YEAR 3 REPORT: SEATTLE PRE-K PROGRAM EVALUATION

Milagros Nores, Ph.D., Steve Barnett, Ph.D., Kwanghee Jung, Ph.D. Gail Joseph, Ph.D., Lea Bachman, Ph.D. & Janet S. Soderberg, Ph.D., The National Institute for Early Education Research & Cultivate Learning

September 2018

About the Authors

Milagros Nores, Ph.D. Dr. Nores is Co-Director for Research at The National Institute for Early Education Research (NIEER) at Rutgers University. Dr. Nores conducts research at NIEER on issues related to early childhood policy, programs, and evaluation, both nationally and internationally. She is also on staff with the Center for Enhancing Early Learning Outcomes (CEELO), a federally funded comprehensive center that provides technical assistance to state agencies around early childhood.

W. Steve Barnett, Ph.D. Dr. Barnett is a Senior Co-Director of the National Institute for Early Education Research (NIEER) and a Board of Governors Professor at Rutgers University. He is also Principal Investigator of the Center for Enhancing Early Learning Outcomes (CEELO). His research includes studies of the economics of early care and education including costs and benefits, the long-term effects of preschool programs on children's learning and development, the economics of human development, practical policies for translating research findings into effective public investments and the distribution of educational opportunities.

Kwanghee Jung, Ph.D. Dr. Jung is an Assistant Professor at The National Institute for Early Education Research (NIEER) at Rutgers University. Her expertise is in quantitative data analysis and the effect of participation in child care and early education on children's learning and development.

Gail Joseph, Ph.D. Dr. Joseph is the Bezos Family Distinguished Professor in Early Learning at the University of Washington. She teaches courses, advises students, provides service and conducts research on topics related to early care and education. She is the Founding Executive Director of Cultivate Learning at the University of Washington (previously known as the Childcare Quality and Early Learning Center for Research and Professional Development, CQEL).

Lea Bachman, Ph.D. Dr. Bachman is a Research Associate at Cultivate Learning (CL) at the University of Washington. She leads CL's work on the SPP Evaluation Study and conducts research on topics related to early childhood education and assessment. She is a psychologist with significant experience in classroom observation, data collection and management.

Janet S. Soderberg, Ph.D. Dr. Soderberg is a Senior Research Scientist Cultivate Learning at the University of Washington. Dr. Soderberg's research includes exploration of the association between classroom quality and children's development, QRIS program evaluation, refinement and support of kindergarten entrance assessments, and dissemination of research. Her interests include child development, assessment, childcare quality, and multi-systems alignment.

Table of Contents

Executive Summary
Third Year Evaluation (2017-18) of the Seattle Preschool Program (SPP)

In 2017–18 SPP grew to 48 classrooms and 13 family child care providers from 32 classrooms in the prior year. SPP quality continued to improve and now reaches levels associated with strong gains in children's learning and development. We recommend that SPP build on its success by seeking further improvements in the quality of instruction with particular attention to language and literacy, integration of content across domains in children's activities, and supports for sustained, reflective thinking as well as personal care routines that contribute to health.
        In Year 3 of the evaluation we addressed four specific questions. Below, we summarize key findings for each question and note specific sections of the report where additional information about the findings can be found.

1. Who enrolled in SPP in 2017–18, and how do they compare demographically to children in Seattle more generally?

SPP children closely resemble the general public-school population in Seattle with respect to gender, language, and income (p. 13). SPP children are somewhat more likely to identify as African American or Black and Asian (and less likely to identify as White) than the overall public-school population. Overall, 74% of the children enrolled in SPP were 4-year-olds, 29% were dual language learners, and they were identified as 12% Hispanic, 22% White, 27% African American, and 28% Asian.

2. What was the observed quality of children's SPP classroom experiences in 2017–18, and did it improve over the prior year?

SPP quality has continued to improve on two separate measures, the Early Childhood Environmental Rating Scale—Third edition (ECERS-3) and the Classroom Assessment Scoring System (CLASS). The average ECERS-3 score increased from 3.89 to 3.99 (on a 7-point scale) and this increase was statistically significant (p.14). CLASS scores also increased, with a particularly large gain for instructional supports that are important for building academic success, moving from 3.06 to 3.42 (also on a 7-point scale) (p.15). Emotional support moved from 6.29 to 6.38 and Classroom Organization moved from 5.55 to 5.96.
        SPP quality as measured by the ECERS-3 and CLASS now exceeds that in some other major city and state pre-k and/or childcare systems. SPP quality is similar to that of the widely recognized New York City and San Antonio programs. Quality must continue to improve if it is to reach the levels in the states and cities with the highest levels of quality observed in research (p. 20).

3. How does quality vary within SPP, and do children from different backgrounds experience different quality?

Average quality does not differ significantly between classrooms and family child care providers (FCCs), which were added this year as part of a pilot (p. 25). Controlling for center and

classroom characteristics (lead teacher qualifications, class size, among others), quality is lower for FCCs on the Emotional Support and Classroom Organization dimensions of CLASS.

There is some variation in quality among classrooms, and continuous improvement should seek to raise the bottom end of the distribution. Average quality as measured by the ECERS-3 and average quality of instructional support as measured by the CLASS are not significantly different by race and ethnicity. Modest differences in quality of classroom organization and emotional supports are observer by children's race and ethnicity, but average quality in those two domains were high for all groups (p. 26).

4. How did children in SPP classrooms and family child care providers progress in 2017–18, and how did it vary with classroom quality? Other program characteristics? How did it vary with child characteristics?

Children in SPP made gains in every domain measured (p.28). Gains in language, literacy and mathematics were larger than would be expected based on maturation (increased age) alone. There were no large differences in gains by gender, race or ethnicity, or language. Children identified as Asian made smaller gains in receptive vocabulary, but larger gains in executive functions, when accounting for other child and school characteristics. No systematic differences in gains were found by income.

Differences in classroom size or curriculum were not found to relate to children's performance (p.31). Better quality of classroom organization was associated with strong gains in math. (Last year's evaluation indicated that better classroom organization also improves literacy scores.) Teacher qualifications were not found to be associated with test score gains. African-American and Asian teachers' students had larger gains in vocabulary, reinforcing the importance of teacher diversity in SPP where most students identify as African-American, Black, or Asian.

The SPP evaluation was conducted by National Institute for Early Education Research at Rutgers University and Cultivate Learning, at the University of Washington. This report focuses on the of the 2017–2018 school year and includes information on the prior two years.

Introduction

The City of Seattle has concluded its third year in its four-year demonstration phase for the Seattle Preschool Program (SPP). SPP was established after voter approval on 2014 of a four-year, $58 million property tax levy. The levy's proposition was of "accessible high-quality preschool services for Seattle children designed to improve their readiness for school and to support their subsequent academic achievement." SPP was subsequently launched in 2015 by the city of Seattle's Department of Education and Early Learning (DEEL) with 14 classrooms operating that year. In the 2016–17 school year, the program more than doubled, operating in 32 classrooms, and by 2017–18, it had expanded to 48 classrooms and 13 family child care providers.

The four-year demonstration phase of SPP included from its very beginning an evaluation component to inform its viability and most significantly support its quality improvement processes. The National Institute for Early Education Research at Rutgers University and Cultivate Learning, at the University of Washington concluded the third year of the evaluation of the demonstration phase of the Seattle Preschool Program (SPP).

This report presents findings for the third year (2017–18) of the program and is centered on classroom quality and children's learning. The report includes information on the children served, children's learning and development during the school year, and program quality across the three years of SPP so far. It also looks at specific subgroups of children and classrooms and examines associations between SPP children's learning gains and their classroom experiences including observed quality.

Study Methods

The SPP evaluation study is a multi-site study that combines various components in order to provide a comprehensive appraisal of the program's quality and its impact on children through the four-year demonstration period. The third year of the study included collection of child and classroom information to address the following four questions:

1. Who enrolled in SPP in 2017–18, and how do they compare demographically to children in Seattle more generally?
2. What was the observed quality of children's SPP classroom experiences in 2017–18, and did it improve over the prior year?
3. How does quality vary within SPP, and do children from different backgrounds experience different quality?
4. How did children in SPP classrooms and family child care providers progress in 2017–18, and how did it vary with classroom quality? Other program characteristics? How did it vary with child characteristics?

The SPP evaluation was framed to understand SPP children's learning and development, as well as how classroom processes evolved over time. In Year 1, the research team measured learning and development at the beginning and at the end of the year, as well as classroom quality. In Year 2, the research team repeated this process, and also recruited a non-equivalent comparison group that is composed of children in the waiting list for SPP together with children attending

centers were some waiting list children ended up enrolled. The team continued to conduct classroom observations. In Year 3, the research team measured learning and development at the beginning and at the end of the year as well as classroom quality in SPP classrooms, and family child care providers (FCCs) which were incorporates this year into SPP as a pilot. FCCs were brought into the program through two hubs, which were tasked with managing up to eight FCCs. In the end, one of the hubs contracted with eight FCCs and the other one with seven, for a total of 13 FCCs being brought into SPP in the 2017-18 school year. Measures and procedures used across all centers, FCCs and children are described below.

   Children were first assessed this year in the Fall of 2017 and assessed again at the end of the school year in 2018. Direct observations of classroom practices were performed to assess overall quality, teacher-child interactions, and engagement. Classroom quality observations were completed between February and March. Quality was assessed using observation protocols widely established in the field. Figure 1 reports the data collection timeline for the 2017–18 school year.

Figure 1. Data Collection Timeline

2017	
September	• Training for data collectors
	• Initial SPP site information gathered
October	• Fall assessment visit scheduling
	• Fall child assessment visits begin
November	• Fall child assessment visits continue
	• Fall assessment visits completed (by December 8)
2018	
January	• Communications to directors to discuss classroom observations (CLASS & ECERS-3)
February	• Unannounced CLASS & ECERS-3 observations
March	• Unannounced CLASS & ECERS-3 observations continue
April - June	• Unannounced CLASS & ECERS-3 observations completed
	• Spring assessment visit scheduling (early April)
	• Spring child assessment visits

Sample

In the school year 2017–18 the research team assessed 761 children in 48 SPP classrooms and 13 SPP family child care providers at pre- and post-test (615 with the full battery). To recruit children in the study, consent forms were distributed forms to parents or guardians of all 943 children enrolled in these classrooms. A total of 913 were consented to participate in the study. We randomly selected 10 children per classroom for the full battery. Figure 1 below shows the study attrition tree. Seven children required language accommodations.

Figure 1. Pre-Post Sample Attrition Tree (includes children assessed with PPVT only)

```
                    ┌─────────────────────────────────┐
                    │ N=943 SPP enrolled children (at  │
                    │   some point during the year)    │
                    └─────────────────────────────────┘
                       ↙                           ↘
        ┌──────────────────────┐        ┌──────────────────────────┐
        │ N=913 with consent   │        │ N=30 without consent     │
        └──────────────────────┘        └──────────────────────────┘
           ↙              ↘
┌────────────────────────┐   ┌──────────────────────┐
│ N=820 with pre-test    │   │ N=93 without pre-     │
│ (666 Full, 154 PPVT    │   │ test                 │
│ Only)                  │   └──────────────────────┘
└────────────────────────┘
   ↙              ↘
┌──────────────────────┐   ┌──────────────────────┐
│ N=761 with post-test │   │ N=59 without post-   │
│ (615 Full, 146 PPVT  │   │ test (51 Full, 8     │
│ Only)                │   │ PPVT)                │
└──────────────────────┘   └──────────────────────┘
```
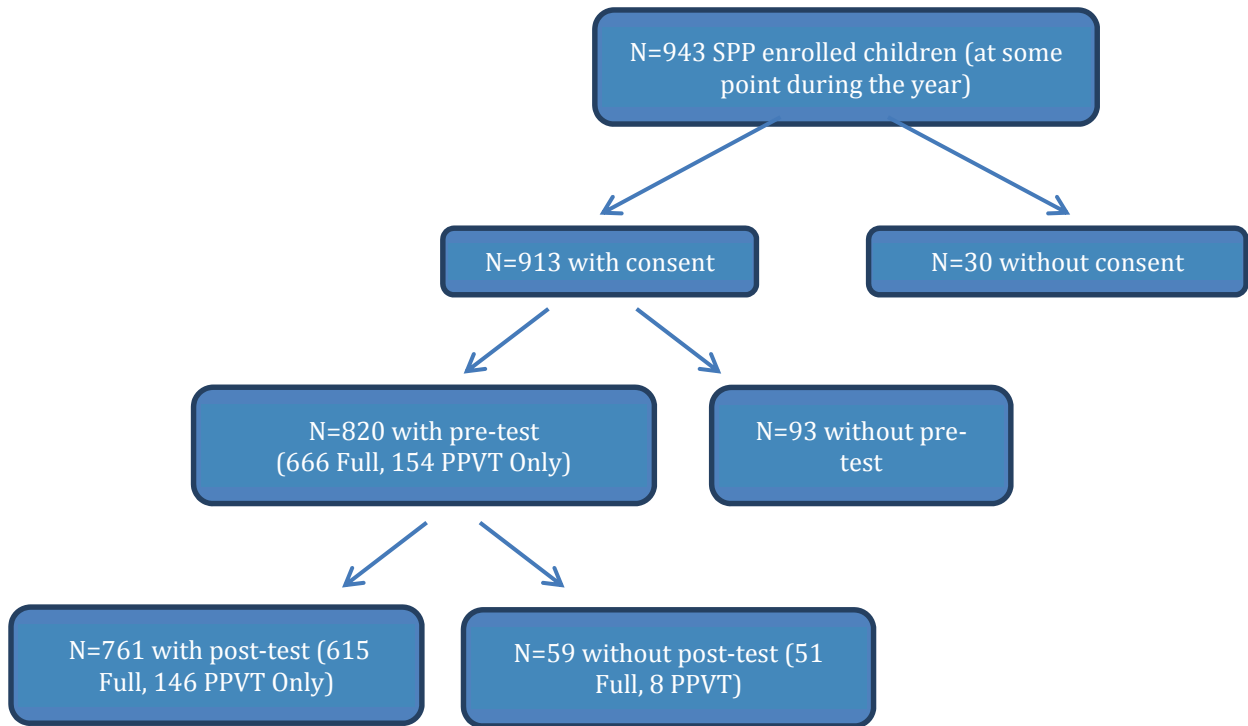
We conducted classroom observations on the 48 SPP classrooms and 13 SPP family child care providers (FCC) in the Spring of 2018. SPP Classrooms and FCCs are described in Table 2. Classrooms in SPP in Year 2 use either Creative Curriculum or HighScope Curriculum, they reported an average class size of about 18 (17.77 in the Spring and 17.31 in the Fall), and they were distributed across eleven agencies, with about 4 classrooms per agency. FCCs are smaller in size with average class sizes of about 9 (not all preschool children), and all of them using Creative Curriculum. Teacher qualifications and race and ethnicity are also reported in Table 1.

Table 1. SPP Classroom characteristics, N=32

Classroom characteristic		SPP Classroom Frequency or Mean (SD[1])	SPP FCC Frequency or Mean (SD[1])
Curriculum	Creative	18	7
	HighScope	30	6
Class Size[a]		17.77 (2.13)	8.92 (2.18)
Agencies/Hubs		11	2
Teacher Qualifications	Less AA/unspecified	20.84%	76.92%
	AA	6.25%	23.08%
	BA	45.83%	0.00%
	MA	27.08%	0.00%
Teacher Race and Ethnicity[b]	Black	18.75%	84.62%
	Hispanic	16.68%	0.00%
	White	33.33%	7.69%
	Asian	14.58%	0.00%
Average No. Classrooms per Agency/Hub		4.36 (4.46)	6.50 (0.71)

[a]Number of children in classroom as reported by director/roster in the Spring (and for FCCs, in the Winter).
[b]Percentages do not add to 100% as information was not available for all teachers.

Measures

Measures on Children

The *Peabody Picture Vocabulary Test—Fourth Edition (PPVT-IV;* Dunn & Dunn, 2007) is a 228-item test of receptive vocabulary in standard English predictive of general cognitive abilities. The test is adaptive and can be used with population ages 2.5 and above. The test has proven reliability based on reported split-half reliabilities or test-retest reliabilities, as well as concurrent validity (e.g., Qi, Kaiser, Milan, & Hancock, 2006). Results on the PPVT have been found to be strongly correlated with school success (Blair & Razza, 2007; Early, et al., 2007). The test is standardized to a mean of 100 and a standard deviation of 15.

The *Woodcock-Johnson Psycho-Educational Battery—Third Edition (WJ-III;* Woodcock, McGrew, Mather, & Schrank, 2001) includes several subtests. Two of these were used in this study: the *Applied Problems* and *Letter-Word Identification* subtests. WJ is also adaptive and for use with populations above the age of 3. The WJ has shown correlations with other tests of cognitive ability and achievement ranging between 0.60 and 0.70. This measure has been used in numerous large-scale preschool studies (e.g., Early, et al., 2007; Wong, Cook, Barnett, & Jung, 2008). The test is standardized to a mean of 100 and a standard deviation of 15.

The *Dimensional Change Card Sort Task* (DCCS; Zelazo, 2006) engages children in reverse categorization by sorting a set of cards based on different criteria provided by the examiner. The test assesses attention-shifting, as well as short term memory. Scores on the DCCS reflect a pass/fail system on three levels of increasing difficulty, and raw scores range

---

[1] SD stands for standard deviation, which is a measure of variation in the data. That is, it measures how close together or spread apart the classrooms are relative to the mean. The larger the value, the farther apart from the mean classrooms are, and the smaller the value, the closer to the mean classrooms are, in a specific indicator, such as classroom size.

between 0 and 3 based on these levels. There are no standard score equivalents. However, in a study of test-retest reliability, means by age for children age 48 months or younger were 1.14, for 48–50 months they were 1.33, for 51–53 months they were 1.42, and for 54–56 months they were 1.58 (Meador et al., 2013).

The *Peg Tapping Test* (PT; Diamond & Taylor, 1996) asks children to tap a peg twice when the experimenter taps once and vice versa. The task requires children to inhibit a natural tendency to mimic the experimenter while remembering the rule for the correct response. Sixteen trials are conducted with 8 one-tap and 8 two-tap trials in random sequence. The task requires two abilities: (a) the ability to hold two things in mind—the rule to tap once when experimenter taps twice and the rule to tap twice when experimenter taps once, and (b) the ability to exercise inhibitory control over one's proponent behavior, the natural tendency to mimic what the experimenter does. The final score for Peg Tapping is a sum of all the 16 items that comprise the test. Again, while there are no standard score equivalents, in a study of test-retest reliability, means by age for children age 48 months or younger were 4.05, for 48–50 months they were 4.57, for 51–53 months they were 6.02, and for 54–56 months they were 7.87 (Meador et al., 2013).

