Attempted	In study	Response rate	Attempted	In study	Response rate
0-star	80	28	35%	25	7	28%
1-star	113	39	35%	108	29	27%
2-star	126	64	51%	169	78	46%
3-star	92	50	54%	105	44	42%
Overall	***411***	***181***	***44%***	***407***	***158***	***39%***

Note: The star ratings for the "Attempted" columns were as of the midpoint of the observation window (February 15) for the year in which the programs would have participated. The star ratings for the "In Study" columns were the rating at the time of the classroom or program observation. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

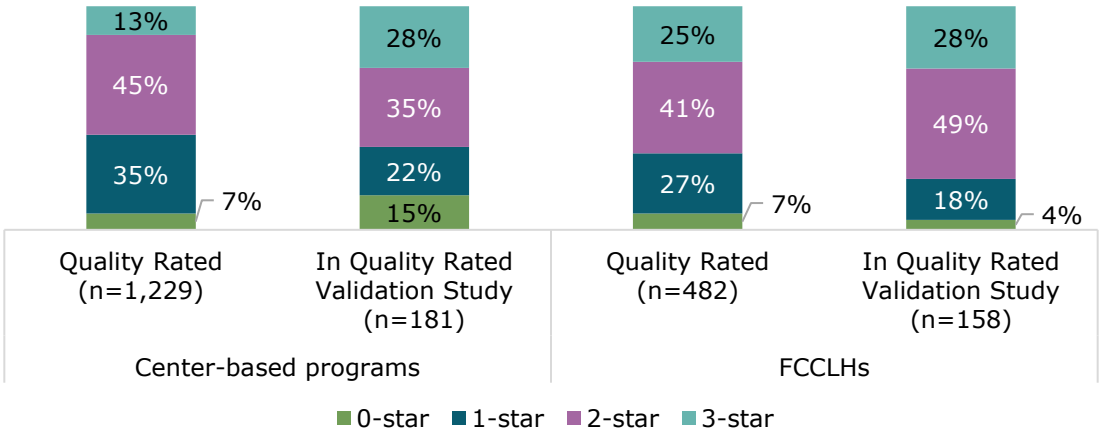
Star rating

The 158 FCCLHs and 181 center-based programs in the study represent 33 percent of all rated FCCLHs and 13 percent of all rated center-based programs. As context for this report, Figure A2 presents the distribution of ratings for the programs in the study and for all Quality Rated programs as of May 15, 2018. The star ratings are from the Quality Rated Administrative Data System, which is maintained by DECAL as part of the process for assigning a star rating.

For the purposes of this report, programs that complete the rating process but do not meet the criteria for 1-, 2-, or 3-stars are referred to as *0-star.* From a policy standpoint, DECAL considers these programs to be participating, but not rated, and does not use the term *0-star.* Because these 0-star programs sought a rating and took part in all aspects of the rating process, we thought it was important to include them when possible. However, very few FCCLH providers with 0-star ratings agreed to participate (*n* = 7), so FCCLHs with 0-star ratings are not included in analyses separated by star rating. They are, however, included in the overall FCCLH demographic information, and children attending 0-star FCCLHs are included in the child-level analyses when center-based programs and FCCLHs are combined.

Figure A2. Star ratings of all programs in Quality Rated and all programs in the Quality Rated Validation Study sample

Programs in the study had a somewhat different star rating distribution than the overall Quality Rated population.



Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated. Programs that were rated after recruitment efforts for the study were not included.
Source: Quality Rated Administrative Data System, May 15, 2018

As described in Report #2 (Orfali et al., 2018), programs may have been rated more than once, either because their rating expired or at their request. Of the 339 programs in the study, 125 (37%) had been rated more than once at any point in time. Of those, 67 programs were re-rated during the school year (August-May) in which they took part in this study. See Table A2 for a description of how many programs had ever been re-rated. This report uses the star rating that was current on the day of the CLASS observation in the preschool classroom or FCCLH, which may not have been the most recent rating if the program was re-rated after the observation occurred.

Table A2. Previous star rating and most recent star rating for programs in the study sample that had been re-rated

Many programs in the study sample increased in rating when re-rated.

Previous star rating	Most recent star rating			
	0-star	1-star	2-star	3-star
0-star	1	3	8	0
1-star	2	8	16	5
2-star	0	3	19	22
3-star	0	2	11	25

Source: Quality Rated Administrative Data System, May 15, 2018

Appendix B: Comparison of study participants to the Quality Rated population

The programs in the study were intended to represent the larger population of all programs in Quality Rated at each star rating. However, the response rate was lower for some star ratings than others. To investigate how similar the sample was to the programs that did not participate and to the population as a whole, we used administrative data to compare the study sample to two groups by star rating: (1) those not participating in the study, and (2) the entire population of Quality Rated programs at the time of recruitment.[2] We used data from the Quality Rated Administrative Data System to compare the groups on their average ERS scores, portfolio scores, and percentage of center-based programs with Head Start funding or Georgia's Pre-K.

The findings indicated that the groups did not differ on most variables, with two exceptions. First, 3-star center-based programs in the study were more likely to receive Head Start funding than 3-star programs that did not participate. Second, 2-star center-based programs in the study had higher portfolio scores and were more likely to receive Head Start funding than 2-star programs that did not participate, or 2-star programs overall. Due to the high level of similarity between the two groups, we concluded that the sample adequately represented the population it was intended to represent.

Average ERS scores

Average ERS scores[3] are the main factor that determines a program's star rating (Early et al., 2017). Within star rating, we used independent samples t-tests to compare the average ERS scores for programs in the study to those that did not participate, and to compare those in the study to the overall population. Findings indicated that there were no differences (see Table B1).

Table B1. Average ERS scores for programs in the study compared to all Quality Rated programs

There were no significant differences in average ERS scores for programs in the study compared to those that declined or the overall population.

		In study			Not in study			In study vs. not	All Quality Rated programs (including in study)			In study vs. all
		n	Mean	SD	n	Mean	SD		n	Mean	SD	
Center-based programs	0-star	28	2.70	0.32	62	2.62	0.28	n.s.	90	2.66	0.29	n.s.
	1-star	39	3.50	0.28	394	3.55	0.29	n.s.	433	3.53	0.31	n.s.
	2-star	64	4.41	0.38	483	4.37	0.38	n.s.	548	4.36	0.38	n.s.
	3-star	50	5.35	0.32	108	5.35	0.29	n.s.	158	5.35	0.30	n.s.

[2] We also compared programs that participated with those that declined. Those results were consistent with the results that compared "in study" versus "not in study" in all instances.
[3] The average ERS score is the average of all the ERS observations conducted at a program for the purpose of the rating. For example, a center-based program might have had multiple observations, including across different age groups. Their score would be the average across classrooms.

		In study			Not in study			In study vs. not	All Quality Rated programs (including in study)			In study vs. all
		n	Mean	SD	n	Mean	SD		n	Mean	SD	
FCCLHs	0-star	7	2.70	0.30	26	2.44	0.42	n.s.	33	2.49	0.41	n.s.
	1-star	29	3.53	0.38	101	3.49	0.34	n.s.	130	3.50	0.35	n.s.
	2-star	78	4.41	0.45	119	4.45	0.41	n.s.	197	4.43	0.43	n.s.
	3-star	44	5.37	0.28	76	5.41	0.36	n.s.	120	5.40	0.33	n.s.

Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated.
Source: Quality Rated Administrative Data System, May 15, 2018

Portfolio scores

As part of the rating process, programs submit evidence in an online portfolio to earn points based on increasingly difficult criteria aligned with five standards. Within star rating, we used independent samples t-tests to compare average portfolio scores of programs in the study to programs that did not participate and to the overall Quality Rated population. There was a significant difference between portfolio points earned by 2-star center-based programs in the study and programs that did not participate and programs in the overall population (see Table B2). There were no significant differences for 0-, 1-, or 3-star center-based programs or any star rating for FCCLHs.

Table B2. Average portfolio score for programs in the study compared to all Quality Rated programs

Two-star center-based programs in the study scored significantly higher on their portfolio compared to the population of all 2-star programs.

		In study			Not in study			In study vs. not	All Quality Rated programs (including in study)			In study vs. all
		n	Mean	SD	n	Mean	SD		n	Mean	SD	
Center-based programs	0-star	28	47.25	20.97	62	44.65	16.89	n.s.	90	45.46	18.17	n.s.
	1-star	39	50.77	17.57	394	45.26	19.50	n.s.	433	45.76	19.38	n.s.
	2-star	64	64.89	24.17	484	55.68	23.74	p<0.01	548	56.76	23.96	p<0.05
	3-star	50	68.86	20.58	108	65.19	23.32	n.s.	158	66.35	22.49	n.s.
FCCLHs	0-star	7	42.29	13.24	28	36.46	15.11	n.s.	35	37.63	14.76	n.s.
	1-star	29	41.66	16.14	101	40.90	14.50	n.s.	130	41.07	14.82	n.s.
	2-star	78	46.85	19.14	119	45.41	16.95	n.s.	197	45.98	17.82	n.s.
	3-star	44	55.16	15.59	76	56.70	17.38	n.s.	120	56.13	16.69	n.s.

Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated.
Source: Quality Rated Administrative Data System, May 15, 2018

Head Start funding

There is some evidence that center-based programs that receive Head Start funding are of higher quality than those that do not (Early et al., 2017). We used chi-squared tests to compare the sample to programs that did not participate and the overall population on the percentage of center-based programs with Head Start funding at each star rating. Significantly more 2-star center-based programs in the study received Head Start funding than 2-star programs that did not participate and the overall population of 2-star programs (see Table B3). In addition, when compared to 3-star programs that did not participate, more 3-star programs in the study had Head Start funding. There were no differences between the study sample and those that did not participate or the overall population for 0- and 1-star programs. Note that FCCLHs are not eligible to receive Head Start funding and were not included in this analysis.

Table B3. Percentage of center-based programs with Head Start funding in the study compared to all Quality Rated programs

A higher percentage of 2-star center-based programs in the study received Head Start funding than the overall population of 2-star programs.

		In study		Declined		In study vs. declined	All Quality Rated programs (including in study)		In study vs. all
		n	Percent	n	Percent		n	Percent	
Center-based programs	0-star	28	4%	62	3%	n.s.	90	3%	n.s.
	1-star	39	5%	394	7%	n.s.	433	6%	n.s.
	2-star	64	33%	484	19%	p<0.01	548	20%	p<0.05
	3-star	50	46%	108	26%	p<0.05	158	32%	n.s.

Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated.
Source: Quality Rated Administrative Data System, May 15, 2018

Georgia's Pre-K classrooms

There is some evidence that center-based programs that include Georgia's Pre-K classrooms are of higher quality than those that do not (Early et al., 2017). We used chi-squared tests to compare the sample to programs that did not participate and the overall population on the percentage of center-based programs with Georgia's Pre-K classrooms at each star rating and found no significant differences (see Table B4). Note that FCCLHs are not eligible to participate in Georgia's Pre-K and were not included in this analysis.

Table B4. Percentage of center-based programs with Georgia's Pre-K in the study compared to all Quality Rated programs

There were no significant differences in the percentage of center-based programs with Georgia's Pre-K in the study compared to the population overall.

		In study		Not in study		In study vs. not	All Quality Rated programs (including in study)		In study vs. all
		n	Percent	n	Percent		n	Percent	
Center-based programs	0-star	28	43%	62	32%	n.s.	90	36%	n.s.
	1-star	39	54%	394	47%	n.s.	433	48%	n.s.
	2-star	64	53%	484	49%	n.s.	548	50%	n.s.
	3-star	50	60%	108	47%	n.s.	158	51%	n.s.

Notes: The ratings for programs not in the study were as of the midpoint of the observation window (February 15) for the year in which they would have participated.
Source: Quality Rated Administrative Data System, May 15, 2018

Appendix C: Detailed data collection process

Data collection took place during the fall, winter, and spring of each school year (2016-17 for FCCLHs, and 2017-18 for FCCLHs and center-based programs). Trained data collectors scheduled visits with FCCLHs and center-based programs via phone call, email, or text within a specified window. Fall child assessments occurred in FCCLHs from September to November during the first year of data collection and the second year of data collection. Spring child assessment visits in FCCLHs occurred from April to June during the first year of data collection and the second year of data collection. Fall visits in center-based programs occurred from August to November 2017, and spring visits occurred from April to June 2018.

For FCCLHs, participating children were classified into age groups that aligned with the age in months they would be at the planned end of each child assessment window (October 31 in the fall and May 31 in the spring) so that the measures appropriate for their age would be administered. During the spring assessment, infants were 1-17.9 months old, toddlers were 18-35.9 months old, and preschoolers were over 36 months old. Children who fell into one age group in the fall and a different age group in the spring received a different battery for post-test than for pre-test. For example, a child who was 31 months old on October 31 and 38 months old on May 31 would have received the toddler battery in the fall and the preschool battery in the spring. In center-based programs, children in the toddler classroom were given the toddler battery, and children in the preschool classroom were given the preschool battery, except for children in the toddler room who turned 36 months old by May 31, who were administered the preschool version of one questionnaire instrument to match their current age. See Table C1 for the mean age in months and range of ages in months for children at post-test.

Table C1. Age in months at post-test for children in the study

Some children in center-based toddler classrooms turned 3 years old during the study school year.

	Center-based programs				FCCLHs			
	Toddler (n=457)		Preschool (n=730)		Infant/Toddler (n=273)		Preschool (n=328)	
	Mean	*Range*	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Range*
Age in months at post-test	33.7	16.9-42.6	54.8	38.8-74.1	23.7	4.5-35.4	46.0	34.5-76.0

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

During the fall and spring child assessment visits, all participating preschool children were given a brief set of assessments designed to measure their expressive vocabulary, early literacy, early math, and executive function skills. Data collectors also gave FCCLH providers and center-based teachers questionnaires to complete about participating infants', toddlers', and preschoolers' skills. When FCCLHs did not have any preschoolers enrolled, data collectors either mailed or dropped off the questionnaires for participating infants and/or toddlers. The questionnaires were accepted through the end of December for the fall assessment windows and the end of June for the spring assessment windows.

FCCLH and center-based classroom observations occurred in the winter. Classroom observations included measures of teacher-child interactions, as well as minute-by-minute coding of children's and teachers' activities. Additionally, audio recordings were made of the language environment. Observations were conducted during the first half of the day, typically beginning when most children arrived at the center and ending before nap time. From November 2017 to April 2018, observations were completed in participating preschool and toddler center-based classrooms. Observations took place in FCCLHs from January to March in both 2017 and 2018.

At the same time the winter classroom observations were taking place, center directors, preschool and toddler teachers, and FCCLH providers were also asked to complete a questionnaire that included their demographic characteristics.

Data collector hiring and training

During the first year of the study, Georgia State University hired and oversaw the graduate research assistants who collected data from FCCLHs, in close collaboration with Child Trends. In the second year of the study, Georgia State hired and supervised data collectors working in FCCLHs, while Child Trends hired and supervised data collectors working in center-based programs. Graduate research assistants at Georgia State primarily covered data collection with Spanish-speaking providers and children, while Child Trends data collectors took on child care programs that were outside the metro Atlanta area. Each year, training and supervising data collectors was a collaborative effort, with the majority of trainings conducted by staff at both institutions. Throughout this report, we use the phrase "validation study team" to refer to the group of researchers who conducted the study at Child Trends and Georgia State. A total of 32 data collectors participated in data collection efforts over a two-year period, including 12 graduate research assistants and four part-time staff at Georgia State and 16 additional part-time staff members at Child Trends. Six of these data collectors worked on the project for two years.

Prior to the start of child assessment data collection each year, all data collectors completed online training related to data security and the protection of research participants, as well as a two-day, in-person training on data collection procedures and the child assessment battery. After practicing the battery for at least one week, each data collector was evaluated using a reliability checklist to ensure study procedures were being followed and each assessment was being administered accurately. The process for assessing reliability varied throughout the course of the project, with some reliability checks occurring in-person and others being videotaped for remote viewing. The majority of data collectors performed their reliability checks with preschool-aged children, but a subset of data collectors administered the battery to another adult due to issues in identifying available young children to serve as practice participants. To the extent possible, inexperienced data collectors were accompanied by more experienced team members on their first visits in the field. Prior to the start of spring child assessments, returning assessors were required to engage in additional reliability checks to ensure they were continuing to administer the battery with fidelity.

All CLASS observers were trained to conduct CLASS Toddler and Pre-K observations by two members of our project team who were certified by Teachstone as affiliate trainers. All CLASS observers achieved reliability in alignment with Teachstone's requirements prior to entering the field. Each CLASS observer also conducted up to three reliability visits with an experienced observer prior to conducting observations independently. To ensure ongoing interrater reliability, observations were conducted in pairs throughout the fielding window, across 10% of all CLASS Toddler and Pre-K observations. On average, the team's scores were within one point of one another 90% of the time during FCCLH CLASS Toddler observations, 92% of the time during center-based CLASS Pre-K observations, and 95% of the time during center-based CLASS Toddler observations. To ensure ongoing fidelity to the CLASS tool, one calibration video was coded by all observers during each observation period. The team conducting observations in FCCLHs achieved 95% reliability in the first year and 100% reliability in the second year, on average. The team conducting center-based observations achieved 96% reliability on CLASS Toddler and 90% on CLASS Pre-K. Because a small subset of the team continued observing into April 2018, one additional CLASS Pre-K (97%, on average) and CLASS Toddler (100%, on average) video were completed by three observers.

COP/TOP observers were trained by a researcher who worked with the Vanderbilt University team that developed COP/TOP. The COP/TOP trainer had extensive experience both in using and training others to use COP/TOP. All COP/TOP observers completed in-person reliability observations with the COP/TOP trainer or another expert from the Vanderbilt team. During these observations, both the data collector and COP/TOP expert coded the entire session. COP/TOP observers were

considered reliable once their scores matched the expert scores 85% of the time across all coding categories. To ensure ongoing interrater reliability throughout the fielding window, 10% of observations were conducted in pairs. On average, the team members were 88% reliable with one another.

Appendix D: Detailed description of measures

This study collected a wide array of data to provide a broad view of how program quality varies as a function of star rating, as well as the extent to which children's growth is linked to star rating. Details regarding all data collection instruments are provided below.

Classroom observations

During the winter of each data collection year, one or two observers visited each classroom and FCCLH to measure teacher-child interactions and gather audio recordings. Table D1 describes the sample size for each type of observation. The tools used are described below.

Table D1. Observed quality measures available across program and classroom type

The language recordings were not conducted in every classroom or FCCLH.

Observed quality	Tool	Center-based classrooms		FCCLHs
		Preschool	*Toddler*	
Teacher-child interactions	CLASS	172	145	147
Child and teacher behavior	COP and TOP	138	-	-
Language environment	LENA[4]	158	136	111

Note: a dash "-" indicates that the construct was not measured for that age group.
Source: validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Classroom Assessment Scoring System

Center-based preschool classrooms were observed using the Classroom Assessment Scoring System Pre-K (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008), an observational tool to assess the quality of the interactions between children and teachers in preschool classrooms (ages 3 to 5 years). Observers rated the classrooms and teachers on each of the dimensions during five 20-minute observation periods throughout the morning.[5] The CLASS Pre-K has 10 subscales, each scored on a 7-point scale with scores of 1 and 2 considered low quality; 3, 4, and 5 considered mid-range quality; and 6 and 7 considered high quality. The subscales are organized into three domains: (1) Emotional Support, (2) Classroom Organization, and (3) Instructional Support.

Center-based toddler classrooms and FCCLHs were observed using the Classroom Assessment Scoring System Toddler (CLASS Toddler; La Paro, Hamre, & Pianta, 2012), an observational tool that is similar to the CLASS Pre-K but assesses the quality of the interactions between children and teachers in toddler classrooms (ages 15 to 36 months). CLASS Toddler includes eight subscales organized into two domains: (1) Engaged Support for Learning, and (2) Emotional and Behavioral Support. The scoring for the CLASS Toddler matches that of CLASS Pre-K, with scores ranging from 1 to 7 and higher values indicating higher quality. The CLASS Toddler is not specifically designed for use with mixed-age groups (Vitiello, 2014), and there is not consensus in the field about the best way to use the CLASS in mixed-age settings. In 2015-2016, according to administrative data, the largest proportion of children in FCCLHs were toddlers, and pre-test data collected during that

[4] Of the 59 total LENA recordings that are not available, 48 were due to the teacher or provider declining to be recorded, and 11 were due to technical difficulties either with the recordings themselves or with transmission of the data to Child Trends.
[5] For all CLASS observations, observers attempted to collect five cycles (20-minute observation periods) of data. If it was not possible to collect five cycles due to the schedule that day, a minimum of four cycles were collected.

same year revealed that the CLASS Toddler items worked adequately in mixed-age FCCLHs. Based on this information, we elected to use the CLASS Toddler in FCCLHs for ease of training, and so we would have comparable data across FCCLH settings with different compositions of infants, toddlers, and preschool-aged children.