Measures on Classrooms

Early Childhood Environment Rating Scale—Third Ed. (ECERS-3; Harms, Clifford & Cryer, 2014). The ECERS-3 is an observation and rating tool for preschool and kindergarten classrooms measuring environmental factors and teacher-child interactions. It emphasizes the role of the teacher in relation to environment and children's developmental gains. The overall ECERS-3 score is an average on 35 items under 6 domains, which are each rated in a scale between 1 and 7. A rating of 1 indicates inadequate quality, a rating of 3 indicates minimal quality, a rating of 5 indicates good quality, and a rating of 7 indicates excellent quality. A general description of each of the 35 items on the ECERS-3 is provided in Appendix Table A.1. A recent validation paper (Early, et. al, 2018) reports a four-factor (Learning Opportunities, Gross Motor, Teacher Interactions, and Math Activities) structure to the ECERS-3, found moderate correlations with the three CLASS Pre-K domains, and positive associations with growth in children's executive functions (while not with children's cognitive measures). The ECERS-3 was only used in classrooms in center-based care.

Classroom Assessment Scoring System Pre-K (CLASS Pre-K; Pianta, La Paro, & Hamre, 2008). The CLASS Pre-K is an observational tool that identifies the classroom interactions that promote children's development and learning. Observations consist of four 20-minute cycles, with 10-minute coding periods between each cycle, which are then averaged for an overall quality score. Interactions are measured through 10 dimensions in three domains. The Emotional Support domain is measured by four dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives. The Classroom Organization domain is measured by 3 dimensions: Productivity, Behavior Management, and Instructional Learning Formats. The Instructional Support domain is measured by three dimensions: Concept Development, Quality of Feedback, and Language Modeling. Each scale uses a 7-point Likert-

type scale, for which a score of 1 or 2 indicates low quality, and a score of 6 or 7 indicates high quality. The CLASS domains and dimensions are outlined in Appendix Table A.2.

Because a CLASS instrument does not exist for mixed aged groupings, Family child care providers were observed with three CLASS instruments using a *Combined CLASS Protocol* (Joseph, Feldman, Phillips & Jackson, 2010),[2] which was designed to be used in any child care facility in a home, with multiple age groups. This protocol integrates the dimensions from Infant, Toddler, and Pre-K CLASS. There are three dimensions that apply only to pre-K children: Productivity, Instructional Learning Formats, and Concept Development. All other dimensions apply to children of different age groups, depending on which children are present.[3] In addition, the combined protocol includes a new dimension, that of Facilitation of Learning and Development, from the CLASS protocols for children under 3. Observers using the combined protocol are trained and reliable in all three CLASS instruments, and the items on the combined protocol draw from the corresponding the Infant CLASS Manual, the Toddler CLASS Manual, and the Pre-K CLASS Manual, which are used by observers throughout the process. The protocol requires paying attention to children of all ages. Therefore, if differentiation by age does not adequately occur (e.g. adequate language modelling is observed for infants and toddlers but not for preschoolers), scores will reflect the average for the whole age-group served, rather than only preschool children.[4] Further information is provided in Appendix Table A.3.

Procedures

Data collection processes were conducted by Cultivate Learning (CL) at the University of Washington. The center trained data collectors on standardized child assessments and classroom observation measures. Data collectors received a two-day training on the measures for child assessments, were given several days to practice, and were then tested for reliability on the assessments before starting data collection.

Observations of classroom quality were conducted by trained and reliable observers. Initial training in administering the observation protocol included the ECERS-3 and the CLASS protocols. ECERS-3 observers were trained by an ECERS-3 certified trainer and met the ERSI[5] reliability requirements for observer certification. The trainee must complete three observations with the trainer with an average of 85% or above exact matches or one-away from the true score. All data collectors met the ECERS-3 reliability requirements with agreement percentages ranging between 89–94%. CLASS observers were trained by a CLASS certified trainer and met the Teachstone reliability certification requirements. CLASS reliability[6] agreement percentages ranged between 93–100%. Assessment and observation score sheets were cleaned and entered at CL by trained staff. Language accommodations were made as necessary in the requested language (N=29). Assessment procedures incorporated culturally sensitive attitudes, knowledge,

---

[2] Protocol designed for Washington State's QRIS, Early Achievers. Also used in Oregon, see Tout, et. al, (2017).
[3] If a given age group is not present or sleeping during the observation, the particular age group will not be considered when scoring.
[4] Although there may be benefits of mixed age-grouping that the CLASS is not designed to capture.
[5] ERSI is the company that sells ECERS-3; for information on the tool and reliability go to http://www.ersi.info/
[6] Teachstone is the company that sells CLASS products and manages CLASS certifications. All training activity is monitored and reported to them. http://www.teachstone.com/about-teachstone/.

interview skills, intervention strategies and evaluation practices specifically informed by the age of the children in the study. Satisfaction surveys were delivered after data collection to providers to follow up on the procedures followed by data collectors, their interactions with the sites, and whether the experience was overall positive and responses to these questions were quite positive.[7]

Methods

To address the descriptive questions on classroom quality and change over time, or differences across types of providers, data were collected and analyzed from the ECERS-3 and the CLASS. Two tailed two sample t-tests assuming unequal variances were used to test changes in quality between years, or between FCCs and Classrooms or to compare the quality received for males versus females. One-way anovas, with Bonferroni multiple-comparison tests are used to tests for differences in quality experienced by different subgroups of children (across race and ethnicity, by language indicators, and by FPL levels).[8]

To address the question concerning children's development over the school year, the child assessments collected from a randomly selected group of children is first described across subgroups and over the years (in terms of standard gains) and then analyzed using multivariate analyses to explore the relationship between children's growth and child demographic information, as well as school and classroom features.

Results

Each of the research questions is addressed individually. Analyses draw from all the SPP classrooms. SPP FCCs are incorporated into comparisons later below in question 3, as well as in analyses in question 4. Questions 3 and 4 also incorporate information on the sample of children in SPP classrooms (although all children were assessed with the PPVT).

1. Who enrolled in SPP in 2017–18, and how do they compare demographically to children in Seattle more generally?

Children's demographics[9] are summarized in Table 2, below, which also summarizes similar demographics for children enrolled in Seattle Public Schools (as these children embody the SPP program target population). Children in the sample were mostly 4-year-olds (74%) and predominantly from English-speaking households (57%), with 29% speaking other languages, including Vietnamese, Amharic, Mandarin, Somali, and Oromo, among others. Children more predominantly represented non-Whites than children in Seattle Public Schools, with 22% White,

---

[7] Only 17 sites answered the survey: 100% agreed data collectors entered the facility and checked in as requested; data collectors' interactions were courteous and professional, and data collectors arrived during the dates/times I expected. 94% found the experience working with the team positive (with one site reporting the testing time was too long).

[8] These categories are limited by what can be identified in this dataset. This is not indicative of importance over other categorizations, nor that there may not be important intersectional groupings as well.

[9] Demographics were provided by DEEL.

27% Black (slightly increasing from last year's 24%), 28% Asian (increasing from last years 17%), 12% Hispanic (also increasing from 8%), and 11% Multiracial/Other. About 77% of the children were under 300% of the Federal Poverty Level (FPL).

Table 2. Child demographics for SPP study children relative to children in Seattle Public Schools

Child Characteristics	SPP Children 2017–18		Seattle Public Schools
	N	%	
Gender			
Female	386	50.7%	51.3%[a]
Male	375	49.3%	48.7%[a]
Age at Pre-Test			
3-Year-Olds	196	25.8%	-
4-Year-Olds	565	74.2%	-
Primary Language			
English	437	57.4%	-
Non-English	219	28.8%	21.7%[a]
Unknown	105	13.8%	-
Income			
20,000 or Less	203	26.7%	
21,000-40,000	157	20.6%	
41,000-60,000	127	16.7%	
61,000-80,000	88	11.6%	
81,000 or more	175	23.0%	
Unknown	11	1.4%	
FPL Percentage			
Less than 100%	236	31.0%	33.9%[a,c]
100 – 199%	157	20.6%	
200 – 299%	190	25.0%	-
≥ 300%	175	23.0%	-
Unknown	3	0.4%	
Race/Ethnicity			
White	164	21.8%	47.2%[a]
Black	200	26.6%	15.0%[a]
Asian	214	28.4%	14.0%[a]
Hispanic	93	12.4%	12.1%[a]
Multi-Racial/Other	82	10.9%	11.7%[a]

[a]Seattle Public Schools as reported in http://www.seattleschools.org/district/district_quick_facts.
[b]Students attending Seattle Public Schools, as reported in Rivers (2016).
[c]Based on Free and Reduce Lunch which is for families <185% FPL.

2. What was the observed quality of children's SPP classroom experiences in 2017–18, and did it improve over the prior year?

Average ECERS-3 Results

ECERS-3 scores for SPP classrooms for 2016 through 2018 are reported in Table 3 below. Mean scores, standard deviations, and minimum and maximum scores are reported for the six ECERS-3 subscales and overall scores. Average ECERS-3 scores and subscale scores in 2018 slightly

increased relative to 2017 (a 0.19 SD increase) even with the program continuing to grow in number of classrooms. Variation also increased. Statistically significant differences in the average compared to the previous year are marked with an asterisk.[10]

Table 3. ECERS-3 Item, Subscale, and Overall Means and Ranges, 2016-2018

ECERS-3 Item and Subscales	Spring 2016 (N=14)				Spring 2017 (N=32)				Spring 2018 (N=48)			
	Mean	(SD)	Min	Max	Mean	(SD)	Min	Max	Mean	(SD)	Min	Max
Overall	**3.57**	**(0.46)**	**2.94**	**4.50**	**3.89***	**(0.55)**	**2.74**	**5.44**	**3.99**	**(0.63)**	**2.47**	**4.94**
Space and Furnishings	3.88	(0.55)	2.86	4.57	3.94	(0.61)	2.71	5.29	**4.25**	**(0.80)**	**2.43**	**5.86**
Personal Care Routines	3.14	(0.65)	1.75	4.25	3.41	(0.86)	1.50	5.50	**2.67**	**(0.85)**	**1.00**	**4.25**
Language & Literacy	3.47	(0.83)	2.40	5.20	3.93	(0.82)	2.40	6.00	**4.22**	**(0.92)**	**2.40**	**5.80**
Learning Activities	2.87	(0.56)	2.10	4.00	3.26	(0.57)	2.40	4.70	**3.45**	**(0.66)**	**2.18**	**4.60**
Interaction	4.49	(0.90)	3.20	5.80	4.99	(1.07)	2.40	6.80	**5.12**	**(0.99)**	**2.60**	**6.60**
Program Structure	4.43	(0.97)	2.67	6.00	4.67	(0.88)	3.00	6.33	**4.76**	**(1.01)**	**2.67**	**6.33**

Average CLASS Scores

Classrooms were observed using the CLASS pre-K. Scores reported below only include overall means for the pre-K classrooms in the SPP program for the spring 2016 through 2018 (scores for the 13 FCCs added as a pilot this year are reported separately below). Table 4 reports mean scores, standard deviations, and minimum and maximum scores for three CLASS domains. All three domains increased in mean scores relative to 2017 (increases were of 0.04SD, 0.41SD and 0.44SD, respectively). Statistically significant differences in the average scores compared to the previous year are indicated in Table 4 by an asterisk.[11]

Table 4. CLASS Domain Means and Ranges, 2016, 2017 and 2018

CLASS Domains	Spring 2016 (N=14)				Spring 2107 (N=32)				Spring 2018 (N=48)			
	Mean	(SD)	Min	Max	Mean	(SD)	Min	Max	Mean	(SD)	Min	Max
Emotional Support	6.14	(0.53)	4.88	6.81	6.29	(0.47)	5.19	7.00	**6.38**	**(0.57)**	**4.19**	**7.00**
Classroom Organization	5.67	(0.74)	4.17	6.58	5.55	(0.76)	3.42	6.83	**5.96***	**(0.77)**	**3.75**	**6.92**
Instructional Support	2.65	(0.71)	1.50	4.25	3.06	(0.88)	1.67	5.75	**3.42***	**(1.05)**	**1.75**	**6.33**

In sum, program has shown continuous improvement in quality as measured by the ECERS-3 (in all areas but the personal care routines scale) and all three CLASS domains. A particularly large gain is observed for instructional supports which are central for building academic success.

---

[10] Two-tailed two-sample t-test assuming unequal variances were used, P-values are reported in Appendix D.
[11] Two-tailed two-sample t-test assuming unequal variances were used, P-values are reported in Appendix D.

Distribution of Classroom Quality across Classrooms

The ECERS-3 and CLASS domains distributions of classroom quality are depicted in Figures 2 and 3 below. While in the spring of 2018, on average, classrooms scored below the good quality threshold of 5 in the ECERS-3, the percentage of classrooms scoring above it increased from 38% in 2017 to 54% in 2018. Classrooms scored high on Emotional Support or ES (92% percent scored above 5.5). Classroom Organization (CO) also had moderately high scores, with a large portion of classrooms scoring above 5.5 (77%). Following national patterns, classrooms scored lower on Instructional Support (IS), with 41% of the classrooms scoring above 3.5. This percentage increased from 25% in 2017.

Figures 2 and 3 present normalized distributions for ECERS-3 and CLASS dimensions for the spring of 2016 (dotted line), 2017 (striped line) and 2018 (solid line). The ECERS-3 distribution of classrooms evidences a larger portion of classrooms scoring higher in the scale but also lower maximum scores and minimum scores.

Figure 2. ECERS-3 distributions of normalized scores, 2016-2018
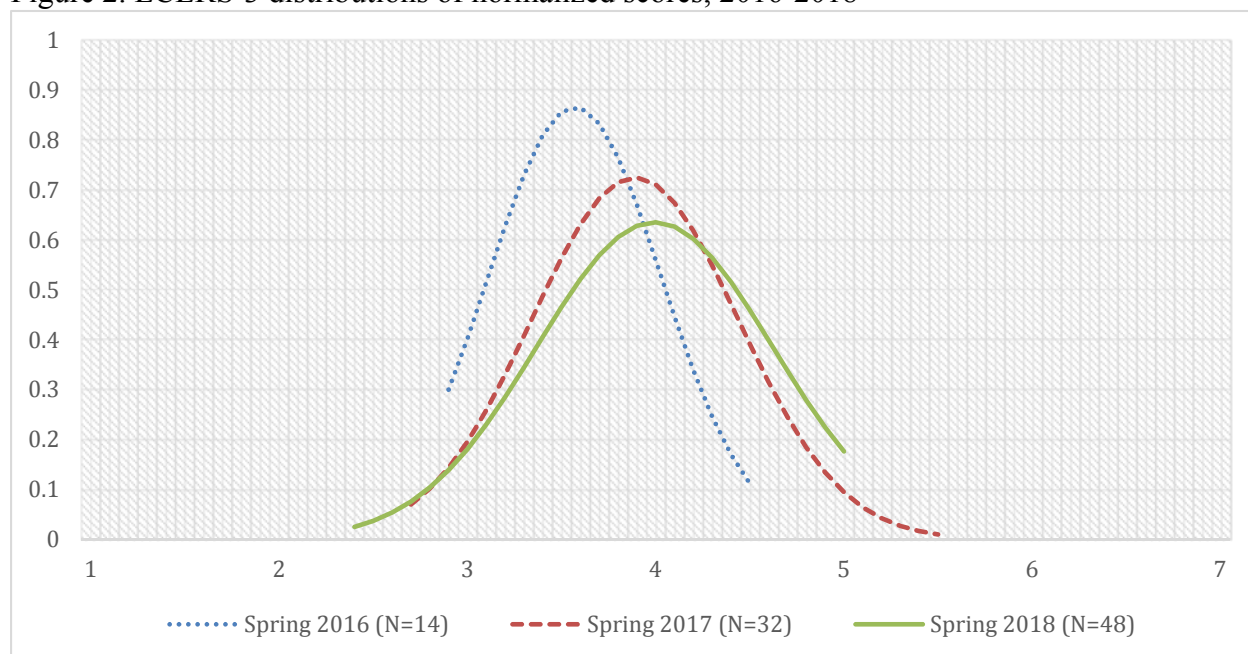
The 2018 CLASS score distributions show an increase in the number of higher-CLASS ES scoring classrooms which drives the increase in average scores even with the minimum scores having decreased (panel a). For CLASS CO (panel b) the distribution shows a shift towards more classrooms scoring in the 5-7 range. For CLASS IS (panel c) there is a shift towards higher scores, and the distribution is starting to spread across the 3-6 score, driving the increase in the mean.

Figure 3. CLASS Domain distributions of normalized scores, 2016-2018

a. CLASS Emotional Support
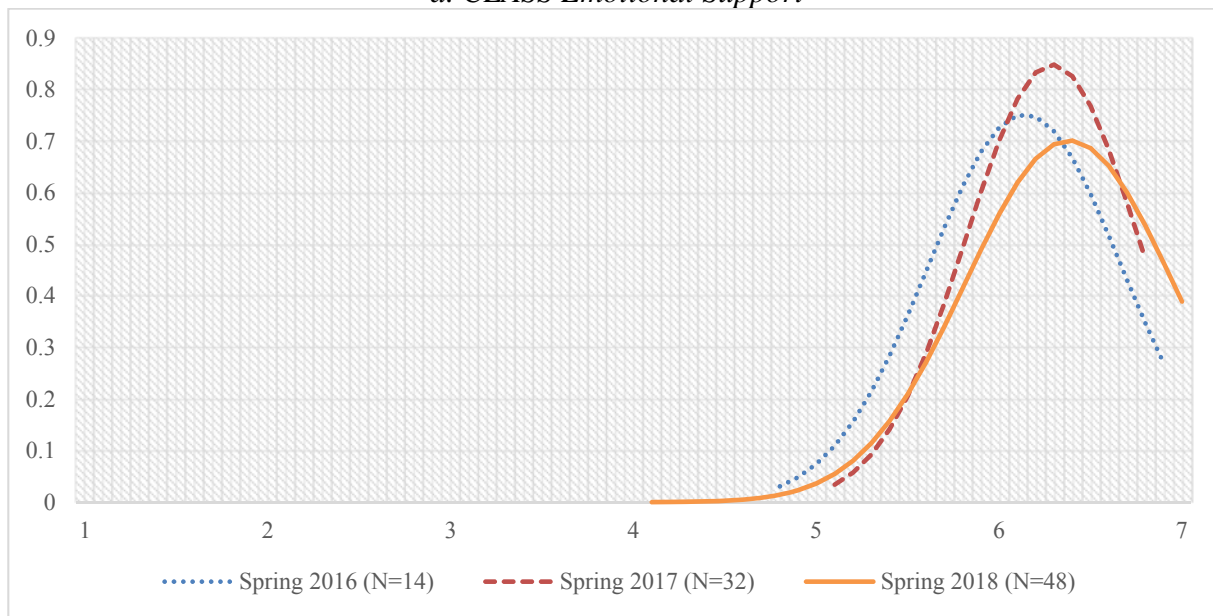
b. CLASS Classroom Organization

c. CLASS Instructional Support

Spring 2016 (N=14)   Spring 2017 (N=32)   Spring 2018 (N=48)
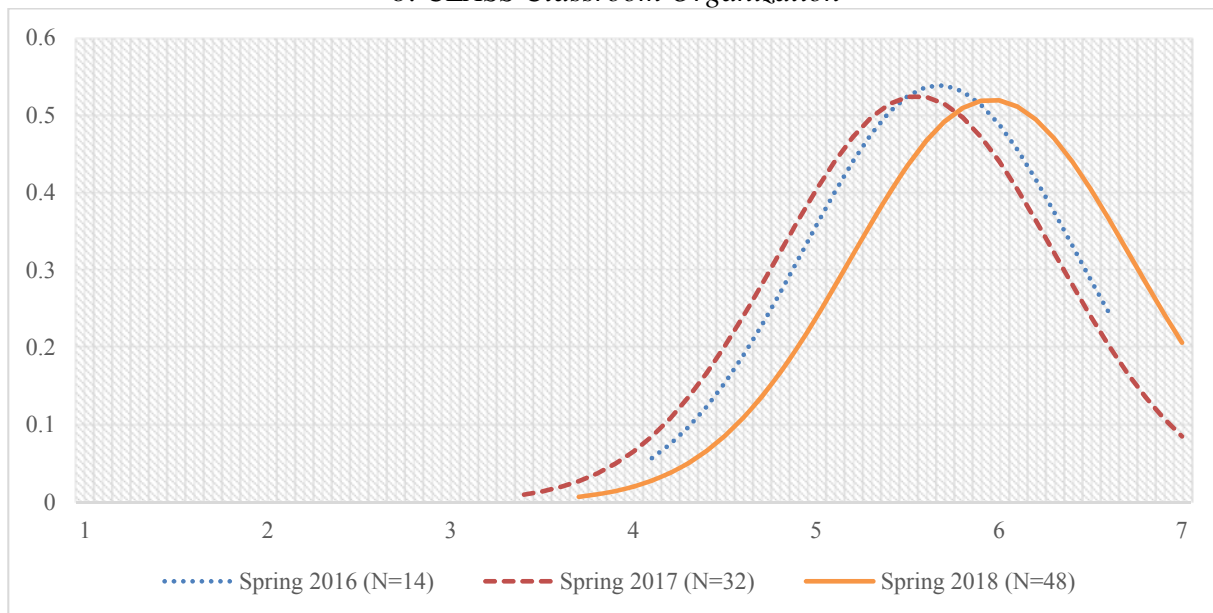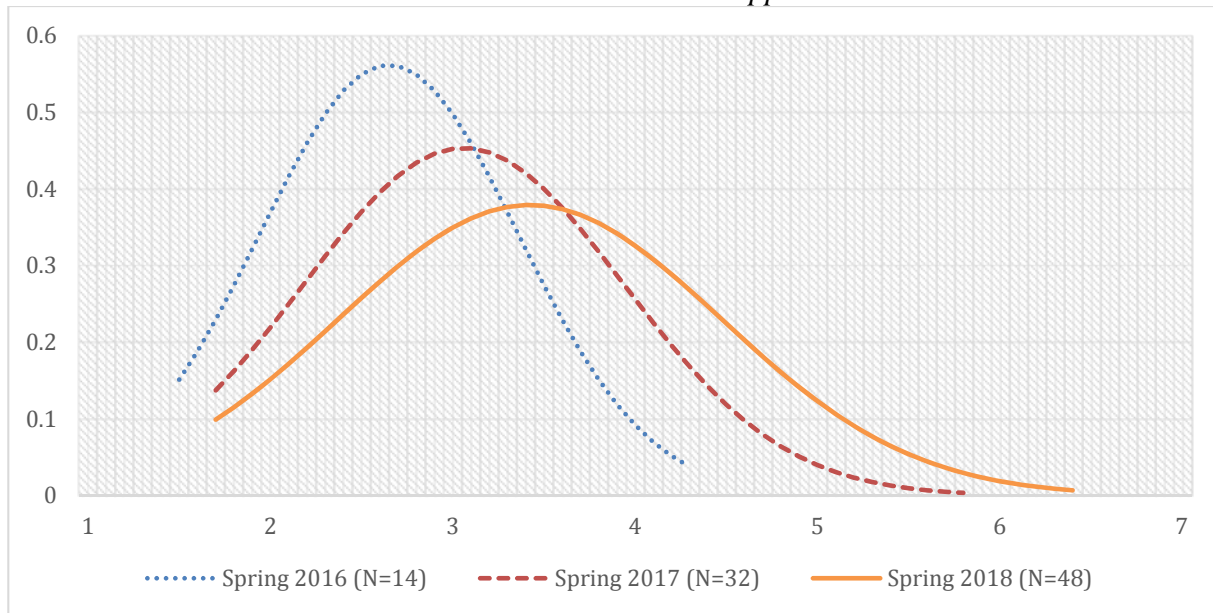
Table 5 and Figure 4 contextualize SPP ECERS-3 scores in relation with other 4 programs/studies: in GA, PA, UW state pre-K and childcare centers and NJ Abbott districts.[12] These are also reported for each subscale (and standard deviations are included when available). SPP classrooms score on average closely to NJ Abbott's average score for Space and Furnishings, Interaction and Program Structure. Areas that underperform the most relative to NJ Abbott are Personal Care Routines, and Learning Activities. This is depicted in Figure 4.

---

[12] The ECERS-3 is still not as widely used as the ECERS-R, which does not allow for comparisons with many high-quality programs.

Table 5. Studies with reported ECERS-3 scores

Study	Space/ Furnishing	Personal Care Routines	Language & Literacy	Learning Activities	Interaction	Program Structure	Average Total
SPP 2018 (N=48)	**4.25**	**2.67**	**4.22**	**3.45**	**5.12**	**5.12**	**3.99**
	(0.80)	(0.85)	(0.92)	(0.66)	(0.99)	(1.01)	(0.63)
SPP 2017 (N=32)	**3.94**	**3.40**	**3.93**	**3.26**	**4.99**	**4.67**	**3.89**
	(0.61)	(0.86)	(0.82)	(0.57)	(1.07)	(0.86)	(0.55)
SPP 2016 (N=12)	**3.88**	**3.14**	**3.47**	**2.87**	**4.49**	**4.43**	**3.57**
	(0.55)	(0.65)	(0.83)	(0.56)	(0.90)	(0.97)	(0.46)
GA[1]	3.49	3.14	3.36	3.14	4.31	3.64	**3.46**
UW state pre-K & childcare (2013-14) (N=299)[2]	3.45	2.89	3.40	2.68	3.88	3.63	**3.23**
PA[3]	3.74	3.77	3.77	2.93	4.72	4.10	**3.68**
GA, PA, WA (2015-16) (N=1063)[4]	3.62	3.36	3.62	2.97	4.41	3.92	**3.53**
NJ Abbott: 2016–17 (N=300)[5]	4.20	4.26	4.70	4.17	5.17	5.02	**4.48**
	(0.84)	(1.14)	(1.10)	(1.11)	(1.30)	(1.38)	**(0.92)**
2015–16 (N=293)[6]	4.43	4.36	4.86	4.22	5.26	5.20	**4.61**
	(1.02)	(1.33)	(1.26)	(1.17)	(1.34)	(1.31)	**(1.03)**

[1]Jenson (2015); [2]CQEL (Unpublished); [3]PAKEYS (Unpublished); [4]Early et. al (2018), subscales estimated from paper [5]NIEER (2017); [6]NIEER (2016).

Figure 4. SPP ECERS-3 scores by subscale in relation to other programs
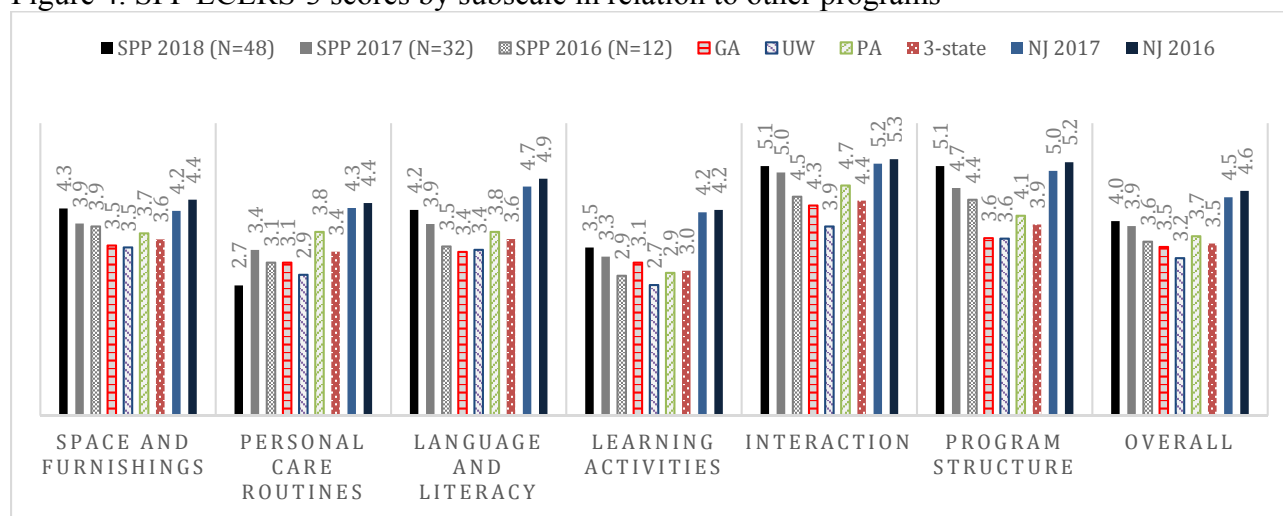
Table 6 and Figure 5 below report CLASS scores for SPP classrooms, 2016-2018, and for selected preschool programs. Seattle is quite at par with the highest scoring programs in the CLASS Emotional Support and Classroom Organization domains (New York and San Antonio) and has increased its CLASS Instructional Support domain scores enough to be the just below previous levels in San Antonio PreK (which actually dropped in scores last year) and the Boston Pre-K program.

Table 6. Classroom quality across the nation, and for selected programs

Study	Emotional Support	Classroom Organization	Instructional Support
SPP classrooms 2018 (N=48)	**6.38 (0.57)**	**5.96 (0.77)**	**3.42 (1.05)**
SPP classrooms 2017 (N=32)	**6.29 (0.47)**	**5.55 (0.76)**	**3.06 (0.88)**
SPP classrooms 2016 (N=14)	**6.14 (0.53)**	**5.67 (0.74)**	**2.65 (0.71)**
Tulsa[1]			
TPS pre-k (N=77)	5.23 (0.57)	4.96 (0.69)	3.21 (0.93)
CAP Head Start (N=28)	5.22 (0.78)	4.80 (0.84)	3.26 (0.94)
Boston[2] (N=83) (2009-2010)	5.63 (0.60)	5.10 (0.68)	4.30 (0.84)
NYC (N=1,570) (2016–17)[3]	6.40	6.20	3.10
NYC (N=1,134) (2015–16)[4]	6.20	6.10	3.30
NYC (N=555) (2012-13 to 2014-15)[5]	6.00	5.80	3.60
National Head Start Overview 2015[6]	6.03 (0.28)	5.80 (0.36)	2.88 (0.54)
Head Start FACES 2009[7]	5.30	4.70	2.30
EA Validation study (N=75) (2013-2014)[8]	5.96 (0.66)	5.26 (0.77)	2.34 (0.71)
NJ Abbott 2013-2014 (N=163)[9]	5.97 (0.63)	5.32 (0.89)	3.15 (0.96)
San Antonio (N=89) (2017)[10]	6.24 (0.52)	5.60 (0.79)	3.55 (1.32)
San Antonio (N=89) (2016)[11]	6.44 (0.51)	5.98 (0.81)	3.67 (1.23)
San Antonio (N=76) (2015)[12]	6.34 (0.64)	5.93 (0.97)	3.02 (1.14)
San Antonio (N=36) (2014)[13]	6.28 (0.35)	5.75 (0.60)	2.82 (0.82)

[1]Phillips et. al (2009); [2]Weiland et. al (2013); [3]NYC Department of Education (2018); [4]NYC Department of Education (2017); [5]NYC Department of Education (2016); [6]Office of Head Start. (2015); [7]Aikens et. al (2013); [8]CQEL (Unpublished); [9]NIEER (2014); [10]EDVANCE (2017); [11]EDVANCE (2016); [12]EDVANCE (2015); [13]EDVANCE (2014).

Figure 5. SPP CLASS scores by domain in relation to other programs
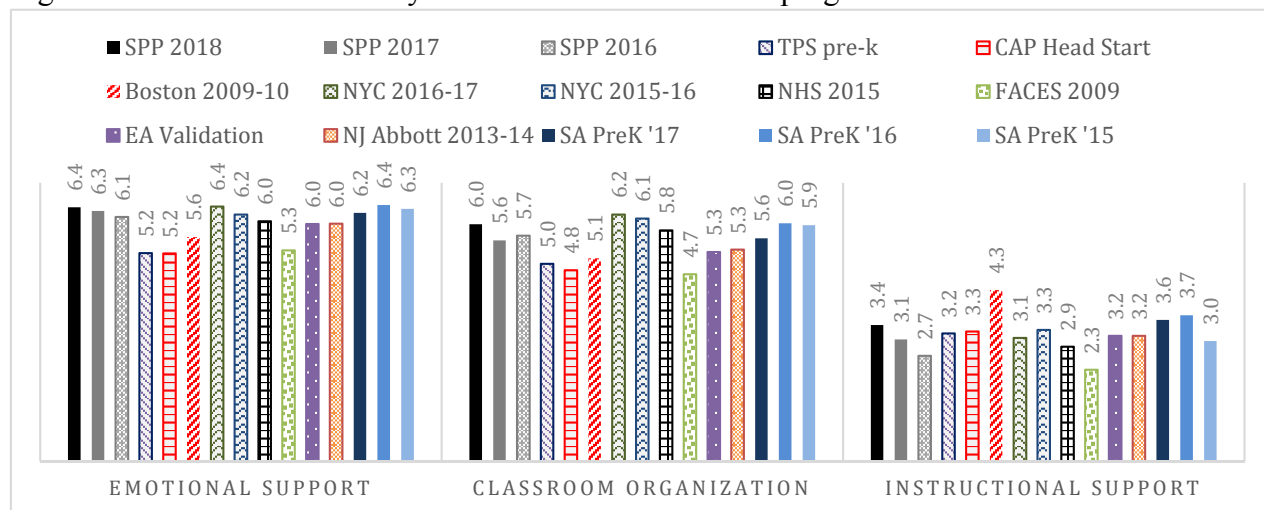
ECERS-3 subscales

Items and subscales for the ECERS-3 are reported in Table 7 for 2016, 2017, and 2018, including the average scores and the ranges, which illustrate the minimum and maximum scored by classrooms.

The *Space and Furnishings subscale* incorporates whether children have enough space and furniture, whether the arrangement of the furniture allows for learning and exploration and whether displays are meaningful and representative of the children in the class. The items for "space for gross motor play" and "gross motor equipment" evidence made up the lowest scores in this subscale. "Indoor space" score have decreased consistently as more programs have been added. All other items have increased scores through the years. Four items under this subscale continue to range starting at 1, indicating classrooms scoring at the inadequate rating. In contrast, this year five items under this subscale showed classrooms scoring at the excellent level.[13]

The *Personal Care Routines* subscale, addresses health, hygiene and safety practices in the classroom. Under *personal care routines*, only "safety practices" is above the minimal threshold score of 3. All other score on average are under it, at the inadequate level. All items in this subscale show reductions in scores with the addition of classrooms. In all items in this scale there were classrooms scoring at "1" (inadequate) and at "7" (excellent).

Language and Literacy focuses on how staff direct activities and materials towards supporting children's development of their language and literacy skills. All but one item under this subscale have continued a positive trend in relation to the previous two years. "Becoming Familiar with Print" remains the lowest scoring item (3.44).[14] The item for "Staff Use of Books" averaged 3.79 this year (up from 3.07 in 2016).[15] On three items in this scale there were classrooms scoring at "1" (inadequate), while in the other two the minimums were at 2 and 3 (minimal). Maximum scores were "7" (excellent) in four items this year and only "5" (good) in one item.

Learning Activities includes the presence, variety, and accessibility of learning materials in the classroom for children, and at the same time captures the extent to which teachers actively engage children with different types of materials. Under this subscale, the average for "fine motor," "art," and "math in daily events" were the highest, 4.88, 4.15, and 4.29, respectively. While still not reaching the level of "good" (5.00), "math in daily events" increased one full point. In the other areas, scores are lower, and three items remain under the minimal score of 3: "nature/science," "math materials and activities," and "understanding written numbers." In seven of the ten items there was improvement relative to 2017.

The *Interaction* subscale assesses children's supervision during gross motor time, teachers' individualization of teaching and learning and interactions between children and teachers. Three items under this subscale showed continued positive trends since 2016, and all two items now score in the good level: "individualized teaching and learning" and "staff-child interaction." For all items there are classrooms scoring at 7 (excellent).

---

[13] "Space for gross motor" and "gross motor equipment" have a time requirement of 15 minutes to receive credit in the "minimal" category of scoring and 30 minutes for "good." It does not count e.g. walking time to the playground.
[14] This item expects observing visible print being combined with pictures and staff taking dictation of children's words in a way that is interesting and engaging to children for the purpose of showing print as a useful tool.
[15] A score in the good (5) to excellent (7) on this item is attained when all children are observed to be actively engaged during story time.

The *Program Structure* subscale is centered on the general formats of the classroom and how children spend their time. "Transitions and waiting times" and "free play" showed increases on average scores, now averaging 5.21 and 4.58.