Child Observation in Preschool and Teacher Observation in Preschool

The Child Observation in Preschool (COP; Farran et al., 2006) and its companion, the Teacher Observation in Preschool (TOP; Bilbrey et al., 2007), were used as additional observational measures of quality in a subset of center-based preschool classrooms.[6] When used together, the COP and TOP form a system for observing the lead teacher's, assistant teacher's, and children's behavior in a preschool classroom. The COP measures seven behaviors: (1) whether children are talking or listening; (2) to whom children speak or listen; (3) learning setting (e.g., whole group, small group, centers, transition, meal); (4) interaction state (e.g., parallel, associative, cooperative); (5) type of task (e.g., passive, sequential, social, disruptive); (6) level of involvement, using a 5-point scale; and (7) learning focus (e.g., literacy, math, science, art). The TOP measures seven behaviors: (1) whether teachers are talking or listening; (2) to whom teachers speak and listen; (3) learning setting; (4) type of task (e.g., instruction, managerial, behavior approving, social, etc.); (5) level of instruction, using a 4-point scale (only if the type of task is coded as instruction); (6) learning focus; and (7) tone of the interactions the teacher has with the class (e.g., vibrant, pleasant, flat, negative). In addition to the behaviors listed above, the COP/TOP also measures behavior approvals used to reinforce a particular behavior, and behavior disapprovals used in reaction to behaviors teachers would like children to stop. Behavior approvals and disapprovals are tallied each time they are observed.

From 11 to 28 rounds of coding (referred to as sweeps) were completed in each preschool classroom. Although most observers completed at least 16 sweeps per observation, a smaller number of sweeps were completed in some classrooms where the group size was particularly large, and more time was needed to observe all the children in the classroom. During each sweep, the observer located every individual teacher or child in the classroom and observed for approximately three seconds, starting with the lead teacher. Each three-second observation was coded on the dimensions described above prior to moving to the next teacher or child.

We used the scores from the COP/TOP to generate eight scores that previous research has indicated are linked to children's outcomes: (1) transition time (routines and wait time for children), (2) quality of instruction, (3) emotional climate, (4) teachers listening to children, (5) sequential activities, (6) social learning interactions, (7) child involvement, and (8) math opportunities (Farran et al., 2017). We also analyzed a ninth score, literacy opportunities, because it measures a construct of particular interest to DECAL and researchers.

Language Environment Analysis

The Language Environment Analysis digital language processor (LENA; Xu, Yapanel, and Gray, 2009) is a recording device intended to capture all the language that is directed at an individual child while the child wears the device. In addition to creating a digital audio recording, LENA uses a specialized processing software to automatically generate a set of language-related variables, including adult word count, child vocalizations, and the proportions of each session that are silent or noisy. In our study, we asked teachers and providers to wear the LENA device to record their speech on the morning of their CLASS observation. We decided to use this measure based on the evidence that LENA captures important information about children's language experiences (Soderberg, 2014) and we wanted to understand the relationship between star ratings and the

[6] COP/TOP observations were conducted in only 138 of the 172 preschool classrooms because one data collector was not able to achieve reliability on the tool, and there was not enough time to train another individual. Once we knew we would not be able to complete all 172 COP/TOP observations, we removed classrooms from the sample at random, based on a stratification by star rating. The goal was to select an even number of classrooms at each star rating. However, about half of the COP/TOP observations had already been conducted or scheduled by the time the random selection occurred, leading to an overrepresentation of 2-star programs.

language environment of child care settings. LENA was originally designed to record speech during one-on-one interactions between caregivers and children; our study was one of the first times LENA was worn by adults to capture speech in child care settings.

Because previous research has not explored the validity of using the LENA automatic variables to capture adult speech when the LENA is worn by adults and because child care settings can be noisy, we conducted a series of checks on the LENA data between the first and second year of our evaluation. Our first step was to examine the relationship between the LENA automatic variables and human transcription of the full recording. We did this by randomly selecting 10% of the FCCLH recordings from the first year of the study. These 12 recordings were stratified by star rating and included four recordings at each star rating (1-, 2-, and 3-star). Two graduate research assistants transcribed each of the recordings, and a third member of the team reconciled any discrepancies between the two transcriptions. Inter-rater reliability for the pairs who transcribed the 12 full recordings was 99 percent.

After the recordings were transcribed, we processed the transcriptions using Systematic Analysis of Language Transcripts (SALT) software (Miller, Andriachi, & Nockerts, 2011). This process yielded several additional variables to measure the diversity of words used in the language environment (e.g., average length of an utterance in words, number of different words, number of nouns) beyond the automatically generated LENA variables. Our team was particularly interested in the adult word count because this variable could be directly compared to a similar automated variable from LENA. We found that there was not a significant difference between the LENA and SALT adult word counts ($p > 0.05$), giving us confidence in the automatic adult word count generated by the LENA.

Our team was also interested in the more detailed variables SALT produces, but we did not have the resources to transcribe the full recording from each observation. Therefore, we wanted to determine whether a shorter segment could adequately represent the full recording. There is some evidence that five-minute segments can be a stable metric of language (Heilmann, Nockerts, & Miller, 2010), so we randomly selected one five-minute segment from each of the 12 FCCLHs with full transcriptions.[7] The first variable we examined was LENA *adult word count*. The LENA adult word count (measured in words per minute) for full segments was strongly correlated with the adult word count for the five-minute segment ($r = 0.84$, $p < .01$). These numbers were also not significantly different. The second variable we examined was *average length of an utterance* in words. Each utterance could be a spoken sound, word, or statement (e.g., "mmhm," "hello," "that is a beautiful dress"). This variable is a proxy for language quality, whereas adult word count is simply a measure of language quantity. We found that the average utterance length for the five-minute segments was not significantly different from the average utterance length for the full recording. Based on these findings, we decided to use five-minute segments as a proxy for the full recording and proceeded with randomly selecting five-minute segments[8] for all of the available recordings from the first and second years of the study.[9]

The variables we chose for our final analysis are 1) adult word count for the full recording, as calculated by the LENA; 2) average length of adult utterances in words, from the five-minute transcriptions processed using SALT; and 3) type token ratio, another SALT variable. Type token ratio is commonly used in language research as a proxy for vocabulary sophistication because it is a proportion on a scale from zero to one of the number of different words spoken in relation to the total number of words spoken (Kemper & Sumner, 2001). A higher type token ratio indicates the

[7] A random number generator was used to select one five-minute segment from each of the 12 FCCLHs where a full transcription was available. We excluded the first and last five-minute segments because these tended to be times when the data collector was talking with the provider. We also excluded any segments with muffled speech, background noise only, or extended pauses.
[8] The same process was used to select five-minute segments for this step as described in the previous footnote.
[9] Note that we were unable to transcribe a five-minute segment for 22 recordings because the LENA device did not produce a usable audio file during the data transfer process. We were able to include data on the LENA automatic variables for these 22 recordings because those variables are generated independently of the audio file.

speaker uses a more varied vocabulary, while a lower type token ratio indicates a more repetitive vocabulary.

Direct child assessments

Preschool-aged children (at least 36 months old by the end of the post-test assessment window) were directly assessed by trained data collectors in the fall and spring. The battery included measures that would gather information about children's counting ability, early math skills, expressive vocabulary, early literacy skills, and executive function (see Table E2).

Counting Bears

Counting Bears (National Center for Early Development and Learning, 2001) measures children's ability to count up to 40 objects using one-to-one correspondence. As part of the Counting Bears assessment, children are presented with a piece of paper showing them 20 bears. They are instructed to point to and count the bears, one at a time. If the child reaches the end of the sheet, they are presented with a second set of 20 bears.

Woodcock Johnson Test of Achievement, 4th edition

We used three subtests of the Woodcock Johnson Test of Achievement, 4th edition (WJ-IV; Schrank et al., 2014): Picture Vocabulary, Letter-Word Identification, and Applied Problems. The items on each subtest increase with difficulty as the child progresses through them, which allows the WJ-IV to be administered with subjects from age 2 through adulthood. Very young children typically complete the first few pages of each subtest before reaching the ceiling.

Picture Vocabulary measures expressive vocabulary by presenting children with pictures and asking them to label the item in each picture. The assessment begins with simple objects (e.g., banana, bicycle, cat) and progresses to complex vocabulary words (e.g., trombone, hinges, megaphone). Letter-Word Identification measures early literacy by first asking children to point to and label certain letters and then to read a series of words. Because the assessment measures identification of words rather than reading skills, the child must know the word on sight; children do not get credit for sounding out words. Applied Problems measures early math skills. The initial items focus on counting, while later items ask children to do simple addition and subtraction followed by more complicated word problems, including concepts such as money and telling time.

Head-Toes-Knees-Shoulders

Head-Toes-Knees-Shoulders (HTKS; Ponitz et al., 2009) measures executive function skills of inhibitory control, working memory, and attention. This assessment is similar to Simon Says and requires children to perform the opposite command from the instructions given by the data collector. HTKS has three parts; the first part involves only two body parts (head and toes) and the later parts include two additional body parts (knees and shoulders). Subsequent parts are only administered if children reach a certain threshold on the previous section.

Battery for Spanish-speaking children

Children whose parents reported that they spoke Spanish as their dominant language were also assessed on Counting Bears in Spanish and the same three subtests of the Woodcock Muñoz-III (Muñoz-Sandoval et al., 2005; see Table D2). Because HTKS is not a measure of language ability, it was only completed once in the child's dominant language.

Table D2. Constructs and tools used in the preschool direct assessment battery

A wide range of measures were used to gather preschool children's skills in the fall and spring.

Construct	Measure	Center-based sample	FCCLH sample
Counting	Counting Bears	722	308
Early math	WJ-IV Applied Problems	719	308
Expressive vocabulary	WJ-IV Picture-Vocabulary	721	308
Early literacy	WJ-IV Letter-word	719	308
Executive Function	HTKS	719	299

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Teacher or FCCLH provider report of children's skills

Teachers or FCCLH providers filled out questionnaires to report on children's skills. The questionnaires contained different measures according to the study age group that the child fell into at the time of the assessment. See Table E3 for the constructs and the tools used to measure them for each battery.

Devereux Early Childhood Assessment

Teachers or providers completed the Devereux Early Childhood Assessment for Toddlers (DECA-T; Mackrain et al., 2007) and the DECA for Preschoolers, Second Edition (DECA-P2; LeBuffe & Naglieri, 2012). The DECA measures children's social-emotional skills over the past four weeks on a series of items answered on a 5-point scale ranging from "never" to "very frequently." Each version of the assessment yields a total protective factors subscale; the preschool version also yields a subscale measuring behavioral concerns. The toddler version contains 36 items about specific behaviors (e.g., *How often did this child handle frustration well*? and *How often did this child try to do things for herself/himself?*) and has high internal validity reported by the authors (alpha of 0.95 for teacher report of total protective factors). The preschool version contains 38 items (e.g., *How often did this child try different ways to solve a problem?* and *How often did this child play well with others?*). The preschool version also had high internal validity reported by the authors (alpha of 0.95 for teacher report of total protective factors and 0.86 for behavioral concerns).

LENA Developmental Snapshot

The LENA Developmental Snapshot (Gilkerson et al., 2016) measures young children's language acquisition skills. Teachers or providers answered the same set of 52 items for both infants and toddlers, ordered in increasing difficulty by age, indicating whether a child has started consistently demonstrating each skill. The items are ordered such that any items marked as "yes" after the child reaches the ceiling (five responses of "not yet" in a row) are not scored. Early items include *Does this child vocalize or make sounds in response to your smile or voice?* and *Does this child recognize his or her name?* Later items include *Does this child name familiar objects in a room?* and *Can this child name common shapes such as circle, triangle, square, and star?* In the development of the tool, the LENA Snapshot total scores correlated highly with other standardized language assessments.

MacArthur-Bates Communicative Development Inventories

The MacArthur-Bates Communicative Development Inventories, short forms (CDI; Fenson et al., 2000) measure children's vocabulary production from a list of developmentally appropriate words. The toddler version contains 100 words (e.g., star, finish, under) along with the question *Has this child begun to combine words yet?* For the toddler version, the teacher or provider marked off whether the child says the word or not. The internal validity reported by the author for the toddler short form was very high (alpha = 0.99).

The child's FCCLH provider filled out the CDI in Spanish if the FCCLH provider regularly spoke Spanish with the child and had knowledge of their language abilities. However, this group was too small (*n* = 18 across all star ratings) to include in any analyses. No center-based teachers in our sample spoke Spanish with the children in their classrooms.

Table D3. Constructs and tools used in the questionnaires about children's skills

A wide range of measures were used to gather infant, toddler, and preschoolers' skills in the fall and spring.

Age group	Center-based programs			FCCLHs		
	Construct	Measure	Sample	Construct	Measure	Sample
Infant (FCCLHs only) and Toddler	Language acquisition	LENA Developmental Snapshot	443	Language acquisition	LENA Developmental Snapshot	269
Toddler	Expressive vocabulary	CDI-Toddler	449	Expressive vocabulary	CDI-Toddler	189
	Social Skills	DECA-Toddler (Under 36 months)	254	Social Skills	DECA-Toddler	203
	Social Skills Behavioral concerns	DECA-P2 (over 36 months)[10]	191	-	-	
Preschool	Social Skills	DECA-P2	712	Social Skills	DECA-P2	325
	Behavioral concerns			Behavioral concerns		

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Work climate

For the current study, center directors, preschool teachers, toddler teachers, and FCCLH providers were given a questionnaire around the time of their classroom or program observation. In these questionnaires, we asked each staff member about several constructs related to work climate.

Perceived Stress Scale

The Perceived Stress Scale (Cohen et al., 1983) is a 14-item scale intended to capture the degree to which situations are considered to be stressful. The original authors tested the reliability and validity of a four-item version of the scale, and this is the version we used in our study (Cohen, Kamarck, & Mermelstein, 1983). It includes questions such as *In the last month, how often have you felt that things were going your way?* which were ranked on a Likert scale of never (0) to very often (4). The mean of the four items was taken to arrive at a score for each participant, with a higher score indicating more stress. Items that were framed negatively were reverse coded. The alpha for the scale across all participants was 0.63, which is consistent with past research (Cohen et al., 1983). The alpha for center directors was 0.62, the alpha for preschool teachers was 0.68, the alpha for toddler teachers was 0.49, and the alpha for FCCLH providers was 0.65.

How Committed Am I? Scale

The How Committed Am I? scale (Jorde-Bloom, 1988) includes ten items such as *I often think of quitting*, and *I put a lot of extra effort into my work.* We chose six items for the FCCLH

[10] The preschool version of the DECA was administered to children in toddler center-based classroom who had turned 36 months old during the study year. In FCCLHs, children who were above 36 months old were considered "preschoolers."

questionnaire because four items did not apply to a family child care context (e.g., *I don't really care what happens to my center after I leave*). Participants rated each item from a scale of strongly disagree (1) to strongly agree (5). Items that were framed negatively were reverse coded so that a higher score indicated a higher level of commitment. The alpha for the scale across participants was 0.78. The alpha for center directors was 0.71, the alpha for preschool teachers was 0.74, the alpha for toddler teachers was 0.85, and the alpha for FCCLH providers was 0.52.

Teacher turnover, pay, and benefits

Center directors reported (1) how many lead and assistant teachers they currently employed, and (2) how many lead and assistant teachers had left their program and had to be replaced in the past 12 months. We divided the number of teachers who needed to be replaced by the number of currently employed teachers to capture turnover for lead and assistant teachers. Directors also reported how much entry-level preschool and toddler teachers are paid per hour. Finally, we asked center-based staff to indicate which benefits they received from their workplace from a list of twelve benefits, such as health insurance and paid vacation.

Appendix E: Detailed analysis

Data entry and validation

Data were entered into a secure website designed for this project prior to being mailed to Child Trends. After arrival, all data were entered a second time 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 data file for each source. These data files were then checked to ensure all values were valid.

Overall data analysis strategy

Because all of our research questions concern how programs with different star ratings differ from one another, for each set of analyses we compared each star rating to every other star rating. In this report, all p-values of 0.05 or smaller are described as statistically significant. We did not adjust our p-values for multiple comparisons because we did not want to inflate the odds of a Type II error (false negative). Our sample size was relatively small, and we are not comparing a large number of groups, so we decided the threat of Type II errors outweighed the threat of Type I errors (Perneger, 1998). Further, only one of 10 states included in the QRIS validation synthesis used an adjustment for multiple comparisons (Tout et al., 2017). We calculated the effect size using Cohen's d for each statistically significant finding. Except where noted, the effect sizes in this report all met the What Works Clearinghouse (2014) definition of *substantively important,* meaning an effect size of 0.25 or higher.

Our approach differed for the analyses in which quality and work climate were the outcomes versus those in which children's skills were the outcomes. There were two main reasons for this difference. First, we measured children's skills twice (at the beginning and end of the school year), so we could control for pre-test when predicting children's post-test scores. Pre-test scores are likely to be linked to program quality because families with more financial, social, and cognitive resources have more early care and education options; they are also more likely to have children with advanced skills as they start the year. Thus, measuring and controlling pre-test scores is critical in non-experimental studies like this one linking children's skills to program characteristics. Quality and work climate were only measured once because we were not interested in change over time. Second, the children in the sample were nested within classrooms, so accounting for their non-independence was important. We selected only one classroom at each age level, so the quality and work climate data are not nested in programs. We provide additional details about each type of analysis below.

Analyses of star ratings as predictors of quality and work climate

For all observational measures (e.g., CLASS, LENA, COP/TOP), as well as measures of teacher, director, and provider stress and commitment, we conducted ANOVAs followed by pairwise comparisons of each star rating (0-, 1-, 2-, or 3-star) to every other star rating. For the pairwise comparisons, we used Fisher's least-significant-differences (LSD) tests, which differ slightly from independent samples t-tests by using the variance from the entire variable, rather than just the variance from each pair. For each statistically significant result, we gauged the size of the effect by calculating Cohen's d by dividing the mean difference between each pair by the standard deviation of the entire variable. The sample was divided into three groups: 1) preschool classrooms in center-based programs, 2) toddler classrooms in center-based programs, and 3) FCCLHs (1-, 2-, and 3-star only).

For the measures of teacher turnover and pay, we categorized the responses because the data were highly skewed.[11] We then used chi-squared tests to compare the groups by star rating.

Analyses of star ratings as predictors of children's growth

To examine the extent to which children's early academic and social development varied by star rating, we conducted multilevel models that compared each star rating to every other star rating and calculated Cohen's d to gauge the size of each statistically significant result by dividing the coefficient (which represents the partial effect of each star rating compared to the comparison group) by the standard deviation of the entire variable. These multilevel models accounted for the fact that children attending the same program were likely to be more similar to one another than to children attending other programs. To determine whether multilevel models were necessary, we calculated the intraclass correlation (ICC) for each academic and social outcome. The ICCs tell us how strongly children within a program resemble each other on these outcomes. The ICCs ranged from 0.22 to 0.55, indicating about 22 to 55 percent of the variance in child outcomes were accounted for by their program. These are relatively high values, indicating that children within programs were similar enough to one another to necessitate multilevel modeling. The multilevel models controlled for children's pre-test scores, as well as family poverty (below 100% of the poverty line, 100-185% of the poverty line, over 185% of the poverty line), children's race (black, white, other[12]), and children's dominant language (English, other[13]). We describe how we selected pre-test scores and control variables later in this section.

Handling missing data

Missing pre-test scores and control variables were imputed using multiple imputation. This method was selected as there was no reason to believe these data were missing systematically. The variables used in the multiple imputation procedure included gender, birth weight, parent's education, age at post-test, family poverty, race, program attendance,[14] and all available pre-test assessment scores. For each age and language group (i.e., infant, toddler, preschool - English, and preschool - Spanish), we created 40 imputed data sets. That is, for each missing value, the imputed values were drawn 40 times from a distribution to account for the uncertainty and range of values that the true value could have taken. Each analysis was conducted on all 40 imputed data sets. Finally, we combined results from the 40 models using Rubin's combination rule (Rubin, 1987).