Table 7. ECERS-3 Item, Subscale, and Overall Means and Ranges by Item, 2016, 2017 and 2018

ECERS-3 Item and Subscales	2016 Mean (Range) N=14	2017 Mean (Range) N=32	2018 Mean (Range) N=48
Space and Furnishings			
1. Indoor space	6.43 (4-7)	5.47 (2-7)	5.40 (2-7)
2. Furnishings for care, play and learning	4.36 (4-7)	4.56 (3-7)	4.44 (3-7)
3. Room arrangement for play and learning	3.64 (2-7)	4.72 (2-7)	5.04 (2-7)
4. Space for privacy	4.14 (2-6)	4.53 (1-7)	4.63 (1-7)
5. Child-related display	3.36 (1-5)	3.09 (1-4)	4.29 (1-7)
6. Space for gross motor play	3.14 (1-4)	3.06 (1-6)	3.10 (1-4)
7. Gross motor equipment	2.07 (1-4)	2.13 (1-5)	2.81 (1-6)
Personal Care Routines			
8. Meals/ snacks	3.07 (1-4)	3.88 (1-7)	2.90 (1-5)
9. Toileting/diapering	2.21 (1-3)	3.19 (1-7)	2.79 (1-6)
10. Health practices	2.93 (2-4)	2.69 (1-5)	1.88 (1-5)
11. Safety practices	4.36 (2-7)	3.88 (1-7)	3.13 (1-7)
Language and Literacy			
12. Helping children expand vocabulary	3.50 (3-5)	3.63 (1-7)	4.63 (3-7)
13. Encouraging children to use language	4.36 (3-7)	4.84 (3-7)	5.15 (2-7)
14. Staff use of books with children	3.07 (1-6)	3.50 (1-6)	3.79 (1-7)
15. Encouraging children's use of books	4.21 (1-7)	4.41 (3-6)	4.08 (1-7)
16. Becoming familiar with print	2.21 (1-4)	3.25 (1-6)	3.44 (1-5)
Learning Activities			
17. Fine motor	4.36 (2-5)	4.47 (2-7)	4.88 (2-7)
18. Art	3.71 (2-6)	4.28 (1-7)	4.15 (1-7)
19. Music and movement	3.50 (2-5)	3.47 (2-6)	3.58 (1-5)
20. Blocks	2.00 (1-4)	2.97 (1-5)	3.13 (1-7)
21. Dramatic Play	2.79 (1-6)	3.50 (1-7)	3.77 (1-7)
22. Nature/science	2.50 (1-4)	2.28 (1-5)	2.73 (1-6)
23. Math materials and activities	1.71 (1-3)	2.25 (1-4)	2.42 (1-6)
24. Math in daily events	2.86 (1-5)	3.34 (1-5)	4.29 (1-7)
25. Understanding written numbers	1.29 (1-2)	1.69 (1-5)	1.44 (1-3)
26. Promoting acceptance of diversity	4.21 (3-6)	4.34 (2-6)	4.06 (3-6)
Interaction			
27. Appropriate use of technology	N/A (1-1)*	N/A	N/A
28. Supervision of gross motor	3.71 (1-7)	4.56 (1-7)	4.67 (1-7)
29. Individualized teaching and learning	4.21 (3-7)	4.94 (2-7)	5.33 (3-7)
30. Staff-child interaction	4.93 (3-7)	5.66 (3-7)	5.96 (2-7)
31. Peer interaction	5.00 (3-7)	4.84 (1-7)	4.85 (1-7)
32. Discipline	4.57 (2-7)	4.97 (2-7)	4.77 (2-7)

Program Structure			
33. Transitions and waiting times	4.86 (3-7)	4.75 (3-7)	5.21 (2-7)
34. Free play	4.50 (3-6)	4.44 (2-7)	4.58 (3-7)
35. Whole - group activities for play and learning	3.93 (2-5)	4.81 (2-6)	4.50 (2-6)

Note: (*) Only 2 classrooms received a score for #27, both were 1. All others were N/A.

CLASS: Emotional Support Domain

Table 8 shows the scores for dimensions under the three CLASS domains. The Emotional Support (ES) domain assesses teacher's promotion of a nurturing and safe environment for children to learn. All dimensions in this domain scored on average above 6. The "Positive Climate" and "Negative Climate" dimensions focus on the emotional connection between teachers and students.[16] Negative Climate scores have been inverted in this report, and scores the highest (6.94), indicating a lack of expressed negativity. The dimension on "Teacher Sensitivity" assesses whether teachers anticipate problems and are able to support children effectively (average 6.23, increasing from 6.04 in 2017). This high range score implies consistency in teachers' awareness of children who need assistance or support, responsiveness to their needs, abilities, problems and emotions, providing individualized support, and generally helping children feel comfortable to seek support and share thoughts. "Regard for Student Perspectives" (average 6.04, with a slight increase from 5.96 in 2017) assesses the degree to which teachers follow children's interests, motivations, and perspectives and encourage student responsibility and autonomy. More consistent opportunities for children to have time to express themselves and move about freely in the classroom, to receive encouragement from the teacher, and to have their interests acknowledged by the teacher, would bring this score even higher.

CLASS: Classroom Organization Domain

The Classroom Organization domain focuses on how teachers manage and redirect behavior, how they manage instructional time and routines, and how they manage activities to expand students' interests. "Behavior Management" assesses whether teachers provide clear behavioral expectations and enforce them consistently, whether they are proactive in preventing problems from arising and effectively redirect misbehavior by focusing on the positive. "Productivity" measures teachers' time management, pacing, and transitions throughout the day and across activities and teachers' preparation for activities. "Instructional Learning Formats" measures how teachers' facilitate student learning during activities, including how effective questions are, having clear learning objectives, and using modalities and materials to engage children. All three dimensions in this domain increased in relation to 2017. "Productivity" scored above 6 this year, and the other two dimensions increased by about 0.40 points each.

CLASS: Instructional Supports Domain

The Instructional Supports Domain assesses the interactions through which teachers enable high-order thinking skills, provide feedback, encourage creativity and reasoning, and promote

---

[16] Positive Climate "reflects the emotional connection between the teacher and students and among students and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions" (Pianta, La Paro & Hamre, p.23). Negative Climate "reflects the overall level of expressed negativity in the classroom" (p. 28).

language development. This domain is the most important in terms of teacher practices that impact on student's learning. It has also proven to be the most challenging. In every published study of pre-K quality, scores on the Instructional Support domain lag the other two domains. Therefore, the pattern has been that it scores lower than the other two across programs as seen above. Two of the three dimensions under this domain maintained its positive trend and continued to increase this year.

"Concept Development" gauges teachers' use of discussions to stimulate reasoning, analysis, and understanding. It also measures teachers' ability to ask questions that encourage children to plan, to connect concepts to their lives, and to integrate information with prior knowledge. Consistency and intentionality are central. Concept Development scored the lowest (average 2.63). Increasing scores in this dimension requires more consistent use of discussions and activities to foster problem solving, prediction, comparison, planning and real-world applications. "Quality of Feedback" (average 3.40) assesses the degree to which teachers' scaffold, engage in of feedback loops, and utilize metacognitive approaches with children, encouraging children to think and explain their thinking. "Language Modeling" measures the quality and quantity of teacher's language used to promote children's language development (average 4.19, up from 3.57). This dimension increased the most under IS, with some classrooms now scoring at 7 (excellent).

Table 8. CLASS Domains and Dimensions Means and Range by Item, 2016, 2017 & 2018

CLASS Dimensions and Domains	2016 Mean (Range) N=14	2017 Mean (Range) N=32	2018 Mean (Range) N=48
Emotional Support Domain	6.14 (4.88-6.81)	6.29 (5.19-7.00)	6.38 (4.19-7.00)
1. Positive Climate	5.80 (4.25-7.00)	6.33 (5.25-7.00)	6.23 (3.00-7.00)
2. Negative Climate*	6.86 (5.75-7.00)	6.95 (6.63-7.00)	6.94 (5.00-7.00)
3. Teacher Sensitivity	5.91 (4.25-6.75)	6.04 (4.25-7.00)	6.23 (4.00-7.00)
4. Regard for Student Perspectives	5.96 (4.25-7.00)	5.96 (4.25-7.00)	6.04 (4.00-7.00)
Classroom Organization Domain	5.67 (4.17-6.58)	5.55 (3.42-6.83)	5.96 (3.75-6.92)
5. Behavior Management	5.73 (3.75-7.00)	5.46 (3.50-6.75)	5.98 (3.00-7.00)
6. Productivity	6.05 (4.50-7.00)	5.91 (3.50-7.00)	6.06 (4.00-7.00)
7. Instructional Learning Formats	5.21 (3.50-6.50)	5.21 (3.00-6.75)	5.69 (3.00-7.00)
Instructional Support Domain	2.65 (1.50-4.25)	3.06 (1.67-5.75)	3.42 (1.75-6.33)
8. Concept Development	2.07 (1.25-3.50)	2.64 (1.25-5.50)	2.63 (1.00-6.00)
9. Quality of Feedback	2.61 (1.50-4.25)	3.03 (1.50-5.50)	3.40 (2.00-6.00)
10. Language Modeling	3.29 (1.75-5.00)	3.57 (1.75-6.25)	4.19 (2.00-7.00)

Note: (*) The Negative Climate dimension was transposed so that on here, high represents "good".

3. How does quality vary within SPP and do children from different backgrounds experience different quality?

Classroom quality for Classrooms and FCCs separately

We also looked at whether classrooms and FCCs differed in quality as measured by the CLASS. While the measures somewhat differ, the standard for what levels are necessary for quality is

consistent across all versions of the CLASS.[17] Average CLASS scores by domains are reported in Table 9 for classrooms and for FCCs. Distributions are depicted in Figure 7. CLASS CO and ES were higher on average on SPP classrooms than SPP FCCs, and these distributions are further to the right. Overall, there were no statistically significant differences in mean scores across domains or dimensions.[18] This was also the case for the three dimensions that are scored only with pre-K children in the protocol used in FCCs; that is, productivity, instructional learning formats and concept development.

Table 9. CLASS Domain and Dimension scores for classrooms in centers and FCCs

	Classrooms in Centers				FCCs			
	Mean	(SD)	Min.	Max.	Mean	(SD)	Min.	Max.
Emotional Support	**6.38**	**(0.57)**	**4.19**	**7.00**	**6.03**	**(0.70)**	**4.50**	**6.80**
1. Positive Climate	6.23	(0.88)	3.00	7.00	6.00	(0.91)	4.00	7.00
2. Negative Climate*	6.94	0.32	5.00	7.00	6.92	0.28	6.00	7.00
3. Teacher Sensitivity	6.23	(0.83)	4.00	7.00	5.85	(0.80)	4.00	7.00
4. Regard for Student Perspectives	6.04	(0.62)	4.00	7.00	5.85	(0.80)	4.00	7.00
Classroom Organization	**5.96**	**(0.77)**	**3.75**	**6.92**	**5.52**	**(0.83)**	**3.50**	**6.38**
5. Behavior Management	5.98	(1.04)	3.00	7.00	5.46	(1.13)	3.00	7.00
6. Productivity[a]	6.06	(0.78)	4.00	7.00	5.85	(0.69)	4.00	7.00
7. Instructional Learning Formats[a]	5.69	(0.83)	3.00	7.00	5.31	(0.95)	4.00	6.00
8. Facilitation of Learning & Dev.	n/a	n/a	n/a	n/a	4.15	(1.63)	2.00	7.00
Instructional Support	**3.42**	**(1.05)**	**1.75**	**6.33**	**3.53**	**(0.96)**	**2.25**	**5.50**
9. Concept Development[a]	2.63	(1.20)	1.00	6.00	2.54	(1.13)	1.00	5.00
10. Quality of Feedback	3.40	(1.25)	2.00	6.00	3.31	(1.25)	2.00	6.00
11. Language Modeling	4.19	(1.18)	2.00	7.00	4.38	(0.87)	3.00	6.00

Note: (*) The Negative Climate dimension was transposed so that on here, high represents "good". [a] These three are scored only for pre-K children in the combined protocol used in FCCs.

---

[17] We also estimated alphas for consistency within domains within the CLASS Pre-K used in the 48 classrooms and the CLASS combined used in the 13 FCCs. Both of these were equally consistent (with alphas between 80%-93%).
[18] Two-tailed two-sample t-test assuming unequal variances. P-values in Appendix D.

Figure 6. CLASS Domain distributions of normalized scores for classrooms and FCCs
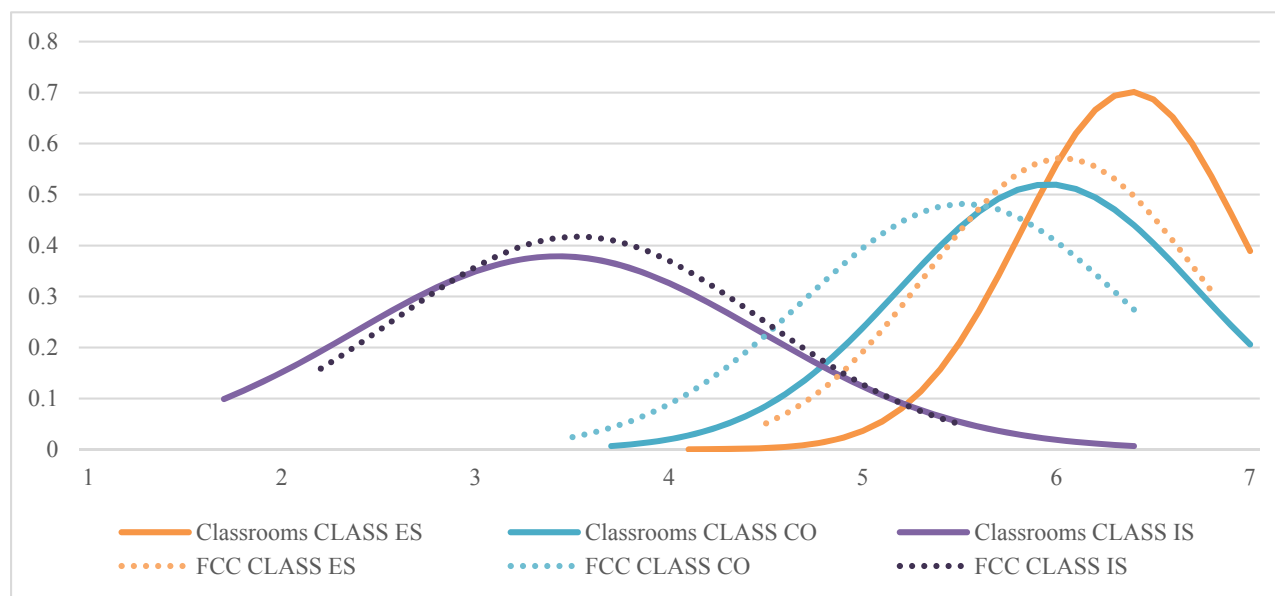
Classroom quality for children from different backgrounds

Figure 7 depicts the quality of care by children's gender, ethnicity/race, language background and FPL for the SPP children in the sample. Tests of statistical significance between groups found no significant differences in quality by gender or language. There were however some differences by race and ethnicity. While there were no differences on the average quality as measured by ECERS by race/ethnicity, there were statistically significant differences in CLASS ES and CLASS CO.[19] On average, children identified as African American or Black experience statistically significantly lower levels of CLASS ES and CLASS CO (at a 5% level) relative to children identified as White. Children identified as Asian, multi-racial or other do so as well for CLASS CO. Children identified as Hispanic experience higher levels of CLASS CO on average than children identified as African American or Black and than children identified as multi-racial or other. Children identified as Hispanic also experience higher average levels of CLASS ES than Black children. For FPL a statistically significant difference was present for CLASS ES and CLASS CO, between families under 100% FPL and families above 300% FPL with children of families above 300% FPL experiencing slightly higher levels of CLASS ES and CLASS CO.

---

[19] One-way anova, with Bonferroni multiple-comparison tests for race/ethnicity, DLL and FPL, and Two-tailed t-test with unequal variances for gender. P-values in Appendix D.

Figure 7. ECERS and CLASS Domain scores by Child Characteristics (N=859 for CLASS, N=910 for ECERS)
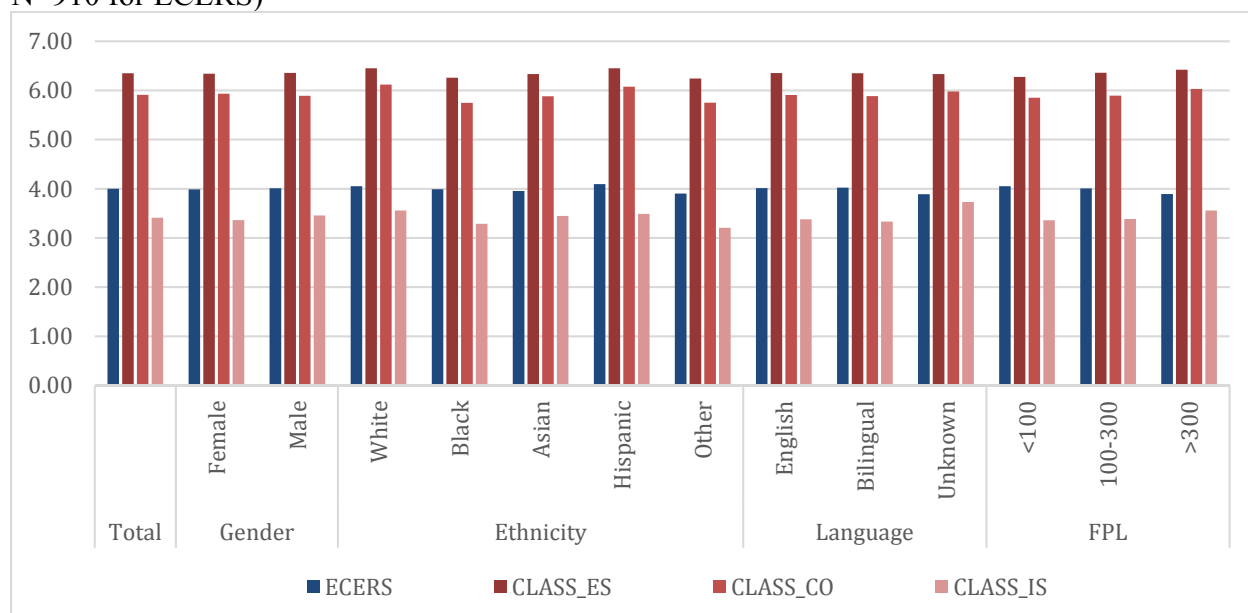
Note: Includes classrooms and FCCs.

Classroom quality by year of entry into SPP

We inquired into whether there were differences in quality between new classrooms in the program, and those with two or three years in the program. Tables 10 and 11 describe ECERS-3 and CLASS scores for classrooms grouped according to the number of years in SPP. Classrooms with three years in the program scored slightly higher on the overall ECERS-3 score than those with two years in the program but did not score higher than new classrooms. Classrooms with three years in the program also scored higher in CLASS ES than classrooms with fewer years in SPP. No clear pattern emerges between year cohort of entry into SPP and scores for the rest of the CLASS domains. Without information on teacher turnover, leadership turnover or other factors that may define individual classroom growth, we cannot identify within this report what factors may be contributing to this lack of patterns.