The analytic sample for each model was limited to children who were assessed in the spring and were in the appropriate age group for the given assessment. We did not impute post-test scores because we did not have repeated measures of multiple timepoints for the assessments (Schafer, 2003).

Handling children with different assessment batteries

To look at children's growth during the year, children were assessed in the fall and again in the spring. As described in the body of the report, different assessment batteries were used for infants, toddlers, and preschoolers. Children were assessed using the battery that was appropriate for their age at each timepoint. The same assessment battery was used with most children at the two timepoints because they did not switch from infant to toddler or toddler to preschooler between the fall and spring. However, 379 children changed age groups during the study and were given different batteries in the fall and the spring. By definition, those children were younger than those who had the same battery at the two timepoints, so we did not impute their pre-test scores. Instead, we used the pre-test measure that was the most similar to the construct measured by the post-test.

[11] The skew of the lead teacher turnover variable was 5.4 with a kurtosis of 42.4. The skew of the assistant teacher turnover variable was 1.9 with a kurtosis of 6.8. Lead teacher hourly wage had a skew of 2.0 and a kurtosis of 9.4. Assistant teacher hourly wage had a skew of 4.5 and a kurtosis of 32.4.
[12] Over half of the children in the "other" category were multi-racial, according to parent report. For more details about the breakdown of this category, see Table F4.
[13] The majority of children whose dominant language was not English spoke Spanish. See Table F4 more details.
[14] Children's attendance records in February were used as a proxy to estimate their exposure to the child care program.

After selecting the appropriate assessment, we tested the association between the proposed alternative pre-test and the post-test. Most of the correlations were moderately sized and significant, giving us confidence that the substitute pre-test was a suitable replacement for the pre-test that matched the post-test. For HTKS, Counting Bears, and the Woodcock Muñoz-III, the correlations were small or non-significant, so those analyses exclude children who changed batteries during the year. We also included a flag in the analysis that controlled for which pre-test was included (same or different from post-test). Table E1 indicates which alternative pre-test was used for children who changed age groups between the two timepoints.

Table E1. Alternative pre-test assessments for children who switched age groups from fall to spring

For most post-test assessments, a suitable alternative pre-test was identified for children who were assessed using different batteries in the fall and spring.

Alternative Pre-test	Post-test
Fall Infant	***Spring Toddler***
CDI vocabulary – infant (understands)	CDI vocabulary – toddler (says)
Fall Toddler	***Spring Preschooler***
LENA Snapshot standard score	WJ IV Applied Problems standard score
LENA Snapshot standard score	WJ IV Letter-Word ID standard score
LENA Snapshot standard score	WJ Picture Vocab standard score
DECA Total Protective Factors- Toddlers	DECA Total Protective Factors-Preschool
DECA Total Protective Factors- Toddlers (reverse-coded)	DECA Behavioral Concerns-Preschool

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Choosing control variables

We expected that pre-test scores would account for most differences between children that existed prior to their enrollment in Quality Rated child care programs. Thus, our goal in choosing control variables was to include only those demographic characteristics that meaningfully explained post-test scores once pre-test scores were included in the model. To select which demographic factors to include in the child outcomes models, we conducted a series of pilot regressions. We conducted five regressions *for each* outcome, using post-test as the dependent variable, pre-test[15] as the independent variable, and one of the following demographic factors as the control variable:

- child gender,
- child race (white, black, and other, with white serving as the reference group),
- child's dominant language (English or other),
- family income (below 100% of the federal poverty line, between 100-185% of the poverty line, and greater than 185% of the poverty line, with the highest income category serving as the reference group), and
- the number of days between the fall and spring assessments.

Note that we did not run pilot regressions for the child assessments conducted in Spanish due to small sample sizes.

We decided in advance that if a demographic factor accounted for less than 5% of the total variance (i.e., the r^2) in post-test scores after controlling pre-test scores for every outcome, we would remove that demographic factor from our covariate list. In contrast, if one covariate accounted for at least 5% of the variance in *any* of the models, we would include that covariate in our list. We made this

[15] For children who switched age groups (and child assessment batteries) between the fall and the spring, we used Table 12 to guide which pre-test we controlled for in the pilot regressions. We also included a flag indicating whether a different measure was administered at pre-test than the one at post-test.

choice to ensure that we used the same set of covariates in our final analysis for all of the child outcomes, while also minimizing the number of control variables due to limited power.

After running 66 pilot regressions (six models for each of 11 child outcomes), we determined that the following covariates would be included in our models: child race, child's dominant language, and family income. None of the controls that were dropped accounted for more than 1% of the variance across all the models, with the exception of the days between indirect assessments, which accounted for a maximum of 1.7% of the variance.

One exception to the process outlined above for selecting covariates was child age at post-test. For models where the outcome was standardized based on the child's age (i.e., Woodcock-Johnson, LENA Snapshot, DECA), we chose not to include age at post-test as a control because age was already accounted for as part of the standardization process. For models where the outcome was not standardized (i.e., Counting Bears, HTKS, CDI), we decided to include age at post-test as a control under the assumption that older children generally score higher on these assessments.

Appendix F: Detailed children, teacher, provider, and program characteristics

This section provides a brief overview of the characteristics of study participants. More detailed information about teachers, providers, and programs in the study can be found in Appendix A of Report #3 (Early et al., 2018). Parents provided information about their children's demographic characteristics during the consent process.

Characteristics of participating FCCLH and center-based programs

Characteristics of FCCLHs were self-reported by FCCLH providers on the provider questionnaire, and characteristics of center-based programs were self-reported by center directors on the director questionnaire. As seen in Table F1, participating center-based programs varied widely in the number of children enrolled, with a median of 88. Over three-quarters (78%) of participating center-based programs served at least one child receiving Child Care and Parent Services (CAPS) scholarships—that is, funding to serve children from low-income families. Over half of the center-based programs (56%) had a Georgia's Pre-K classroom, and nearly one-third (30%) received Head Start funding.

Almost half (46%) of FCCLH providers reported having at least one staff member in addition to themselves. FCCLH providers served a median of six children. Almost half (42%) of FCCLHs had at least one child enrolled who received a CAPS scholarship.

Table F1. Characteristics of programs in the study

Over three-quarters of center-based programs in the study had children enrolled receiving CAPS scholarships.

		Center-based programs (n=164-174)		FCCLHs (n=149-155)	
		Percentage/ Median	*Range*	*Percentage/ Median*	*Range*
Number of classrooms		6	1 - 19	-	-
Number of teachers	Lead	6	1 - 20	-	-
	Assistant	5	0 - 32	-	-
Additional staff	Did not have additional staff	-	-	54%	
	Had paid additional staff only	-	-	25%	
	Had unpaid additional staff only	-	-	17%	
	Had both paid and unpaid additional staff	-	-	4%	
Children enrolled[16]	Total children enrolled	88	12 - 332	6	1 - 13
	Served infants	79%		62%	
	Served toddlers	83%		86%	
	Served preschoolers	98%		86%	
	Served school-aged children	62%		38%	

[16] The questionnaires asked for the total children enrolled. The total number of children reported may not attend every day.

		Center-based programs (n=164-174)		FCCLHs (n=149-155)	
		Percentage/ Median	*Range*	*Percentage/ Median*	*Range*
CAPS scholarships	Enrolled at least one child receiving a CAPS scholarship	78%		42%	
	% CAPS[17]	20%	<1%- 100%	33%	1% - 100%
	1% - 24%[18]	59%		37%	
	25% - 49%	27%		27%	
	50% - 74%	9%		21%	
	75% - 100%	6%		16%	
Sources of funding	Pre-K	56%		-	-
	Head Start[19]	30%		-	-
Profit status	Not-for-profit	48%		-	-
	For-profit	52%		-	-

Note: Medians are reported instead of means because there are some extreme values that would unduly influence the mean. A dash " – " indicates that the question was not asked of this group. Source: Child Trends' director questionnaire, winter 2017-18; Child Trends' provider questionnaire, winter 2016-17 and winter 2017-18

Characteristics of participating FCCLH providers, center directors, and teachers

Demographic characteristics were self-reported by FCCLH providers, center directors, and teachers on the winter questionnaire. As seen in Table F2, the majority of center directors, teachers, and FCCLH providers in the study were female, black or African American, and spoke English with parents/children. Participants had a wide range of years of experience and professional development. Over half of center directors had a Bachelor's degree or higher, and almost half of preschool teachers had a Bachelor's degree or higher. Much lower percentages of toddler teachers and FCCLH providers had a Bachelor's degree or higher. Most FCCLH providers, center directors, preschool teachers, and toddler teachers majored in early childhood education, and about one-quarter of center directors studied business (not tabled). Over half of FCCLH providers had a Child Development Associate, compared to about one-third of center directors, preschool teachers, and toddler teachers (not tabled).

[17] Average percentage of children with a CAPS scholarship among programs that had at least one child enrolled receiving CAPS.
[18] Distribution of CAPS scholarships, among programs that had at least one child receiving CAPS.
[19] Note that FCCLHs may have been receiving Head Start funds through Early Head Start partnerships, but we did not include this item on the provider questionnaire.

Table F2. Demographic characteristics of FCCLH providers, center directors, and teachers in the study

There was a wide range of educational attainment among participants.

		Center directors (n=174)	Preschool teachers (n=169-172)	Toddler teachers (n=143-146)	FCCLH providers (n=151-155)
Gender	Female	95%	98%	99%	100%
	Male	5%	2%	1%	0%
Race/ Ethnicity	Black/ African American	50%	57%	66%	63%
	Hispanic or Latino	0%	4%	1%	11%
	White	43%	35%	28%	21%
	Other[20]	3%	1%	1%	4%
	Multi-racial[21]	4%	3%	3%	1%
Education	Some high school	0%	0%	2%	2%
	High school diploma/GED	6%	12%	26%	17%
	Some college	22%	24%	38%	37%
	Associate's degree (AA)	18%	19%	18%	19%
	Bachelor's degree (BA/BS)	27%	32%	11%	18%
	Beyond Bachelor's degree	27%	13%	5%	6%

Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017-18; Child Trends' provider questionnaire, winter 2016-17 and winter 2017-18

Characteristics of participating center-based classrooms and FCCLHs

Characteristics of participating center-based classrooms were self-reported by teachers on the teacher questionnaire. As seen in Table F3, median enrollment in preschool classrooms was 17, but the range was large. Toddler classrooms were smaller, with a median enrollment of 10. FCCLHs had a median enrollment of 6 children. The median number of paid adults in both preschool and toddler classrooms was 2 with a range of 1 to 4. In FCCLHs, there was a median of 1 adult present (the provider) but there was a range of 1 to 4 total adults, including paid and unpaid staff members. The median ratio of adults to children was 1:9 in preschool classrooms, 1:7 in toddler classrooms, and 1:5 in FCCLHs. Georgia licensing standards for CCLCs require a staff-to-child ratio of 1:10 for 2-year-old children, 1:15 for 3-year-old children, and 1:18 for 4-year-old children. In mixed-age group classrooms, the ratio should be based on the age of the youngest group of children that includes more than 20 percent of the total number of children.[22] The regulations for staff-to-child ratios in FCCLHs are more complicated, based primarily on ratios of infants, toddlers, and older children. Among preschool classrooms, about one-quarter were part of Georgia's Pre-K, and almost that same percentage were Head Start.

[20] Other includes respondents who selected American Indian/Alaskan Native, Asian, Native Hawaiian or Pacific Islander, or Other.
[21] Multi-racial includes participants who selected more than one of the options presented.
[22] See http://decal.ga.gov/documents/attachments/CCLCRulesandRegulations.pdf for more details about Georgia's licensing regulations. These guidelines apply specifically to CCLCs because Others are exempt from licensing standards.

Table F3. Demographic information about the classrooms in center-based programs in the study

About one-quarter of study classrooms were part of Georgia's Pre-K.

		Preschool classrooms (n=169-172)		Toddler classrooms (n=143-148)		FCCLHs (n=147-155)	
		Percentage/ Median	*Range*	*Percentage/ Median*	*Range*	*Percentage/ Median*	*Range*
Total children enrolled		17	4-30	10	4-21	6	1-13
Number of adults[23]		2	1-4	2	1-4	1	1-4
Adult-to-child ratio		1:9	1:4 – 1:24	1:7	1:2 – 1:16	1:5	2:1 – 1:13
Sources of funding	Georgia's Pre-K[24]	26%		-	-	-	-
	Head Start	22%		5%		-	-

Source: Child Trends' teacher questionnaires, winter 2017-18; Child Trends' provider questionnaire, winter 2016-17 and winter 2017-18

Characteristics of participating children

Table F4 shows the demographic characteristics of the children in the study, as reported by their parents during the consent process. There were slightly more boys than girls, and roughly half were black/African American. Between one-fifth and one-third of children were from families with incomes below the poverty line for their family size. From 11 to 29 percent of children received a CAPS scholarship, according to their parent.

Table F4. Demographic information about the children in the study

From 19 to 40 percent of the families in the study had incomes that fell below the federal poverty level. The majority of children were either black or white and spoke English.

		Center-based programs		FCCLHs	
		Toddler (n=374-457)	Preschool (n=604-730)	Infant/ Toddler (n=221-272)	Preschool (n=270-328)
		Percentage	*Percentage*	*Percentage*	*Percentage*
Gender	Boy	48%	51%	60%	53%
	Girl	52%	49%	40%	47%
Race/Ethnicity	Black/ African American	46%	46%	57%	56%
	White/Caucasian	37%	33%	26%	22%
	Hispanic/Latino	3%	8%	5%	8%
	Other	1%	2%	0%	1%
	Multi-racial	13%	11%	11%	13%
Family poverty level	Below 100%	31%	40%	19%	20%
	100-185%	24%	23%	20%	23%
	Above 185%	44%	38%	62%	57%

[23] Center-based teachers were asked how many paid adults were in the classroom, including themselves, when most of the children were present. The adults in the FCCLHs include the provider and any paid or unpaid assistants.
[24] Since the questionnaires were the same for all teachers in the study, teachers in center-based toddler classrooms were asked if they were in a Georgia's Pre-K classroom. Even though Georgia's Pre-K does not apply to toddler classrooms, 5% of these teachers replied "Yes."

		Center-based programs		FCCLHs	
		Toddler (n=374-457)	Preschool (n=604-730)	Infant/ Toddler (n=221-272)	Preschool (n=270-328)
		Percentage	*Percentage*	*Percentage*	*Percentage*
Received CAPS scholarship	Yes	29%	13%	11%	12%
	No	71%	87%	89%	88%
Has a disability or condition that makes the child eligible for special services	Yes	5%	4%	1%	2%
	No	95%	96%	99%	98%
Parent's highest level of education	No diploma	1%	3%	0%	0%
	High school diploma or equivalent	22%	25%	9%	14%
	Some college or technical training	17%	21%	16%	14%
	Associate's or two-year degree	13%	17%	19%	18%
	Bachelor's degree	24%	20%	25%	30%
	Graduate degree	23%	15%	31%	25%
Language(s) spoken at home	English	97%	94%	98%	97%
	Spanish	5%	10%	7%	11%
	Other	4%	3%	2%	1%

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Appendix G: Descriptive information and statistical comparisons for observations

Table G1. Descriptive information about CLASS scores

Classrooms in 3-star programs generally scored higher in all domains compared to those in lower-rated programs.

		0-star			1-star			2-star			3-star		
		n	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*
CLASS Pre-K	Emotional Support	25	4.98	0.89	39	5.03	0.85	62	5.34	0.72	46	5.60	0.79
	Classroom Organization	25	4.50	0.99	39	4.36	0.92	62	4.85	0.87	46	5.14	1.00
	Instructional Support	25	1.95	0.74	39	2.03	0.60	62	2.16	0.71	46	2.41	0.79
CLASS Toddler	Emotional and Behavioral Support	26	4.64	0.78	36	4.82	0.83	48	4.99	0.88	35	5.41	0.85
	Engaged Support for Learning	26	2.43	0.51	36	2.51	0.67	48	2.56	0.73	35	2.99	0.92
CLASS Toddler - FCCLHs	Emotional and Behavioral Support	-	-	-	27	5.28	0.68	76	5.08	0.89	44	5.44	0.71
	Engaged Support for Learning	-	-	-	27	3.00	0.73	76	2.86	0.73	44	3.35	0.81

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table G2. Descriptive information about LENA scores

The different domains of language environment were generally consistent across settings.

		0-star			1-star			2-star			3-star		
		n	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*
Words per minute	Preschool	24	50.00	22.32	34	58.38	20.61	57	50.99	19.90	42	58.32	20.59
	Toddler	26	41.65	21.76	34	54.94	21.02	43	49.92	26.74	32	60.22	23.12
	FCCLH	-	-	-	23	59.69	22.54	56	58.59	19.30	32	72.21	27.95
Length of utterances in words	Preschool	22	4.68	1.51	30	4.44	1.24	56	4.52	0.97	40	4.63	0.97
	Toddler	25	3.78	0.73	25	3.78	0.73	42	3.89	0.92	32	4.31	0.78
	FCCLH	-	-	-	22	3.98	0.94	53	4.09	0.97	32	4.28	1.08
Proportion of different words spoken	Preschool	22	0.38	0.07	30	0.37	0.06	56	0.37	0.08	40	0.37	0.07
	Toddler	25	0.32	0.07	29	0.35	0.07	42	0.36	0.09	32	0.32	0.05
	FCCLH	-	-	-	22	0.36	0.09	53	0.36	0.07	32	0.34	0.07

Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table G3. Comparison of CLASS scores across star ratings

There was some evidence that classroom quality was higher in 3-star programs compared to those in lower-rated programs.

		Comparison	Difference between means	Significance	Effect size (d)
CLASS Pre-K	Emotional Support	3-star vs. 2-star	0.25		
		3-star vs. 1-star	0.57	*	0.69
		3-star vs. 0-star	0.62	*	0.75
		2-star vs. 1-star	0.32		
		2-star vs. 0-star	0.37	*	0.45
		1-star vs. 0-star	0.05		
	Classroom Organization	3-star vs. 2-star	0.30		
		3-star vs. 1-star	0.78	*	0.66
		3-star vs. 0-star	0.64	*	0.81
		2-star vs. 1-star	0.49	*	0.50
		2-star vs. 0-star	0.35		
		1-star vs. 0-star	-0.14		
	Instructional Support	3-star vs. 2-star	0.25		
		3-star vs. 1-star	0.38	*	0.53
		3-star vs. 0-star	0.46	*	0.63
		2-star vs. 1-star	0.13		
		2-star vs. 0-star	0.21		
		1-star vs. 0-star	0.08		
CLASS Toddler	Emotional and Behavioral Support	3-star vs. 2-star	0.42	*	0.48
		3-star vs. 1-star	0.59	*	0.67
		3-star vs. 0-star	0.77	*	0.88
		2-star vs. 1-star	0.17		
		2-star vs. 0-star	0.35		
		1-star vs. 0-star	0.18		
	Engaged Support for Learning	3-star vs. 2-star	0.44	*	0.58
		3-star vs. 1-star	0.49	*	0.64
		3-star vs. 0-star	0.57	*	0.75
		2-star vs. 1-star	0.05		
		2-star vs. 0-star	0.13		
		1-star vs. 0-star	0.08		
CLASS Toddler - FCCLHs	Emotional and Behavioral Support	3-star vs. 2-star	0.36	*	0.44
		3-star vs. 1-star	0.17		
		2-star vs. 1-star	-0.20		
	Engaged Support for Learning	3-star vs. 2-star	0.49	*	0.63
		3-star vs. 1-star	0.34		
		2-star vs. 1-star	-0.14		

Source: *Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years*

Table G4. Comparison of LENA scores across star ratings

Generally, there were no differences in the richness of the language environment across star ratings. However, FCCLH providers in 3-star programs spoke more words per minute than those in 1-star and 2-star programs.