Table 10. ECERS-3 Subscale, and Overall Means and Ranges, 2017 (N=48)

ECERS-3 Item and Subscales	3 years in SPP (N=9)		2 year in SPP (N=27)		1 year in SPP (N=12)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Overall	4.11	0.59	3.91	0.65	4.11	0.63
Space and Furnishings	4.38	0.79	4.15	0.78	4.37	0.88
Personal Care Routines	3.08	0.79	2.70	0.76	2.29	0.99
Language and Literacy	4.07	0.95	4.10	0.94	4.58	0.80
Learning Activities	3.58	0.68	3.36	0.62	3.57	0.74
Interaction	5.36	0.85	4.99	1.05	5.22	0.96
Program Structure	4.59	0.97	4.63	1.03	5.19	0.97

Table 11. CLASS Domain Means and Ranges, 2018 (N=61)

CLASS Domains	3 years in SPP (N=9)		2 year in SPP (N=27)		1 year in SPP (N=25)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Emotional Support	6.40	0.46	6.38	0.54	6.20	0.72
Classroom Organization	5.79	0.80	5.97	0.80	5.77	0.80
Instructional Support	3.42	1.11	3.22	0.87	3.70	1.13

Associations between program features and quality

Lastly, we also estimated the association between program features and classroom quality through multi-level regression models that accounted for classrooms clustering at the agency level. First, we assessed these associations for the ECERS-3, then for CLASS pre-K only (classrooms) and then for CLASS pre-K and the CLASS combined (in classrooms and FCCs). None of the indicators included showed any statistically significant association with quality, with the exception of a positive association with a teacher meeting or exceeding required qualifications and the ECERS-3 and missing teacher information and CLASS CO and IS levels. Results did not quite vary whether we constrained analyses to classrooms (assessed with the CLASS Pre-K) protocol, or whether we included the FCCs (assessed with the combined protocol).[20] It is critical to acknowledge that the modest number of classrooms provides low statistical power to detect relationships between classroom characteristics and classroom quality. FCCs however did show a negative association with CLASS ES and CO quality, after controlling for classroom characteristics.

Table 12. Association between classroom quality and program features (N=61)

	ECERS	CLASS ES	CLASS ES	CLASS CO	CLASS CO	CLASS IS	CLASS IS
Class Size	0.039	-0.032	-0.044	-0.058	-0.080	-0.055	-0.052
	(0.05)	(0.05)	(0.04)	(0.06)	(0.05)	(0.08)	(0.07)
Creative Curriculum	0.265	-0.020	0.002	-0.262	-0.281	-0.370	-0.274
	(0.19)	(0.18)	(0.16)	(0.25)	(0.21)	(0.31)	(0.26)
Teacher Qual. Meets	0.483	0.306	0.223	0.450	0.361	0.432	0.558
	(0.28)	(0.27)	(0.27)	(0.36)	(0.35)	(0.47)	(0.45)
Teacher Qual. Exc.	**0.513***	0.124	0.127	0.509	0.526	0.297	0.280
	(0.23)	(0.22)	(0.23)	(0.29)	(0.29)	(0.38)	(0.38)
Missing T. Qual.	0.618	0.644	**0.679***	**0.859***	**0.898***	**1.313***	**1.290***
	(0.34)	(0.33)	**(0.34)**	**(0.43)**	**(0.43)**	**(0.57)**	**(0.56)**
T Black	0.130	0.029	-0.172	-0.181	-0.393	-0.637	-0.464
	(0.33)	(0.31)	(0.26)	(0.41)	(0.33)	(0.54)	(0.43)
T Hispanic	-0.104	0.042	-0.058	-0.129	-0.231	-0.298	-0.194
	(0.30)	(0.29)	(0.28)	(0.37)	(0.35)	(0.50)	(0.47)
T Asian	0.029	0.109	-0.008	-0.242	-0.377	-0.512	-0.400
	(0.33)	(0.32)	(0.31)	(0.42)	(0.40)	(0.55)	(0.51)
FCC			**-1.144***		**-1.453***		-0.990
			(0.49)		**(0.62)**		(0.80)
N	48	48	61	48	61	48	61

Note: Omitted groups are teacher not meeting qualifications, teacher identifies as White and classroom is center-based. * p<0.05; ** p<0.01; *** p<0.001.

---

[20] While not shown, estimations without including program features showed a negative association between home provision and CLASS ES and CO. However, this negative association disappeared once we accounted for class sizes, and for teacher education.

4. How did children in SPP classrooms and family child care providers progress in 2017–18, and how did it vary with classroom quality? Other program characteristics? How did it vary with child characteristics?

This evaluation measured child outcomes in receptive vocabulary (using the Peabody Picture Vocabulary Test), literacy (using the Woodcock-Johnson Tests of Achievement Letter-Word subtest), and math (using the Woodcock-Johnson Tests of Achievement Applied Problems subtest). In addition, it measured executive functioning (EF) using two measures: the Dimensional Change Card Sort Game (DCCS) and the Peg Tapping task (PT). The latter two assess a combination of short-term memory, the ability to inhibit automatic response tendencies that can interfere with achieving a task, and the capacity for set shifting.

Child gains for the 2017–18 school year for the all children in the SPP sample (all children were assessed with PPVT and only a random sample was assessed with the rest of the battery) and then for various child subgroups of interest are reported in Appendix B. The PPVT (vocabulary) and Woodcock-Johnson (literacy and math) assessments provide standardized scores that provide comparisons to expected gains after controlling for age. Positive gains in these standard scores indicate that children gained more than other children from a similar background adjusting for age. Overall, children's standard scores increased on all three measures. Children also improved on the executive function measure. The other trends that stand out are: (a) growth in gains across all measures compared to the prior year, except for math (panel c), (b) larger fall to spring gains for children identified as Black, Hispanic, DLL, and low FPL. Gains for the 2016-17 and 2017-18 school years are reported by race/ethnicity, language and FPL in Figure 8 below.

Figure 8. Child gains across the different measures by child demographics

a. Standard Score PPVT gains   b. Standard Score LW gains

c. Standard Score AP gains

d. DCCS gains

e. PT gains
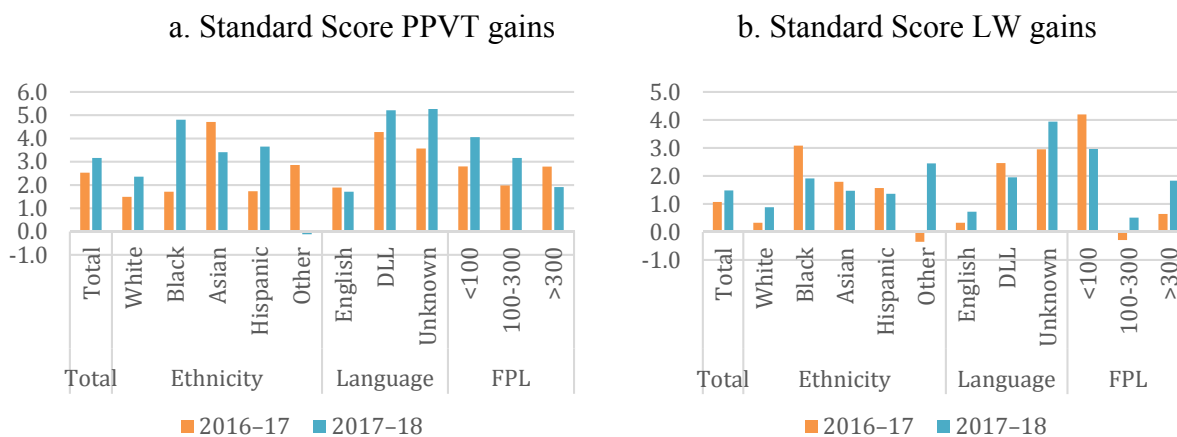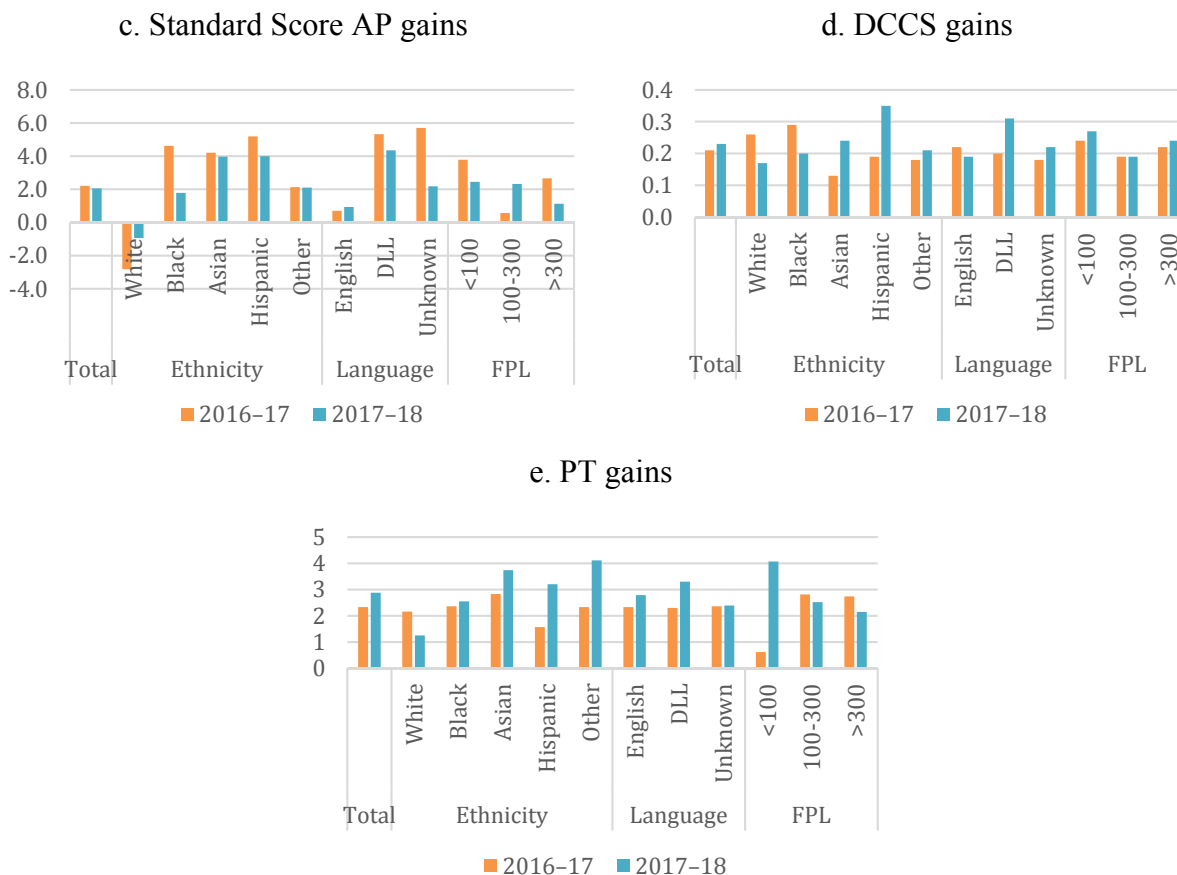
This next section focuses on assessing if differences (if any) in the school year trajectory of children across these subgroups exist and doing so through estimations that relate various children's characteristics to children's gains in the various measures included in the study and controlling for school features.

Multivariate analyses also allow exploring whether there are associations between children's learning gains and program features while taking into account children's characteristics. We incorporate demographics on the children such as their age, gender, race and ethnicity, and home language, as well as household demographics such as income, household size and Federal Poverty Level (FPL). Program features for SPP include class size, agency, curriculum used (whether it is Creative or High Scope) teacher race and ethnicity, teacher degree and classroom quality. We also account for the fact that children that are grouped together in the same classroom or FCC program should not be considered to be independent of each other.

Table 13-15 present the estimates of the associations of program features and child characteristics with children's development. We performed separate analyses with the two measures of quality, one controlling for quality as measured by the ECERS-3 (Table 13), and the other for quality as measured by the CLASS dimensions for classrooms in centers only (Table 14), as well as including FCCs (Table 15). Statistically significant results are highlighted in bold. For categorical variables, such as female, the results need to be interpreted in relation to the omitted group (i.e. males).

In terms of children's characteristics, this year[21] we do not find evidence of disadvantages for children that identify as Black or Hispanic across any of the outcomes. Children identified as Asian evidence lower receptive vocabulary gains (standard and raw) than their peers, yet they outperform their peers in one of the executive function measures (Peg Tapping, or PT).[22] The latter is also the case for children identified as other. No systematic differences were found for dual language children (in comparison to English speaking children), by income or FPL. Agency selected children (usually enrolled by the agency to maintain continuity with previous years) showed higher receptive vocabulary scores.[23] There is no evidence that the program is creating consistent patterns of advantages or disadvantages that emerge from these results for the analyzed subgroups of children and across the various areas of development measured.

In terms of program or classroom features, there are no differences by curriculum, with results shown for HighScope in relation to the omitted group being classrooms implementing Creative Curriculum. No association was found between classroom size and children's performance. There are some positive associations between teachers' who identify as people of color[24] and children's vocabulary and literacy gains. No associations are observed between lead teacher qualifications and children's outcomes, or between the ECERS-3 measure of quality and the different measures of child progress. Positive associations were found between CLASS CO scores and Math standard and raw gains (see Appendix Table C.2 and C.3). Results are quite consistent in estimations with and without family child care providers (Tables 12 and 13). FCCs (Table 15) on average show smaller gains in their children's executive function levels (as measured by the DCCS), although in estimations including FCCs CLASS CO is positively associated with changes in DCCS.

Table 13. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and overall ECERS-3, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.271	**3.435*****	0.679	**-0.129***	**-1.432***
	(1.12)	**(1.01)**	(1.10)	**(0.06)**	**(0.65)**
Returning Status	-2.232	1.283	-0.664	-0.018	-0.060
	(1.37)	(1.33)	(1.45)	(0.08)	(0.82)
Asian	**-2.923***	-0.682	0.501	-0.037	**1.693***
	(1.37)	(1.24)	(1.37)	(0.07)	**(0.77)**
Black	-0.852	-0.222	-1.214	-0.117	-0.544
	(1.40)	(1.29)	(1.44)	(0.07)	(0.81)
Hispanic	-0.043	-1.167	1.565	0.044	1.014
	(1.52)	(1.43)	(1.59)	(0.08)	(0.89)
Other	-2.426	0.594	1.177	0.002	**2.448****
	(1.50)	(1.36)	(1.51)	(0.08)	**(0.85)**
DLL	-0.467	0.831	0.926	0.030	-0.588
	(1.15)	(1.01)	(1.11)	(0.06)	(0.63)
Agency Selected	**2.360***	-0.006	0.270	0.007	-0.211

[21] Last year Blacks and Hispanics evidenced lower gains in receptive vocabulary and children categorized as Other evidenced lower literacy scores (Nores, et. al, 2017).

[22] Intersectional estimations for gender and race (not shown) found that the negative receptive effect observed for children identified as Asian is driven by males and the positive executive functions effect is driven by females.

[23] This may be due to accumulated benefits of programming, or to selection biases when programs "select" children.

[24] Self identification as having an ethnic or minority background.

Variables					
	(1.08)	(1.02)	(1.08)	(0.06)	(0.61)
HH Income<20k	-1.805	-3.456	2.420	-0.157	-0.736
	(2.64)	(2.55)	(2.80)	(0.15)	(1.59)
HH Income 21-40k	-1.785	-1.462	-1.196	-0.188	-1.369
	(2.07)	(1.88)	(2.09)	(0.11)	(1.18)
HH Income 41-60k	-0.381	-1.348	0.223	-0.098	-0.666
	(1.99)	(1.85)	(2.04)	(0.11)	(1.15)
HH Income 61-80k	0.551	-2.166	0.132	-0.060	-0.075
	(1.99)	(1.81)	(1.99)	(0.11)	(1.13)
FPL < 100	0.289	2.275	-4.232	0.097	0.994
	(2.62)	(2.57)	(2.80)	(0.15)	(1.58)
FPL 100 to 300	-0.563	-1.280	-1.666	0.015	-0.208
	(1.83)	(1.68)	(1.84)	(0.10)	(1.04)
High Scope	0.518	0.094	-0.281	-0.042	0.087
	(1.09)	(1.04)	(1.07)	(0.06)	(0.61)
Class Size	0.396	-0.020	-0.370	-0.001	-0.080
	(0.27)	(0.25)	(0.26)	(0.01)	(0.15)
Teacher Qual. Exceeds	2.697	2.173	1.106	0.010	-0.127
	(1.52)	(1.51)	(1.55)	(0.08)	(0.88)
Teacher Qual. Meets	2.471	-0.419	0.634	0.035	0.907
	(1.38)	(1.30)	(1.35)	(0.07)	(0.76)
Teacher Black	**4.869****	1.747	-1.194	-0.066	-0.504
	(1.66)	(1.65)	(1.69)	(0.09)	(0.96)
Teacher Hispanic	1.932	-0.536	-1.147	-0.130	-1.015
	(1.54)	(1.51)	(1.54)	(0.08)	(0.87)
Teacher Asian	**4.527***	2.154	-0.094	-0.082	-1.784
	(1.87)	(1.80)	(1.86)	(0.10)	(1.05)
Teacher Other	2.000	**3.312***	0.843	-0.002	-0.580
	(1.30)	**(1.30)**	(1.33)	(0.07)	(0.75)
ECERS-3	0.451	-0.237	0.492	-0.052	-0.694
	(0.76)	(0.74)	(0.77)	(0.04)	(0.43)
N	702	573	573	571	574

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualifications and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table 14. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and CLASS dimensions, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.255	**3.217****	0.555	**-0.125***	**-1.349***
	(1.12)	**(1.01)**	(1.11)	**(0.06)**	**(0.65)**
Returning Status	-2.200	1.236	-0.481	-0.007	0.045
	(1.38)	(1.32)	(1.45)	(0.08)	(0.83)
Asian	**-2.867***	-0.819	0.553	-0.025	**1.817***
	(1.37)	(1.24)	(1.37)	(0.07)	**(0.77)**
Black	-0.830	-0.135	-1.030	-0.106	-0.477
	(1.41)	(1.29)	(1.44)	(0.07)	(0.81)
Hispanic	-0.053	-1.304	1.362	0.040	1.004
	(1.52)	(1.43)	(1.58)	(0.08)	(0.89)
Other	-2.372	0.650	1.380	0.007	**2.475****
	(1.50)	(1.36)	(1.50)	(0.08)	**(0.85)**