		Comparison	Difference between means	Significance	Effect size (d)
Words per minute	Preschool	3-star vs. 2-star	7.33		
		3-star vs. 1-star	-0.07		
		3-star vs. 0-star	8.31		
		2-star vs. 1-star	-7.39		
		2-star vs. 0-star	0.98		
		1-star vs. 0-star	8.38		
	Toddler	3-star vs. 2-star	10.30		
		3-star vs. 1-star	5.28		
		3-star vs. 0-star	18.57	*	0.77
		2-star vs. 1-star	-5.02		
		2-star vs. 0-star	8.28		
		1-star vs. 0-star	13.30	*	0.55
	FCCLHs	3-star vs. 2-star	13.63	*	0.57
		3-star vs. 1-star	12.52	*	0.53
		2-star vs. 1-star	-1.11		
Length of utterances in words	Preschool	3-star vs. 2-star	0.11		
		3-star vs. 1-star	0.19		
		3-star vs. 0-star	-0.05		
		2-star vs. 1-star	0.08		
		2-star vs. 0-star	-0.16		
		1-star vs. 0-star	-0.24		
	Toddler	3-star vs. 2-star	0.42		
		3-star vs. 1-star	0.53		
		3-star vs. 0-star	0.53	*	0.56
		2-star vs. 1-star	0.11		
		2-star vs. 0-star	0.11		
		1-star vs. 0-star	0.00		
	FCCLHs	3-star vs. 2-star	0.19		
		3-star vs. 1-star	0.30		
		2-star vs. 1-star	0.11		
Proportion of different words spoken	Preschool	3-star vs. 2-star	0.00		
		3-star vs. 1-star	0.00		
		3-star vs. 0-star	-0.01		
		2-star vs. 1-star	0.01		
		2-star vs. 0-star	-0.01		
		1-star vs. 0-star	-0.01		
	Toddler	3-star vs. 2-star	-0.05	*	-0.61
		3-star vs. 1-star	-0.04		
		3-star vs. 0-star	-0.01		
		2-star vs. 1-star	0.01		
		2-star vs. 0-star	0.04	*	0.51

		Comparison	Difference between means	Significance	Effect size (d)
		1-star vs. 0-star	0.03		
	FCCLHs	3-star vs. 2-star	-0.02		
		3-star vs. 1-star	-0.03		
		2-star vs. 1-star	0.00		

Source: *Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years*

Appendix H: COP and TOP means by star rating

The bolded headings in Table G1 correspond to the nine COP/TOP scores we compared by star rating. Although we did not compare the underlying behaviors (e.g., teacher tone, child associative interaction) by star rating, we present them in the table below because they are easier to interpret than the collapsed factors.

Table H1. COP and TOP descriptive statistics in preschool classrooms

Classrooms in 2-star programs scored significantly higher on overall emotional climate than those in 0- and 1-star programs.

	0-star (n=25)	1-star (n=31)	2-star (n=49)	3-star (n=33)
Transitions				
Child - Transitions (%)	30.7	30.7	29.7	27.2
Quality of instruction				
Teacher - Level of instruction (1-4)	1.7	1.8	1.7	1.7
Positive emotional climate*				
Teacher - Tone (1-5)	3.3	3.2	3.4	3.3
Teacher - Behavior approving (%)	1.2	1.8	2.3	1.1
Teacher - Behavior disapproving (%)	9.4	11.3	6.8	6.4
Teachers listening to children				
Teacher - Listening, total (%)	6.9	7.4	5.9	7.4
Teacher - Listening to child (%)	5.1	4.6	3.2	4.5
Child - Talking, total (%)	13.0	13.2	11.8	13.5
Sequential activities				
Child – Non-sequential activities (%)	22.0	20.3	22.0	22.3
Child - Sequential activities (%)	15.6	18.1	17.6	16.3
Social-learning interactions				
Child - Associative interaction (%)	5.3	6.3	6.0	6.1
Child - Cooperative interaction (%)	0.8	0.7	0.9	0.8
Level of involvement				
Child - Level of involvement (1-5)	2.1	2.1	2.1	2.1
Math opportunities				
Child - Math focus (%)	3.8	4.2	4.3	3.2
Literacy opportunities				
Child - Literacy focus (%)	8.1	8.1	8.8	9.5

Note: One classroom in a 3-star program did not have any observed instruction; therefore, the quality of instruction could not be coded, leading to a sample size of 32.
*The analysis variable for positive emotional climate is a z-score, which is standardized within the sample. For that reason, the differences between 2-star and 0- or 1-star programs (0.49 and 0.58, respectively) can be interpreted as effect sizes similar to Cohen's d.
Source: Validation study team data collection in center-based programs, 2017-18 school year

Appendix I: Regression tables for child outcome analyses

Table I1. Executive functioning in preschool children

Children's executive functioning did not vary across star rating.

Measure	Comparison	All children		Children in center-based programs		Children in FCCLHs	
		B	SE *B*	*B*	SE *B*	*B*	SE *B*
HTKS	0-star vs. 1-star	0.3	1.7	-0.4	1.6	-	-
	0-star vs. 2-star	-0.4	1.5	-0.5	1.4	-	-
	0-star vs. 3-star	-0.5	1.6	0.1	1.5	-	-
	1-star vs. 2-star	-0.8	1.2	-0.1	1.2	-0.9	2.0
	1-star vs. 3-star	-0.9	1.2	0.5	1.3	-2.2	2.1
	2-star vs. 3-star	-0.1	1.0	0.6	1.1	-1.2	1.6

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I2. Language and literacy in preschool children

Preschoolers' expressive vocabulary and early literacy did not vary across star rating.

Measure	Comparison	All children		Children in center-based programs		Children in FCCLHs	
		B	SE *B*	*B*	SE *B*	*B*	SE *B*
WJ Picture Vocabulary	0-star vs. 1-star	1.6	1.8	1.5	1.8	-	-
	0-star vs. 2-star	0.7	1.6	2.2	1.7	-	-
	0-star vs. 3-star	1.4	1.6	2.2	1.7	-	-
	1-star vs. 2-star	-0.9	1.3	0.7	1.4	-4.1	2.5
	1-star vs. 3-star	-0.2	1.3	0.7	1.5	-2.8	2.6
	2-star vs. 3-star	0.7	1.1	0.0	1.3	1.4	1.8

Measure	Comparison	All children		Children in center-based programs		Children in FCCLHs	
		B	SE *B*	*B*	SE *B*	*B*	SE *B*
WJ Letter Word	0-star vs. 1-star	2.2	2.0	1.2	2.1	-	-
	0-star vs. 2-star	1.6	1.9	0.2	1.9	-	-
	0-star vs. 3-star	3.4	1.9	1.9	1.9	-	-
	1-star vs. 2-star	-0.6	1.5	-1.0	1.6	-0.1	2.8
	1-star vs. 3-star	1.2	1.5	0.6	1.6	2.8	3.0
	2-star vs. 3-star	1.8	1.3	1.6	1.4	2.9	2.1

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I3. Language and literacy in infants and toddlers

Infants' and toddlers' language acquisition and toddlers' expressive vocabulary did not vary by star rating.

Measure	Comparison	All children		Children in center-based programs		Children in FCCLHs	
		B	SE *B*	*B*	SE *B*	*B*	SE *B*
LENA Developmental Snapshot	0-star vs. 1-star	-1.7	3.1	-0.1	3.8	-	-
	0-star vs. 2-star	-0.4	2.9	0.7	3.8	-	-
	0-star vs. 3-star	0.2	3.0	1.6	3.8	-	-
	1-star vs. 2-star	1.3	2.1	0.8	3.0	1.5	3.1
	1-star vs. 3-star	1.9	2.2	1.7	3.1	1.6	3.3
	2-star vs. 3-star	0.6	2.0	0.9	3.1	0.2	2.6
CDI Toddler	0-star vs. 1-star	-2.0	4.7	1.3	5.3	-	-
	0-star vs. 2-star	-2.6	4.5	2.9	5.2	-	-
	0-star vs. 3-star	-2.5	4.6	0.9	5.3	-	-
	1-star vs. 2-star	-0.7	3.4	1.6	4.2	-3.5	6.0
	1-star vs. 3-star	-0.6	3.5	-0.4	4.3	0.4	6.4
	2-star vs. 3-star	0.1	3.2	-2.0	4.2	3.8	5.2

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I4. Math skills in preschool children

There is some evidence that children's early math skills were higher in 3-star programs compared to those of children attending lower-rated programs.

Measure	Comparison	All children			Children in center-based programs			Children in FCCLHs	
		B	SE *B*	*Effect size (d)*	*B*	SE *B*	*Effect size (d)*	*B*	SE *B*
Counting Bears	0-star vs. 1-star	0.9	1.4		1.2	1.2		-	-
	0-star vs. 2-star	1.0	1.2		0.9	1.1		-	-
	0-star vs. 3-star	2.0	1.3		2.5 *	1.1	0.22	-	-
	1-star vs. 2-star	0.1	1.0		-0.3	0.9		0.8	1.7
	1-star vs. 3-star	1.1	1.0		1.3	1.0		1.1	1.8
	2-star vs. 3-star	1.1	0.8		1.6	0.9		0.2	1.3
WJ Applied Problems	0-star vs. 1-star	4.9 *	2.1	0.32	4.2	2.3		-	-
	0-star vs. 2-star	2.4	1.9		2.4	2.1		-	-
	0-star vs. 3-star	5.1 *	2.0	0.34	4.7 *	2.2	0.31	-	-
	1-star vs. 2-star	-2.5	1.5		-1.8	1.8		-3.8	3.0
	1-star vs. 3-star	0.2	1.6		0.6	1.9		-0.7	3.2
	2-star vs. 3-star	2.7 *	1.3	0.18	2.3	1.6		3.1	2.2

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I5. Social and emotional development in preschool and older toddler children

Children attending 3- and 2- star programs had higher social skills and lower behavioral concerns than those attending lower-rated programs.

Measure	Comparison	All children			Children in center-based programs			Children in FCCLHs		
		B	SE *B*	*Effect size (d)*	*B*	SE *B*	*Effect size (d)*	*B*	SE *B*	*Effect size (d)*
DECA P2 Protective Factors	0-star vs. 1-star	2.6	1.4		2.7	1.5		-	-	
	0-star vs. 2-star	2.8 *	1.3	0.28	3.2 *	1.4	0.31	-	-	
	0-star vs. 3-star	4.1 **	1.3	0.40	3.6 *	1.5	0.35	-	-	
	1-star vs. 2-star	0.2	1.0		0.5	1.2		0.2	1.8	
	1-star vs. 3-star	1.5	1.0		0.9	1.2		3.1	2.0	
	2-star vs. 3-star	1.3	0.9		0.4	1.1		2.9 *	1.4	0.31
DECA Behavioral Concerns	0-star vs. 1-star	0.7	1.4		1.1	1.6		-	-	
	0-star vs. 2-star	-2.1	1.3		-1.6	1.5		-	-	
	0-star vs. 3-star	-1.8	1.4		-1.1	1.5		-	-	
	1-star vs. 2-star	-2.8 **	1.0	-0.29	-2.7 *	1.3	-0.27	-3.5	1.9	
	1-star vs. 3-star	-2.5 *	1.1	-0.25	-2.2	1.3		-4.2 *	2.1	-0.27
	2-star vs. 3-star	0.3	0.9		0.5	1.2		-0.6	1.5	

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I6. Social and emotional development in toddlers

Generally, there were no differences in social skills across star rating. However, toddlers attending 0-star programs had higher social skills compared to those attending 3-star programs.