Variables					
DLL	-0.466	0.904	0.977	0.030	-0.618
	(1.15)	(1.00)	(1.11)	(0.06)	(0.63)
Agency Selected	**2.433***	0.118	0.425	-0.016	-0.533
	(1.06)	(0.98)	(1.06)	(0.06)	(0.60)
HH Income<20k	-1.860	-3.203	2.717	-0.137	-0.584
	(2.64)	(2.54)	(2.79)	(0.15)	(1.59)
HH Income 21-40k	-1.809	-1.201	-1.026	-0.184	-1.398
	(2.07)	(1.88)	(2.08)	(0.11)	(1.18)
HH Income 41-60k	-0.410	-0.997	0.371	-0.094	-0.680
	(2.00)	(1.85)	(2.03)	(0.11)	(1.15)
HH Income 61-80k	0.535	-1.946	0.279	-0.058	-0.099
	(1.99)	(1.81)	(1.99)	(0.10)	(1.13)
HH Income Missing	4.936	3.699	4.288	-0.056	0.317
	(4.18)	(3.77)	(4.14)	(0.22)	(2.36)
FPL < 100	0.385	2.036	-4.456	0.068	0.745
	(2.61)	(2.56)	(2.78)	(0.15)	(1.58)
FPL 100 to 300	-0.511	-1.412	-1.669	0.005	-0.296
	(1.83)	(1.67)	(1.83)	(0.10)	(1.04)
High Scope	0.396	-0.480	-1.000	-0.042	0.322
	(1.11)	(1.03)	(1.09)	(0.06)	(0.62)
Class Size	0.425	-0.011	-0.284	0.000	-0.082
	(0.27)	(0.25)	(0.26)	(0.01)	(0.15)
Teacher Qual. Exceeds	2.777	1.751	0.670	-0.024	-0.461
	(1.48)	(1.42)	(1.52)	(0.08)	(0.87)
Teacher Qual. Meets	2.632	-1.208	-0.128	-0.029	0.402
	(1.34)	(1.23)	(1.32)	(0.07)	(0.75)
Teacher Black	**4.943****	2.267	-0.821	-0.090	-0.873
	(1.67)	(1.59)	(1.69)	(0.09)	(0.96)
Teacher Hispanic	1.889	-0.117	-0.924	-0.130	-1.098
	(1.56)	(1.45)	(1.55)	(0.08)	(0.88)
Teacher Asian	**4.493***	2.930	0.428	-0.085	-2.015
	(1.91)	(1.77)	(1.89)	(0.10)	(1.07)
Teacher Other	1.976	**2.711***	0.413	-0.002	-0.442
	(1.33)	**(1.26)**	(1.35)	(0.07)	(0.77)
CLASS ES average	0.322	-2.054	-1.610	-0.100	-0.353
	(1.25)	(1.20)	(1.28)	(0.07)	(0.73)
CLASS CO average	0.146	1.029	**2.063***	0.103	0.516
	(1.00)	(0.95)	**(1.01)**	(0.05)	(0.57)
CLASS IS average	-0.043	0.930	0.213	-0.041	-0.521
	(0.52)	(0.48)	(0.51)	(0.03)	(0.29)
N	702	573	573	571	574

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualifications and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table 15. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and CLASS dimensions, including FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.476	**2.520****	0.560	**-0.128***	**-1.671****
	(1.08)	**(0.97)**	(1.05)	**(0.06)**	**(0.62)**
Returning Status	-2.450	0.835	-0.514	0.003	-0.043

	(1.37)	(1.32)	(1.43)	(0.08)	(0.81)
Asian	**-2.691***	-0.117	0.857	-0.041	**1.788***
	(1.36)	(1.23)	(1.34)	(0.07)	**(0.75)**
Black	-0.848	0.273	-0.640	-0.138~	-0.360
	(1.38)	(1.26)	(1.39)	(0.07)	(0.78)
Hispanic	-0.585	-0.894	1.726	0.014	0.955
	(1.49)	(1.40)	(1.53)	(0.08)	(0.86)
Other	-2.380	1.219	1.714	-0.005	**2.442****
	(1.48)	(1.34)	(1.47)	(0.08)	**(0.83)**
DLL	-0.606	0.622	0.529	0.046	-0.595
	(1.11)	(0.97)	(1.06)	(0.06)	(0.60)
Agency Selected	**2.238***	-0.077	0.415	-0.023	-0.576
	(1.06)	(1.00)	(1.05)	(0.06)	(0.60)
HH Income<20k	-1.537	-2.598	1.594	-0.194	-0.980
	(2.58)	(2.46)	(2.67)	(0.14)	(1.52)
HH Income 21-40k	-1.297	-1.164	-0.526	-0.188	-1.342
	(2.05)	(1.87)	(2.05)	(0.11)	(1.16)
HH Income 41-60k	-0.438	-1.163	0.409	-0.102	-0.531
	(1.98)	(1.84)	(2.01)	(0.11)	(1.14)
HH Income 61-80k	0.356	-1.806	0.300	-0.079	-0.085
	(1.99)	(1.81)	(1.97)	(0.11)	(1.12)
FPL < 100	-0.622	1.107	-3.517	0.101	1.027
	(2.55)	(2.47)	(2.66)	(0.14)	(1.51)
FPL 100 to 300	-0.687	-1.454	-1.841	-0.006	-0.309
	(1.82)	(1.68)	(1.82)	(0.10)	(1.03)
FCC	0.429	-5.721	-1.494	**-0.349***	-2.276
	(3.38)	(3.06)	(3.18)	**(0.17)**	(1.81)
High Scope	-0.463	-0.720	-1.250	-0.043	0.283
	(1.06)	(1.00)	(1.02)	(0.05)	(0.58)
Class Size	0.389	-0.111	-0.230	-0.005	-0.108
	(0.25)	(0.24)	(0.25)	(0.01)	(0.14)
Teacher Qual. Exceeds	2.290	0.941	0.699	-0.060	-0.629
	(1.47)	(1.45)	(1.49)	(0.08)	(0.85)
Teacher Qual. Meets	2.474	-1.315	-0.218	-0.035	0.291
	(1.34)	(1.27)	(1.31)	(0.07)	(0.74)
Teacher Black	**3.432***	0.773	-0.851	-0.150	-1.138
	(1.59)	(1.54)	(1.58)	(0.08)	(0.90)
Teacher Hispanic	1.268	-0.752	-0.921	**-0.160***	-1.183
	(1.54)	(1.49)	(1.52)	**(0.08)**	(0.86)
Teacher Asian	3.615	1.863	0.717	-0.137	**-2.103***
	(1.87)	(1.78)	(1.83)	(0.10)	**(1.04)**
Teacher Other	1.958	**2.646***	0.371	-0.010	-0.534
	(1.32)	**(1.31)**	(1.34)	(0.07)	(0.76)
CLASS ES average	0.257	-1.583	-1.697	-0.076	-0.447
	(1.22)	(1.20)	(1.23)	(0.07)	(0.70)
CLASS CO average	0.362	1.077	**2.202***	**0.112***	0.662
	(0.98)	(0.96)	**(0.98)**	**(0.05)**	(0.56)
CLASS IS average	-0.175	0.824	0.221	-0.045	-0.530
	(0.51)	(0.49)	(0.51)	(0.03)	(0.29)
N	735	606	606	604	607

* $p<0.05$; ** $p<0.01$; *** $p<0.001$. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualifications and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Sensitivity Analyses

We also conducted three types of sensitivity checks to assess the robustness of findings. First, we repeated the analyses with raw scores because imperfections in the standardization could affect results. Second, we investigated whether a quality threshold made a difference, third, we replicated the analyses with fixed effects for agencies, which can be interpreted as understanding differences within agencies.

The results of the three types of sensitivity analyses are summarized as follows.

(1) Results of analyses on raw scores for the PPVT, LW and AP measures (Tables C.1 using ECERS and C.2 and C.3. using CLASS) are consistent with the standard score analyses.

(2) Analyses investigating thresholds of quality are reported in Appendix Tables C.4 for ECERS and C.5 for CLASS.[25] We find no association between the ECERS-3 threshold above 3 and children's standard score gains (or raw score gains, either, although these are not reported).[26] We also find no associations between the CLASS thresholds and children's outcomes.

(3) Analyses with agency fixed effects (Tables C.6 and C.7) revealed that on average some few agencies under or over performed in specific few areas of development (not shown), while the majority seem to have no specific effects on children. That is, for the most part, children attending most agencies did not perform any different than children attending other agencies. However, within agencies, ECERS scores were actually negatively associated with the DCCS measure. On the other hand, CLASS IS differences showed a statistically significant positive association with letter-word identification changes in children.

Summary

The evaluation finds that SPP quality has continued to improve on two separate measures, the ECERS-3 and the CLASS. SPP quality as measured by the ECERS-3 and CLASS now exceeds that in some other major city and state pre-k and/or childcare systems. Average quality does not differ significantly between classrooms and family child care providers, the latter having been added to SPP this year as a pilot. Average quality as measured by the ECERS-3 and the CLASS instructional support does not significantly differ by race and ethnicity. Modest differences in the CLASS classroom organization and emotional supports were observed for race and ethnicity, albeit high for all children regardless. Children in SPP made gains in all measured domains with gains in language, literacy and mathematics larger than expected based on maturation. High CLASS classroom organization was associated with strong gains in math for children in the program. African-American and Asian teachers' students had larger gains in vocabulary, pointing to the importance of teacher diversity in SPP. We recommend that the Seattle Preschool Program builds on its success by focusing further improvement efforts in the

---

[25] Burchinal et al. (2010) found evidence of CLASS IS thresholds at 3.25, and CLASS ES in the 5-7 range, and Hatfield et al. (2016) found evidence of CLASS IS threshold at 3 and CLASS ES and CO at 6. Given the distributions of quality in the sample, we chose to use a level of 3 for the ECERS and levels of 5.5 for CLASS emotional support and classroom organization scales, and a level of 3 for CLASS instructional supports.
[26] We also tested the higher level of 5 considered high in the instrument and did not find positive associations either. These are not reported.

quality of instruction with particular attention to language and literacy, integration of content across domains in children's activities, and supports for sustained, reflective thinking as well as personal care routines that contribute to health.

References

Aikens, N., Klein, A. K., Tarullo, L., & West, J. (2013). Getting ready for kindergarten: Children's progress during Head Start. FACES 2009 Report. (OPRE Report 2013–21a) Office of Planning. Research and Evaluation, Administration for Children and Families. Washington, D.C.: US Department of Health and Human Services.

Barnett, W.S. (2013). Expanding Access to Quality Pre-K is Sound Public Policy. New Brunswick, NJ: National Institute for Early Education Research.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child development, 78(2), 647-663.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. Early Childhood Research Quarterly, 25(2), 166-176.

Childcare Quality & Early Learning Center for Research & Professional Development (Unpublished). Early Achievers Standards Validation Study. Seattle: University of Washington

Childcare Quality & Early Learning Center for Research & Professional Development (Unpublished). Large Scale Psychometric Assessment of ECERS 3. Seattle: University of Washington

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". Developmental psychobiology, 29(4), 315-334.

Dunn, L. M., & Dunn, D. M. (2007). PPVT-4: Peabody picture vocabulary test. Pearson Assessments.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D. Cai, K., Clifford, R.M., Ebanks, C., Griffin, J.A & Henry, G. T. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. Child development, 78(2), 558-580.

Early, D. M., Sideris, J., Neitzel, J., LaForett, D. R., & Nehler, C. G. (2018). Factor structure and validity of the Early Childhood Environment Rating Scale–Third Edition (ECERS-3). Early Childhood Research Quarterly, *44*, 242-256.

Edvance Research (2014). Pre-K 4 SA Evaluation Report. YEAR 1. Final Report Submitted to Early Childhood Education Municipal Development Corporation. San Antonio, TX: Author.

Edvance Research (2015). Pre-K 4 SA Evaluation Report. YEAR 2. Final Report Submitted to Early Childhood Education Municipal Development Corporation. San Antonio, TX: Author.

Edvance Research (2016). Pre-K 4 SA Evaluation Report. YEAR 3. Final Report Submitted to Early Childhood Education Municipal Development Corporation. San Antonio, TX: Author.

Harms, T., Clifford, R. M., & Cryer, D. (2014). Early childhood environment rating scale. Teachers College Press.

Hatfield, B. E., Burchinal, M. R., Pianta, R. C., & Sideris, J. (2016). Thresholds in the association between quality of teacher–child interactions and preschool children's school readiness skills. Early Childhood Research Quarterly, 36, 561-571.

Jenson, D. (2015) ECERS-3: One year out. Four States' Experiences with Planning and
        Implementing Use of ECERS-3. Presented at the QRIS National Meeting 2015. Available
        at http://www.qrisnetwork.org/sites/all/files/conference-
        session/resources/703ECERS3_0.pdf

Joseph, G. E., Feldman, E., Phillips, J. J., & Jackson, E. (2010). The combined CLASS:
        Assessing the adult-child interactions in mixed age family childcare. A procedure
        manual. Designed for Washington State's QRIS, Early Achievers. Seattle, WA: Cultivate
        Learning.

Lamy, C.E., Frede, E. Seplocha, H., Strasser, J., Jambunathan, S., Juncker, J. A., Ferrar, H.
        Wiley, L., & Wolock, E. (2004). Inch by Inch, Row by Row Gonna Make This Garden
        Grow. Classroom Quality and Language Skills in the Abbott Preschool Program. Year
        One Report, 2002-2003 Early Learning Improvement Consortium. New Jersey: Early
        Learning Improvement Consortium. Available at
        http://www.state.nj.us/education/ece/research/inch.pdf

Meador, D. N., Turner, K. A., Lipsey, M. W., & Farran, D. C. (2013). Administering Measures
        from the PRI Learning-Related Cognitive Self- Regulation Study. Nashville, TN: Peabody
        Research Institute. Available at
        https://my.vanderbilt.edu/cogselfregulation/files/2012/11/SR-Measure-Training-Manual-
        final.pdf

Nores, M., Barnett, W.S., Joseph, G., Stull, S., Jung, K. & Soderberg, J.S. (2017). Year 2 report:
        Seattle Pre-k program evaluation. New Brunswick, NJ: National Institute for Early
        Education Research & Seattle, WA: Cultivate Learning.

NIEER (2014). New Jersey Abbott Preschool Quality Evaluation Study. Summary Report. New
        Brunswick, NJ: National Institute for Early Education Research.

NIEER (2016). New Jersey Abbott Preschool Quality Evaluation Study. Summary Report. New
        Brunswick, NJ: National Institute for Early Education Research.

NYC Department of Education (2017). Pre-K Program Assessments Early Childhood
        Environmental Rating Scale –Revised (ECERS-R) and Classroom Assessment Scoring
        System (CLASS) Release. New York: Author. Available at
        http://schools.nyc.gov/NR/rdonlyres/5FEA3D5B-E615-4E16-83A8-
        C4E58A4D6F02/0/201516ProgramAssessmentResultsSummary.pdf

NYC Department of Education. (2015). Pre-K Program Assessments Classroom Assessment
        Scoring System (CLASS) and Early Childhood Environmental Rating Scale – Revised
        (ECERS-R) Release. New York: Author. Available at
        http://schools.nyc.gov/NR/rdonlyres/A8A27BFE-7C58-4F03-8EB7-
        B90E01BA3D0D/0/CLASSandECERSRReleaseDeckFinal.pdf

NYC Department of Education. (December 18, 2015). Mayor de Blasio Announces Over 68,500
        Students Enrolled in Pre-K for All. Press office. New York: Author. Available at
        http://www1.nyc.gov/office-of-the-mayor/news/954-15/mayor-de-blasio-over-68-500-
        students-enrolled-pre-k-all

Office of Head Start. U.S. A National Overview of Grantee CLASS® Scores in 2015.
        Washington, D.C.: Department of Health and Human Services. Available at
        http://eclkc.ohs.acf.hhs.gov/hslc/data/class-reports/docs/national-class-2015-data.pdf

PAKEYS (Unpublished). What does the data tell us? The evolution of environment rating scale
        (ERS) use within QRIS. Available at https://qrisnetwork.org/sites/all/files/conference-
        session/resources/651DataTellsUs.pdf.

Phillips, D. A., Gormley, W. T., & Lowenstein, A. E. (2009). Inside the pre-kindergarten door: Classroom climate and instructional time allocation in Tulsa's pre-K programs. Early Childhood Research Quarterly, 24(3), 213-228.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System: Manual Pre-K. Education Review//Reseñas Educativas.

Qi, C. H., Kaiser, A. P., Milan, S., & Hancock, T. (2006). Language performance of low-income African American and European American preschool children on the PPVT–III. Language, Speech, and Hearing Services in Schools, 37(1), 5-16.

Rivers, N. M. (2016). Seattle Public Schools and Housing Report. Seattle: Seattle Public Schools. Available at https://www.seattleschools.org/cms/one.aspx?portalId=627&pageId=15652.

Tout, K., Magnuson, K. Lipscomb, S., Karoly, L, Starr, R., Quick H., Early, D. Epstein, D., Joseph, G., Maxwell, K, Roberts, J., Swanson, C. & Wenner, J. (2017). Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies. OPRE Report #2017-92. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. Early Childhood Research Quarterly, 28(2), 199-209.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. Journal of policy Analysis and management, 27(1), 122-154.

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. (2001). Woodcock-Johnson III NU tests of achievement. Rolling Meadows, IL: Riverside Publishing.

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. Nature Protocols, 1, 297-301.

Appendices

Appendix A. ECERS-3 and CLASS, additional details.
Appendix B. Child Scores, pre, post and gains.
Appendix C. Sensitivity Analyses.
Appendix D. P-values for tests of differences in means.

Appendix A. ECERS-3 and CLASS, additional details.

Table A.1. ECERS-3 Subscale and Item Descriptions.

Subscale	Items	Description
Space for Furnishings	1. Indoor Space	Considers enough indoor space for children, staff, and basic furnishings for routines, play, and learning.
	2. Furnishings for care, play, and learning	Focuses on ample furniture for routine care, play, and learning, including convenient cubbies for individual use.
	3. Room arrangement for play and learning	Space is arranged so that classroom pathways generally do not interrupt play and supervision.
	4. Space for privacy	Considers an indoor space for privacy available and set up physically in the classroom to discourage interruptions.
	5. Child-related display	Focuses on appropriate materials displayed for children throughout the classroom, including simple pictures, posters, and artwork.
	6. Space for gross motor play	Gross motor area is spacious, generally safe, and easily accessible to children.
	7. Gross motor equipment	Equipment is age appropriate, accessible, and ample enough to interest every child.
Personal Care Routines	Meals/Snacks	Schedule and sanitary procedures are appropriate during meal times. Staff sit with children to encourage learning.
	Toileting/diapering	Proper sanitary procedures usually followed with pleasant supervision.
	Health practices	Proper sanitary procedures used consistently as needed, with a few lapses.
	Safety practices	Considers no more than 2 major safety hazards present indoors or outdoors.
Language and Literacy	Helping children expand vocabulary	Measures how frequent staff uses specific words for objects and actions and descriptive words as children experience routines and play.
	Encouraging children to use language	Assesses how frequent staff asks questions that children are interested in answering and that require longer answers. Includes many conversations during gross motor free play and routines.
	Staff use of books with children	Staff read appropriate books to children that relate to current classroom activities or themes, showing interest and enjoyment while doing so.
	Encouraging children's use of books	Many books are accessible and organized in a defined interest center.
	Becoming familiar with print	Focuses on how most visible print is combined with pictures, relates to current classroom topics, and shows a variety of words.
Learning Activities	Fine motor	Focuses on the accessibility for children of fine motor materials, including interlocking building materials, manipulatives, puzzles, and art materials.
	Art	Art materials, including drawing materials, paints, 3D objects, collage materials, and tools, must be accessible for children.
	Music and movement	Measures how many music materials and activities are accessible for children during free play.
	Blocks	Enough space, unit blocks and accessories from 3 different categories for 2-3 children to build at once.
	Dramatic play	Many and varied dramatic play materials, including dolls, furniture, play food and dress-up clothes must be accessible for children during free play.

	Nature/science	At least 15 nature/science materials, including living things, natural objects, factual books, tools, or sand/water must be accessible for children.
	Math materials and activities	At least 10 different appropriate math materials accessible, including materials to count/compare quantities, measure/compare sizes, and familiarize children with shapes.
	Math in daily events	Assess how staff encourages math learning as part of daily routines.
	Understanding written numbers	At least 3-5 different materials should be present in the classroom that shows children the meaning of print numbers.
	Promoting acceptance of diversity	At least 10 examples of diversity accessible, including books, displayed pictures and materials.
	Appropriate use of technology	All observed materials used are appropriate and limited to 10-15 minutes per child during the observation.
Interaction	Supervision of gross motor	Focuses on careful supervision in order to ensure children's safety.
	Individualized teaching and learning	Many activities observed are open- ended and most allow children to be successful.
	Staff-child interaction	Evaluates frequent positive staff- child interactions, with no long periods of no interaction.
	Peer interaction	Captures positive peer interactions during at least half of the observation.
	Discipline	Children appear to be aware of classroom rules, and generally follow them with reasonable amount of teacher control.
Program Structure	Transitions and waiting times	Classroom transitions are usually smooth and productively engaging.
	Free play	Free play takes place for 1 hour during observation, including some time indoors and some time outdoors (weather permitting).
	Whole - group activities for play and learning	Staff are responsive and flexible in ways that maximize child engagement during whole group activities.

Table A.2. CLASS Domains and Dimension Descriptions.

Domain	Dimension	Description
Emotional Support	Positive Climate	Reflects the emotional connection between teachers and children and among children, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.
	Negative Climate	Reflects the overall level of expressed negativity in the classroom. The frequency, quality, and intensity of teacher and peer negativity are key to this dimension
	Teacher Sensitivity	Encompasses the teacher's awareness of and responsiveness to students' academic and emotional needs.
	Regard for Student Perspectives	Captures the degree to which the classroom activities and teacher's interactions with students place an emphasis on students' interests, motivations, and points of view and encourage student responsibility and autonomy.
Classroom Organization	Behavior Management	Encompasses the teacher's ability to provide clear behavior expectations and use effective methods to prevent and redirect misbehavior.
	Productivity	Considers how well the teacher manages instructional time and routines and provides activities for students so that they have the opportunity to be involved in learning activities.
	Instructional Learning Formats	Focuses on the ways in which teachers maximize students' interest, engagement, and abilities to learn from lessons and activities.

Instructional Support	Concept Development	Measures the teacher's use of instructional discussions and activities to promote students' higher-order thinking skills and cognition and the teacher's focus on understanding rather than on rote instruction.
	Quality of Feedback	Assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation.
	Language Modeling	Captures the effectiveness and amount of teacher's use of language-stimulation and language-facilitation techniques.

Table A.3. Considerations on The Combined CLASS protocol

The protocol for using Combined CLASS manuals (Joseph, Feldman, Phillips & Jackson, 2010) integrates dimensions from all three CLASS tools (Infant, Toddler, and Pre-K) to allow for multi-age groupings, most often present in family child care homes. Each of the individual CLASS protocols contain differing numbers of dimensions i.e., Infant has 4, Toddler has 8, and Pre-K has 10. Therefore, some dimensions in the Combined CLASS process apply only to certain age groups. For example, three dimensions apply only to preschool children, four dimensions apply only to toddlers and preschoolers, and the remaining four dimensions apply to all age groups. When coding dimensions that span all age groups, consideration is given to how many children are present within an age group and the relative breadth/depth of interactions that impact each. For example, imagine half the attendees are infants and half are preschoolers. In such a scenario, if the caregiver provides appropriate language stimulation to infants but only provides low level language modeling for Pre-K children, the score on this dimension may fall in the mid-range when using the Combined Class process even though it may fall in the low range for Pre-K CLASS children. Only children present are counted and infants sleeping are not considered "present". Please note this process is a hybrid model designed for Washington State's QRIS and utilized in this study. For information about other Family Child Care CLASS models, please see "Using the CLASS Measure in Family Child Care Homes" (Vitiello, 2014) via Teachstone.com

Table A.3. ECERS and CLASS Dimension and Domain Means by Child Demographics, 2018

		N	ECERS		N	CLASS_ES		CLASS_CO		CLASS_IS	
			Mean	SD		Mean	SD	Mean	SD	Mean	SD
Total		859	4.00	0.62	910	6.35	0.59	5.91	0.78	3.41	1.04
Gender	Female	409	3.99	0.63	435	6.34	0.60	5.93	0.77	3.36	0.98
	Male	450	4.01	0.61	475	6.35	0.58	5.89	0.79	3.46	1.09
Ethnicity	White	181	4.05	0.58	192	6.45	0.41	6.12	0.60	3.56	0.97
	Black	228	3.99	0.66	257	6.26	0.69	5.75	0.87	3.29	0.97
	Asian	240	3.95	0.63	242	6.33	0.62	5.88	0.81	3.45	1.17
	Hispanic	109	4.09	0.54	115	6.45	0.51	6.08	0.66	3.49	1.07
	Other	89	3.90	0.62	92	6.24	0.55	5.75	0.81	3.21	0.88
Language	English	479	4.01	0.62	520	6.35	0.54	5.91	0.79	3.38	0.94
	Bilingual	259	4.02	0.57	268	6.35	0.54	5.89	0.77	3.33	0.99
	Unknown	121	3.89	0.70	122	6.33	0.83	5.98	0.80	3.73	1.41
FPL	<100	271	4.05	0.62	289	6.28	0.67	5.85	0.79	3.36	0.97
	100-300	401	4.01	0.60	427	6.36	0.56	5.89	0.80	3.38	1.04
	>300	181	3.89	0.65	188	6.42	0.49	6.03	0.73	3.56	1.13

Appendix B. Child Scores, pre, post and gains.

Receptive vocabulary results

Table B.1. reports children's receptive vocabulary scores for the fall (pre-test) and spring (post-test) and fall to spring gains. Standardized scores—which are adjusted for age—are reported in this section (raw scores are reported in section further below). The mean standard score for this measure is set at 100 which represents the average child in the U.S. population at any age. The standard deviation is 15. Thus, positive gains are an indication that children improved more over the course of the preschool year than is expected based on the change in age alone. Only valid scores for children assessed in both fall and spring of the school year are included.

Table B.1. Receptive vocabulary means and gains by child characteristics

| | | PPVT 2017 Fall | | | PPVT 2018 Spring | | PPVT Gains 2017–18 | |
		N	Mean	SD	Mean	SD	Mean	SD
Total		735	96.17	19.27	99.33	18.31	3.16	11.66
Gender	Male	371	95.02	19.36	99.04	18.13	4.02	11.46
	Female	364	97.35	19.13	99.63	18.52	2.28	11.80
Age	3-Year-Old Cohort	185	92.70	16.69	95.63	15.55	2.93	10.83
	4-Year-Old Cohort	550	97.34	19.95	100.58	19.01	3.24	11.93
Ethnicity	White	161	108.79	17.69	111.15	15.08	2.36	11.89
	Black	190	89.07	15.17	93.88	16.52	4.81	12.91
	Asian	203	89.67	18.98	93.08	17.43	3.41	10.24
	Hispanic	92	95.63	18.49	99.28	18.54	3.65	9.65
	Other	81	104.17	16.13	104.05	17.55	-0.12	13.07
Language	English	432	102.80	17.36	104.51	17.40	1.71	12.45
	DLL	200	83.88	15.70	89.09	14.84	5.21	9.83
	Unknown	103	92.26	20.51	97.53	19.54	5.27	10.59
FPL	<100	228	89.24	17.58	93.29	17.23	4.06	12.31
	100-300	334	95.34	18.74	98.50	17.94	3.16	11.82
	>300	170	107.26	17.68	109.17	16.51	1.91	10.41

Children's pre-test and post-test vocabulary standard scores for selected center characteristics are reported in Table B.2. (raw scores are reported further below).

Table B.2. Receptive vocabulary means and gains by center characteristics

| | | PPVT 2017 Fall | | | PPVT 2018 Spring | | PPVT Gains 2017–18 | |
		N	Mean	SD	Mean	SD	Mean	SD
Total		735	96.17	19.27	99.33	18.31	3.16	11.66
Curriculum	High Scope	464	96.85	19.63	99.97	18.51	3.12	11.28
	Creative Curriculum	271	95.01	18.62	98.25	17.96	3.24	12.29
ECERS	Less than 3	72	93.60	17.14	96.29	16.68	2.69	10.32
	3 or More	630	96.72	19.57	99.97	18.46	3.24	11.87
CLASS ES	Less than 5.5	38	88.87	15.85	92.53	16.00	3.66	8.40
	5.5 or More	697	96.57	19.37	99.71	18.37	3.13	11.81
CLASS CO	Less than 5.5	158	94.39	16.05	97.49	15.40	3.10	10.12
	5.5 or More	577	96.66	20.05	99.84	19.02	3.18	12.05
CLASS IS	Less than 3	270	93.88	17.37	97.53	17.62	3.65	12.33
	3 or More	465	97.50	20.19	100.38	18.65	2.88	11.25

Literacy results

Children's WJ-III letter-word (LW) identification scores for the overall sample and by selected child characteristics are reported in Table B.3. The LW subtest measures children's ability to identify letters and subsequently read a list of words of increasing difficulty. The test also has a mean standard (i.e., age adjusted score) of 100 and a standard deviation of 15 (raw scores are reported further below).

Table B.3. Literacy means and gains by child characteristics

| | | WJ-LW 2017 Fall | | | WJ-LW 2018 Spring | | WJ-LW Gains 2017–18 | |
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	101.20	15.32	102.68	14.92	1.48	9.77
Gender	Male	298	101.43	16.04	102.47	16.12	1.04	9.85
	Female	308	100.97	14.60	102.87	13.69	1.90	9.70
Age	3-Year-Old Cohort	167	102.40	15.29	105.31	14.78	2.91	12.34
	4-Year-Old Cohort	439	100.74	15.32	101.68	14.87	0.94	8.55
Ethnicity	White	129	103.09	13.87	103.97	12.99	0.88	7.94
	Black	152	100.16	16.04	102.07	14.71	1.91	9.44
	Asian	178	102.52	17.33	103.98	17.27	1.47	11.49
	Hispanic	70	97.56	13.37	98.91	14.36	1.36	8.65
	Other	71	99.94	12.06	102.39	12.26	2.45	9.68
Language	English	357	102.31	15.07	103.03	15.11	0.72	8.47
	DLL	171	100.61	15.66	102.56	14.85	1.95	11.56
	Unknown	78	97.41	15.17	101.35	14.30	3.94	10.67
FPL	<100	176	97.32	15.65	100.28	14.91	2.96	11.48
	100-300	288	101.16	14.05	101.67	13.70	0.51	8.86
	>300	140	106.10	16.16	107.93	16.12	1.83	8.72

Table B.4. reports SPP children's pre- and post-test letter-word identification standard scores across selected center characteristics (raw scores are reported further below).

Table B.4. Literacy means and gains by center characteristics

| | | WJ-LW 2017 Fall | | | WJ-LW 2018 Spring | | WJ-LW Gains 2017–18 | |
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	101.20	15.32	102.68	14.92	1.48	9.77
Curriculum	High Scope	382	100.40	14.00	102.29	13.90	1.89	9.83
	Creative Curriculum	224	102.56	17.27	103.34	16.53	0.78	9.65
ECERS	Less than 3	58	100.19	13.74	101.34	13.48	1.16	11.06
	3 or More	515	101.36	15.66	102.96	15.14	1.61	9.67
CLASS ES	Less than 5.5	30	99.10	12.59	101.00	10.85	1.90	9.61
	5.5 or More	576	101.31	15.45	102.76	15.10	1.46	9.79
CLASS CO	Less than 5.5	129	100.39	14.06	100.65	14.00	0.26	9.62
	5.5 or More	477	101.42	15.64	103.22	15.13	1.81	9.79
CLASS IS	Less than 3	227	100.02	14.18	101.45	13.99	1.43	10.24
	3 or More	379	101.90	15.93	103.41	15.42	1.51	9.49

Early math results

Children's pre- and post-test math scores, as measured by the applied problems (AP) subscale of the WJ-III are reported in Table B.5. Like the two measures above, AP is normed with a mean of 100 and a standard deviation of 15.

Table B.5. Math means and gains by child characteristics

		WJ-AP 2017 Fall			WJ-AP 2018 Spring		WJ-AP Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	101.54	14.40	103.60	13.84	2.06	10.77
Gender	Male	298	100.84	14.98	102.72	14.55	1.88	11.45
	Female	308	102.22	13.80	104.46	13.08	2.24	10.08
Age	3-Year-Old Cohort	167	99.90	14.53	103.49	13.70	3.58	11.16
	4-Year-Old Cohort	439	102.16	14.31	103.65	13.91	1.49	10.57
Ethnicity	White	129	110.60	11.21	109.66	10.93	-0.94	9.32
	Black	152	97.22	13.08	99.01	12.71	1.79	9.34
	Asian	178	98.10	15.35	102.07	13.90	3.97	11.21
	Hispanic	70	99.91	12.80	103.91	14.57	4.00	11.48
	Other	71	103.94	12.99	106.04	15.38	2.10	12.97
Language	English	357	104.77	13.04	105.71	13.19	0.94	10.20
	DLL	171	95.85	14.82	100.20	14.15	4.35	10.90
	Unknown	78	99.22	15.22	101.40	14.37	2.18	12.24
FPL	<100	176	97.64	13.79	100.09	14.74	2.45	11.76
	100-300	288	100.38	14.25	102.71	12.28	2.33	9.60
	>300	140	108.81	12.95	109.94	13.80	1.13	11.75

Table B.6. shows children's pre- and post-test standardized math scores and gains by selected center characteristics (raw scores are reported below).

Table B.6. Math means and gains by center characteristics

		WJ-AP 2017 Fall			WJ-AP 2018 Spring		WJ-AP Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	101.54	14.40	103.60	13.84	2.06	10.77
Curriculum	High Scope	382	101.68	14.46	103.68	13.88	2.00	11.43
	Creative Curriculum	224	101.30	14.31	103.47	13.79	2.17	9.55
ECERS	Less than 3	58	97.69	13.87	100.36	11.51	2.67	10.20
	3 or More	515	102.17	14.58	104.22	14.10	2.05	10.96
CLASS ES	Less than 5.5	30	93.87	9.19	97.00	11.76	3.13	7.80
	5.5 or More	576	101.94	14.51	103.95	13.86	2.01	10.90
CLASS CO	Less than 5.5	129	98.51	12.74	99.71	12.83	1.19	11.35
	5.5 or More	477	102.36	14.72	104.66	13.93	2.30	10.60
CLASS IS	Less than 3	227	100.09	12.83	101.97	13.68	1.88	10.68
	3 or More	379	102.41	15.21	104.58	13.86	2.17	10.83

Executive functions

We used two measures of executive functions. The DCCS is an attention shifting test which taps into a child's short-term memory. Table B.7. reports children's pre- and post-test DCCS scores by selected child characteristics. As a reference, the Learning-Related Cognitive Self-Regulation School Readiness Measures for Preschool Children Study (aka the Self-Regulation Measurement Study) (Meador, et. al, 2013) tested alternative measures of executive functions and included the DCCS. The authors found average DCCS scores of 1.42 at 51–53 months and 1.62 at 57–59 months (an average difference of 0.20 between these two ages); ranges which include the average ages at fall and spring testing in this study (53.2 months in the fall and 59.3 in the spring). Table B.8. report children's pre- and post-test DCCS scores by selected center characteristics.

Table B.7. DCCS means and gains by child characteristics

			DCCS 2017 Fall		DCCS 2018 Spring		DCCS Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		604	1.41	0.61	1.63	0.63	0.23	0.59
Gender	Male	298	1.36	0.60	1.59	0.65	0.23	0.62
	Female	306	1.45	0.62	1.67	0.60	0.22	0.57
Age	3-Year-Old Cohort	168	1.09	0.53	1.33	0.57	0.24	0.60
	4-Year-Old Cohort	436	1.53	0.60	1.75	0.61	0.22	0.59
Ethnicity	White	129	1.69	0.58	1.86	0.57	0.17	0.53
	Black	151	1.19	0.56	1.39	0.59	0.20	0.64
	Asian	178	1.37	0.62	1.60	0.63	0.24	0.60
	Hispanic	69	1.30	0.60	1.65	0.56	0.35	0.54
	Other	71	1.58	0.55	1.79	0.65	0.21	0.63
Language	English	357	1.52	0.61	1.71	0.64	0.19	0.59
	DLL	170	1.21	0.61	1.51	0.59	0.31	0.63
	Unknown	77	1.34	0.50	1.56	0.60	0.22	0.53
FPL	<100	175	1.25	0.56	1.53	0.60	0.27	0.62
	100-300	287	1.38	0.62	1.57	0.62	0.19	0.59
	>300	140	1.66	0.58	1.90	0.59	0.24	0.57

Table B.8. DCCS means and gains by center characteristics

			DCCS 2017 Fall		DCCS 2018 Spring		DCCS Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		604	1.41	0.61	1.63	0.63	0.23	0.59
Curriculum	High Scope	380	1.43	0.60	1.66	0.63	0.22	0.58
	Creative Curriculum	224	1.37	0.64	1.60	0.62	0.23	0.63
ECERS	Less than 3	91	1.32	0.58	1.48	0.70	0.16	0.64
	3 or More	513	1.42	0.62	1.66	0.61	0.24	0.59
CLASS ES	Less than 5.5	30	1.27	0.58	1.50	0.73	0.23	0.82
	5.5 or More	574	1.42	0.61	1.64	0.62	0.22	0.58
CLASS CO	Less than 5.5	129	1.36	0.62	1.54	0.67	0.19	0.68
	5.5 or More	475	1.42	0.61	1.66	0.61	0.24	0.57
CLASS IS	Less than 3	226	1.37	0.59	1.59	0.65	0.23	0.62
	3 or More	378	1.43	0.62	1.66	0.61	0.22	0.58

Children were also assessed with the Peg Tapping (PT) measure. PT is a measure of inhibitory control. Table B.9. reports children's pre- and post-test Peg Tapping scores by selected child characteristics. No norms exist for this measure, either. The Self-Regulation Measurement Study (Meador, et. al, 2013) included this measure as well. Authors reported average scores of 6.02 at 51–53 months and 8.80 at 57–59 months, with a difference of 2.78. SPP children advanced similarly throughout the preschool year. Table B.10. reports pre- and post-test Peg-Tapping scores for children in the sample across selected center characteristics.