Measure	Comparison	All children		Children in center-based programs			Children in FCCLHs	
		B	SE *B*	*B*	SE *B*	*Effect size (d)*	*B*	SE *B*
DECA Toddler Protective Factors	0-star vs. 1-star	-3.4	2.4	-5.7	3.2		-	-
	0-star vs. 2-star	-2.2	2.3	-3.4	3.1		-	-
	0-star vs. 3-star	-2.6	2.4	-7.5 *	3.2	-0.65	-	-
	1-star vs. 2-star	1.2	1.7	2.3	2.5		0.9	2.2
	1-star vs. 3-star	0.8	1.8	-1.7	2.6		2.8	2.3
	2-star vs. 3-star	-0.4	1.5	-4.1	2.5		1.9	1.8

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Table I7. Children's skills for Spanish-speaking children

Spanish-speaking children's early math, language, and literacy outcomes did not vary across star rating.

Measure	Comparison	All children	
		B	SE *B*
Counting Bears (Spanish)	2-star vs. 3-star	1.5	1.5
WM Letter-Word	2-star vs. 3-star	3.3	6.2
WM Picture Vocabulary	2-star vs. 3-star	-2.6	3.9
WM Applied Problems	2-star vs. 3-star	-5.9	3.2

Note: * = p < .05, ** = p < .01, *** = p < .001. Each line represents up to three HLM models. A dash "-" indicates that the model was not run for that comparison. Source: Validation study team data collection in center-based programs, 2017-18 school year; Validation study team data collection in FCCLHs, 2016-17 and 2017-18 school years

Appendix J: Descriptive information and statistical comparisons for perceived stress scale and job commitment scale

Table J1. Descriptive information about perceived stress and job commitment

Provider stress was generally higher in FCCLH providers and job commitment was very high on average for all survey participants.

		0-star			1-star			2-star			3-star		
		n	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*Mean*	*SD*
Perceived Stress Scale (PSS)	Director	23	1.02	0.73	39	1.24	0.75	60	1.12	0.58	46	1.01	0.76
	Preschool teacher	24	1.04	0.67	38	1.20	0.69	62	0.92	0.72	45	1.01	0.72
	Toddler teacher	26	0.80	0.64	35	1.03	0.60	48	1.19	0.68	36	0.89	0.61
	FCCLH provider	-	-	-	26	2.23	0.48	75	2.11	0.44	43	2.15	0.49
Job commitment	Director	22	4.50	0.57	38	4.66	0.44	62	4.55	0.43	47	4.63	0.39
	Preschool teacher	22	4.24	0.51	37	4.39	0.49	61	4.49	0.45	45	4.53	0.39
	Toddler teacher	25	4.44	0.50	35	4.36	0.58	47	4.18	0.78	37	4.45	0.65
	FCCLH provider	-	-	-	28	4.21	0.31	75	4.09	0.44	44	4.16	0.36

Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018; Child Trends' provider questionnaire, winter 2016–2017 and winter 2017–2018

Table J2. Comparison of perceived stress and job commitment across star ratings

Generally, there were no differences in perceived stress or job commitment across star ratings. However, toddler teachers in 2-star programs had higher stress compared to those in 0-star or 3-star programs, and preschool teachers in 3-star programs were more committed to their jobs than those in 0-star programs.

		Comparison	Difference between means		Effect size (d)
Perceived Stress Scale (PSS)	Director	3-star vs. 2-star	-0.11		
		3-star vs. 1-star	-0.24		
		3-star vs. 0-star	-0.02		
		2-star vs. 1-star	-0.13		
		2-star vs. 0-star	0.09		
		1-star vs. 0-star	0.22		
	Preschool teacher	3-star vs. 2-star	0.09		
		3-star vs. 1-star	-0.19		
		3-star vs. 0-star	-0.03		
		2-star vs. 1-star	-0.28		
		2-star vs. 0-star	-0.12		
		1-star vs. 0-star	0.16		
	Toddler teacher	3-star vs. 2-star	-0.30	*	-0.45
		3-star vs. 1-star	-0.14		
		3-star vs. 0-star	0.09		
		2-star vs. 1-star	0.16		
		2-star vs. 0-star	0.39	*	0.64
		1-star vs. 0-star	0.23		
	FCCLH provider	3-star vs. 2-star	0.04		
		3-star vs. 1-star	-0.09		
		2-star vs. 1-star	-0.13		
Job commitment	Director	3-star vs. 2-star	0.08		
		3-star vs. 1-star	-0.02		
		3-star vs. 0-star	0.13		
		2-star vs. 1-star	-0.10		
		2-star vs. 0-star	0.05		
		1-star vs. 0-star	0.16		
	Preschool teacher	3-star vs. 2-star	0.04		
		3-star vs. 1-star	0.14		
		3-star vs. 0-star	0.29	*	0.60
		2-star vs. 1-star	0.10		
		2-star vs. 0-star	0.25		
		1-star vs. 0-star	0.15		
	Toddler teacher	3-star vs. 2-star	0.27		
		3-star vs. 1-star	0.09		
		3-star vs. 0-star	0.01		
		2-star vs. 1-star	-0.18		
		2-star vs. 0-star	-0.26		
		1-star vs. 0-star	-0.08		
	FCCLH provider	3-star vs. 2-star	0.07		
		3-star vs. 1-star	-0.05		
		2-star vs. 1-star	-0.12		

Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018; Child Trends' provider questionnaire, winter 2016–2017 and winter 2017–2018

Appendix K: Means, medians, and ranges for teacher turnover and entry-level teacher hourly wages

Center-based directors were asked how many lead and assistant teachers they had at their center and how many had left and had to be replaced in the past 12 months; those numbers were used to estimate teacher turnover. Directors were also asked how much an entry-level preschool and toddler teacher would be paid hourly at their center. These outcomes were categorized in the main text due to a high level of skew. Means, medians, and ranges for these measures are presented in Table I1.

Table K1. Mean, median, and range for lead and assistant teacher turnover and entry-level hourly wages for preschool and toddler teachers

Turnover tended to be higher and wages tended to be lower in lower-rated programs.

		0-star (n=21-22)			1-star (n=36-38)			2-star (n=55-62)			3-star (n=43-45)		
		Mean	*Median*	*Range*	*Mean*	*Median*	*Range*	*Mean*	*Median*	*Range*	*Mean*	*Median*	*Range*
Teachers	Lead teachers	5.0	5	(2-9)	6.4	6	(2-16)	6.6	6	(1-20)	7.2	7	(1-16)
	Lead teachers left	2.2	2	(0-8)	2.1	2	(0-11)	1.5	1	(0-6)	1.5	1	(0-6)
	Assistant teachers	5.8	4	(1-17)	5.6	5	(1-14)	6.4	5	(0-18)	7.0	6	(0-32)
	Assistant teachers left	3.9	2	(0-28)	2.9	2	(0-25)	1.8	1	(0-8)	1.7	1	(0-6)
Turnover	Lead teacher	51%	33%	(0%-400%)	37%	29%	(0%-300%)	22%	20%	(0%-75%)	24%	20%	(0%-100%)
	Assistant teacher	64%	50%	(0%-200%)	58%	50%	(0%-250%)	31%	23%	(0-2)	33%	20%	(0%-200%)
Hourly wage	Preschool teacher	$9.92	$9.25	($7.25-$17.00)	$11.27	$10.00	($7.25-$22.00)	$11.32	$9.50	($7.25-$22.50)	$13.40	$11.63	($8.00-$26.00)
	Toddler teacher	$9.04	$9.00	($7.25-$14.00)	$9.63	$9.00	($7.25-$25.00)	$10.07	$9.00	($7.25-$22.00)	$10.80	$10.00	($8.00-$20.00)

Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018

Appendix L: Benefits by star rating for center directors and toddler teachers

Center-based staff were also asked to indicate which benefits they received from their workplace from a list of twelve benefits, such as health insurance and paid vacation (See Table L1).

Significantly more directors in 3- and 2-star programs had health insurance than directors in 1-star programs, but higher-rated programs did not differ significantly from directors in 0-star programs. Significantly more directors in 3- and 2-star programs had retirement benefits than directors in 1-star programs; the percentage who had retirement benefits in 3-star programs also differed significantly from those in 0-star programs. Significantly more directors in 2-star programs received dental insurance than those in 0- and 1-star programs. Significantly more directors in 3- and 2-star programs received paid sick leave than directors in 0- and 1-star programs. Paid vacation or personal leave did not differ among the star ratings. Information about the significant differences between benefits that preschool teachers in differently rated programs received can be found in the main body of this report.

Significantly more toddler teachers in 3-star programs had health insurance, retirement benefits, dental insurance, and paid sick leave than toddler teachers in 0-, 1-, or 2-star programs. Comparisons among toddler teachers in 0-, 1-, and 2-star programs were not significantly different.

Table L1. Percentage of center directors and teachers with each benefit across star rating

In general, the percentage of staff members who received benefits tended to increase as the star rating increased.

Benefit	Center directors				Preschool teachers				Toddler teachers			
	0-star (n=22)	1-star (n=39)	2-star (n=60)	3-star (n=48)	0-star (n=22)	1-star (n=38)	2-star (n=63)	3-star (n=46)	0-star (n=26)	1-star (n=33)	2-star (n=47)	3-star (n=37)
Health insurance	36%	26%	47%	58%	9%	26%	40%	67%	15%	18%	19%	49%
Retirement benefits	23%	15%	40%	58%	9%	13%	24%	37%	4%	9%	9%	32%
Dental insurance	27%	26%	40%	52%	5%	21%	32%	50%	8%	15%	15%	41%
Paid sick leave	45%	44%	78%	85%	36%	21%	44%	70%	19%	24%	28%	51%
Paid vacation or personal leave	68%	69%	72%	73%	32%	32%	48%	50%	35%	39%	36%	51%

Source: Child Trends' director questionnaire and teacher questionnaires, winter 2017–2018