Table B.9. Peg Tapping means and gains by child characteristics

| | | | PT 2017 Fall | | PT 2018 Spring | | PT Gains 2017–18 |
		N	Mean	SD	Mean	SD	Mean	SD
Total		607	5.50	5.97	8.38	6.52	2.88	6.15
Gender	Male	299	5.74	5.84	8.02	6.04	2.28	5.28
	Female	308	5.27	6.09	8.73	6.95	3.47	6.85
Age	3-Year-Old Cohort	167	1.83	4.00	5.10	5.14	3.27	5.01
	4-Year-Old Cohort	440	6.89	6.00	9.63	6.57	2.74	6.53
Ethnicity	White	130	8.18	5.93	9.42	5.77	1.25	5.23
	Black	152	3.52	5.19	6.07	5.71	2.55	5.37
	Asian	178	4.72	5.86	8.46	5.91	3.74	5.16
	Hispanic	70	5.06	5.78	8.26	5.80	3.20	5.07
	Other	71	7.21	5.85	11.32	9.47	4.11	10.58
Language	English	357	6.41	6.02	9.20	6.79	2.79	6.58
	DLL	171	3.85	5.58	7.16	5.94	3.30	5.31
	Unknown	79	4.94	5.79	7.33	5.96	2.39	5.82
FPL	<100	176	3.51	5.22	7.58	7.77	4.07	7.83
	100-300	289	5.26	5.82	7.78	5.92	2.52	5.24
	>300	140	8.53	6.01	10.68	5.38	2.15	5.29

Table B.10. Peg-Tapping means and gains by center characteristics

| | | | PT 2017 Fall | | PT 2018 Spring | | PT Gains 2017–18 |
		N	Mean	SD	Mean	SD	Mean	SD
Total		607	5.50	5.97	8.38	6.52	2.88	6.15
Curriculum	High Scope	383	5.87	6.07	8.66	6.80	2.79	6.51
	Creative Curriculum	224	4.87	5.74	7.91	6.01	3.04	5.50
ECERS	Less than 3	58	5.22	6.40	8.00	5.85	2.78	5.99
	3 or More	516	5.64	5.95	8.61	6.60	2.97	6.21
CLASS ES	Less than 5.5	30	3.53	5.31	6.40	5.60	2.87	5.26
	5.5 or More	577	5.60	5.98	8.49	6.55	2.89	6.20
CLASS CO	Less than 5.5	129	4.99	5.97	7.61	5.75	2.62	5.42
	5.5 or More	478	5.64	5.96	8.59	6.71	2.96	6.33
CLASS IS	Less than 3	227	5.00	5.92	8.30	7.31	3.31	7.28
	3 or More	380	5.80	5.98	8.43	6.01	2.63	5.35

Raw Scores

Table B.11. Receptive vocabulary raw score means and gains by child characteristics

		PPVT 2017 Fall			PPVT 2018 Spring		PPVT Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		735	66.31	27.31	79.41	26.56	13.10	14.26
Gender	Male	371	65.60	27.55	79.58	26.71	13.99	15.05
	Female	364	67.04	27.08	79.24	26.44	12.20	13.36
Age	3-Year-Old Cohort	185	48.58	20.93	62.17	20.59	13.59	13.20
	4-Year-Old Cohort	550	72.27	26.62	85.21	25.82	12.94	14.60
Ethnicity	White	161	84.43	25.86	96.81	22.94	12.37	15.48
	Black	190	55.50	21.35	71.07	22.74	15.57	14.66
	Asian	203	57.31	25.85	69.93	25.11	12.62	12.80
	Hispanic	92	66.27	26.45	79.75	27.76	13.48	12.30
	Other	81	77.70	23.69	87.56	24.20	9.85	15.87
Language	English	432	75.52	25.09	87.32	24.63	11.80	15.30
	DLL	200	48.70	22.04	63.52	22.84	14.82	11.80
	Unknown	103	61.86	27.76	77.12	26.98	15.25	13.56
FPL	<100	228	56.77	23.73	70.50	24.33	13.73	14.72
	100-300	334	64.19	26.50	77.55	25.81	13.36	14.23
	>300	170	83.58	25.67	95.28	24.09	11.70	13.78

Table B.12. Receptive vocabulary raw score means and gains by center characteristics

		PPVT 2017 Fall			PPVT 2018 Spring		PPVT Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		735	66.31	27.31	79.41	26.56	13.10	14.26
Curriculum	High Scope	464	68.28	27.96	80.99	27.24	12.71	14.32
	Creative Curriculum	271	62.93	25.85	76.71	25.16	13.78	14.15
ECERS	Less than 3	72	63.31	26.51	75.64	26.50	12.33	13.78
	3 or More	630	67.05	27.57	80.30	26.52	13.25	14.34
CLASS ES	Less than 5.5	38	58.26	23.89	71.03	25.26	12.76	11.27
	5.5 or More	697	66.75	27.43	79.87	26.57	13.12	14.41
CLASS CO	Less than 5.5	158	64.11	24.71	77.14	24.31	13.03	13.47
	5.5 or More	577	66.91	27.97	80.03	27.13	13.12	14.48
CLASS IS	Less than 3	270	63.15	25.25	77.13	24.81	13.98	14.45
	3 or More	465	68.14	28.30	80.74	27.46	12.59	14.14

Table B.13. Literacy raw score means and gains by child characteristics

		WJ-LW 2017 Fall			WJ-LW 2018 Spring		WJ-LW Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	8.05	6.13	11.12	7.53	3.06	4.56
Gender	Male	298	8.39	6.58	11.23	7.37	2.84	3.25
	Female	308	7.72	5.65	11.00	7.69	3.28	5.55
Age	3-Year-Old Cohort	167	6.07	4.48	8.75	5.31	2.68	3.48
	4-Year-Old Cohort	439	8.81	6.49	12.02	8.04	3.21	4.91
Ethnicity	White	129	8.76	6.59	11.71	6.85	2.95	2.99
	Black	152	7.56	5.62	10.31	6.25	2.75	3.04
	Asian	178	8.64	7.13	11.71	7.80	3.07	3.88
	Hispanic	70	6.49	4.26	9.30	5.74	2.81	3.18
	Other	71	7.80	4.80	12.11	11.07	4.31	9.65
Language	English	357	8.50	6.62	11.44	8.48	2.94	5.12
	DLL	171	7.60	5.50	10.54	5.81	2.95	3.57
	Unknown	78	6.99	4.81	10.87	6.11	3.88	3.65
FPL	<100	176	6.70	5.40	10.30	8.84	3.60	6.86
	100-300	288	7.68	5.26	10.18	5.66	2.50	2.90
	>300	140	10.54	7.78	14.15	8.32	3.61	3.50

Table B.14. Literacy raw score means and gains by center characteristics

		WJ-LW 2017 Fall			WJ-LW 2018 Spring		WJ-LW Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	8.05	6.13	11.12	7.53	3.06	4.56
Curriculum	High Scope	382	7.89	5.53	11.23	7.39	3.34	5.18
	Creative Curriculum	224	8.33	7.03	10.92	7.77	2.60	3.21
ECERS	Less than 3	58	7.53	4.91	10.52	5.69	2.98	3.64
	3 or More	515	8.16	6.35	11.27	7.77	3.11	4.69
CLASS ES	Less than 5.5	30	7.13	4.18	10.40	4.93	3.27	3.84
	5.5 or More	576	8.10	6.21	11.15	7.64	3.05	4.60
CLASS CO	Less than 5.5	129	7.74	5.37	10.26	5.99	2.52	3.18
	5.5 or More	477	8.14	6.32	11.35	7.88	3.21	4.86
CLASS IS	Less than 3	227	7.52	5.39	10.68	8.11	3.17	6.00
	3 or More	379	8.37	6.51	11.37	7.16	3.00	3.44

Table B.15. Math raw score means and gains by child characteristics

		WJ-AP 2017 Fall			WJ-AP 2018 Spring		WJ-AP Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	10.30	5.22	13.50	6.63	3.19	5.21
Gender	Male	298	10.33	5.44	13.26	5.60	2.93	3.58
	Female	308	10.28	5.01	13.73	7.49	3.45	6.41
Age	3-Year-Old Cohort	167	7.03	4.29	10.70	9.24	3.67	8.21
	4-Year-Old Cohort	439	11.55	5.00	14.56	4.92	3.01	3.44
Ethnicity	White	129	13.57	4.44	15.78	4.11	2.22	3.12
	Black	152	8.44	4.57	11.98	9.61	3.54	8.45
	Asian	178	9.29	5.52	12.78	5.35	3.48	3.43
	Hispanic	70	9.50	4.35	13.17	5.14	3.67	3.81
	Other	71	11.49	4.72	14.75	5.39	3.25	3.92
Language	English	357	11.47	4.86	14.47	7.25	3.00	6.11
	DLL	171	8.13	5.24	11.76	5.28	3.63	3.40
	Unknown	78	9.71	5.22	12.82	5.35	3.12	3.83
FPL	<100	176	8.81	4.79	11.99	5.22	3.18	3.81
	100-300	288	9.66	5.05	13.08	7.64	3.42	6.43
	>300	140	13.54	4.75	16.31	4.96	2.77	3.76

Table B.16. Math raw score means and gains by center characteristics

		WJ-AP 2017 Fall			WJ-AP 2018 Spring		WJ-AP Gains 2017–18	
		N	Mean	SD	Mean	SD	Mean	SD
Total		606	10.30	5.22	13.50	6.63	3.19	5.21
Curriculum	High Scope	382	10.60	5.40	3.04	3.61	13.65	5.19
	Creative Curriculum	224	9.79	4.86	3.45	7.16	13.24	8.55
ECERS	Less than 3	58	9.17	5.32	12.26	5.34	3.09	3.16
	3 or More	515	10.53	5.24	13.76	6.82	3.23	5.49
CLASS ES	Less than 5.5	30	7.73	3.79	11.20	5.36	3.47	3.12
	5.5 or More	576	10.44	5.25	13.61	6.67	3.18	5.30
CLASS CO	Less than 5.5	129	9.42	4.82	12.07	5.27	2.65	3.52
	5.5 or More	477	10.54	5.30	13.88	6.90	3.34	5.58
CLASS IS	Less than 3	227	9.77	4.94	13.14	8.59	3.37	7.29
	3 or More	379	10.62	5.36	13.71	5.10	3.09	3.42

Appendix C. Sensitivity Analyses.

Table C.1. Multivariate analyses of children's 2017–18 raw score gains in relation to child and site or classroom characteristics and ECERS-3, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)
3-year-olds	-4.260**	0.720	-0.378
	(1.44)	(0.53)	(0.61)
Returning Status	-3.730*	1.674*	-0.943
	(1.71)	(0.68)	(0.78)
Asian	-3.205	-0.249	0.725
	(1.70)	(0.64)	(0.74)
Black	0.209	-0.412	1.013
	(1.75)	(0.66)	(0.77)
Hispanic	0.217	-0.128	0.715
	(1.89)	(0.73)	(0.85)
Other	-2.175	1.027	0.748
	(1.87)	(0.70)	(0.81)
DLL	-1.115	0.319	0.035
	(1.43)	(0.51)	(0.60)
Agency Selected	2.540	0.156	0.269
	(1.34)	(0.53)	(0.58)
HH Income<20k	-1.731	-1.436	0.486
	(3.30)	(1.30)	(1.51)
HH Income 21-40k	-2.294	-0.658	-0.638
	(2.58)	(0.96)	(1.12)
HH Income 41-60k	0.312	-0.704	-0.307
	(2.49)	(0.95)	(1.09)
HH Income 61-80k	1.303	-0.761	1.381
	(2.48)	(0.92)	(1.07)
FPL < 100	-0.385	1.885	-0.970
	(3.27)	(1.31)	(1.50)
FPL 100 to 300	-0.855	-0.217	-0.162
	(2.28)	(0.86)	(0.99)
High Scope	0.076	0.430	-0.410
	(1.36)	(0.55)	(0.58)
Class Size	0.413	0.042	-0.148
	(0.33)	(0.13)	(0.14)
Teacher Qual. Exceeds	1.810	1.250	-1.309
	(1.89)	(0.80)	(0.84)
Teacher Qual. Meets	2.389	0.019	-0.403
	(1.72)	(0.69)	(0.72)
Teacher Black	4.486*	0.659	-1.863*
	(2.08)	(0.87)	(0.91)
Teacher Hispanic	1.328	-0.312	-1.578
	(1.92)	(0.80)	(0.83)
Teacher Asian	3.828	0.641	-1.277
	(2.34)	(0.95)	(1.00)
Teacher Other	2.134	1.078	0.045
	(1.62)	(0.69)	(0.72)
ECERS-3	1.011	-0.244	0.350
	(0.94)	(0.39)	(0.41)
N	702	573	573

* $p<0.05$; ** $p<0.01$; *** $p<0.001$. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days

between tests and an indicator for missing language, income, race, FPL, and teacher qualification and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.2. Multivariate analyses of children's 2017–18 raw score gains in relation to child and site or classroom characteristics and CLASS dimensions, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)
3-year-olds	-4.192**	0.645	-0.448
	(1.45)	(0.53)	(0.62)
Returning Status	-3.686*	1.671*	-0.880
	(1.72)	(0.68)	(0.78)
Asian	-3.079	-0.290	0.772
	(1.71)	(0.64)	(0.74)
Black	0.267	-0.389	1.132
	(1.76)	(0.66)	(0.77)
Hispanic	0.250	-0.178	0.677
	(1.89)	(0.73)	(0.85)
Other	-2.081	1.042	0.820
	(1.88)	(0.70)	(0.81)
DLL	-1.102	0.325	0.067
	(1.43)	(0.51)	(0.60)
Agency Selected	2.726*	0.158	0.403
	(1.32)	(0.52)	(0.57)
HH Income<20k	-1.889	-1.343	0.597
	(3.29)	(1.30)	(1.50)
HH Income 21-40k	-2.364	-0.583	-0.527
	(2.59)	(0.96)	(1.12)
HH Income 41-60k	0.231	-0.601	-0.232
	(2.49)	(0.95)	(1.09)
HH Income 61-80k	1.244	-0.691	1.463
	(2.48)	(0.92)	(1.07)
FPL < 100	-0.160	1.799	-1.098
	(3.26)	(1.31)	(1.50)
FPL 100 to 300	-0.741	-0.269	-0.196
	(2.28)	(0.86)	(0.99)
High Scope	-0.167	0.321	-0.826
	(1.39)	(0.56)	(0.59)
Class Size	0.469	0.038	-0.108
	(0.34)	(0.13)	(0.14)
Teacher Qual. Exceeds	2.114	1.048	-1.352
	(1.85)	(0.78)	(0.82)
Teacher Qual. Meets	2.823	-0.293	-0.726
	(1.68)	(0.67)	(0.71)
Teacher Black	4.612*	0.767	-1.660
	(2.09)	(0.87)	(0.91)
Teacher Hispanic	1.207	-0.192	-1.442
	(1.94)	(0.80)	(0.83)
Teacher Asian	3.771	0.836	-0.914
	(2.39)	(0.96)	(1.02)
Teacher Other	2.123	0.928	-0.170
	(1.66)	(0.69)	(0.72)
CLASS ES average	0.509	-0.569	-1.285
	(1.57)	(0.66)	(0.69)
CLASS CO average	0.362	0.256	1.246*
	(1.25)	(0.52)	(0.54)

CLASS IS average	-0.199	0.253	0.033
	(0.65)	(0.26)	(0.28)
N	702	573	573

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualification and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.3. Multivariate analyses of children's 2017–18 raw score gains in relation to child and site or classroom characteristics and CLASS dimensions, including FCCs

	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)
3-year-olds	-4.359**	0.290	-0.464
	(1.39)	(0.50)	(0.58)
Returning Status	-3.947*	1.471*	-0.863
	(1.71)	(0.67)	(0.76)
Asian	-2.900	0.033	0.821
	(1.70)	(0.63)	(0.71)
Black	0.195	-0.247	1.188
	(1.73)	(0.64)	(0.74)
Hispanic	-0.472	-0.059	0.852
	(1.86)	(0.71)	(0.81)
Other	-2.177	1.233	0.917
	(1.86)	(0.68)	(0.78)
DLL	-1.145	0.226	-0.046
	(1.38)	(0.49)	(0.56)
Agency Selected	2.461	0.050	0.409
	(1.32)	(0.53)	(0.56)
HH Income<20k	-1.568	-1.045	0.176
	(3.23)	(1.25)	(1.42)
HH Income 21-40k	-1.695	-0.642	-0.403
	(2.57)	(0.95)	(1.09)
HH Income 41-60k	0.147	-0.668	-0.182
	(2.48)	(0.94)	(1.07)
HH Income 61-80k	0.993	-0.659	1.464
	(2.49)	(0.92)	(1.05)
FPL < 100	-1.386	1.233	-0.688
	(3.19)	(1.25)	(1.41)
FPL 100 to 300	-0.960	-0.317	-0.240
	(2.29)	(0.85)	(0.97)
FCC	0.103	-0.802	-0.103
	(4.24)	(1.64)	(1.69)
High Scope	-1.248	0.080	-0.828
	(1.32)	(0.54)	(0.54)
Class Size	0.436	-0.015	-0.093
	(0.32)	(0.13)	(0.13)
Teacher Qual. Exceeds	1.498	0.633	-1.342
	(1.84)	(0.79)	(0.79)
Teacher Qual. Meets	2.623	-0.341	-0.740
	(1.68)	(0.69)	(0.69)
Teacher Black	2.656	-0.124	-1.611
	(1.99)	(0.84)	(0.84)
Teacher Hispanic	0.387	-0.577	-1.417
	(1.93)	(0.81)	(0.81)
Teacher Asian	2.604	0.207	-0.805

	(2.35)	(0.96)	(0.97)
Teacher Other	2.106	0.934	-0.187
	(1.66)	(0.72)	(0.71)
CLASS ES average	0.461	-0.320	-1.241
	(1.53)	(0.65)	(0.65)
CLASS CO average	0.623	0.252	1.256*
	(1.23)	(0.52)	(0.52)
CLASS IS average	-0.404	0.188	0.045
	(0.64)	(0.27)	(0.27)
N	735	606	606

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualification and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.4. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and ECERS-3 threshold, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.238	3.471***	0.719	-0.132*	-1.448*
	(1.12)	(1.01)	(1.11)	(0.06)	(0.65)
Returning Status	-2.272	1.346	-0.673	-0.015	0.021
	(1.37)	(1.33)	(1.45)	(0.08)	(0.83)
Asian	-2.944*	-0.652	0.523	-0.038	1.708*
	(1.37)	(1.24)	(1.37)	(0.07)	(0.77)
Black	-0.849	-0.215	-1.216	-0.118	-0.535
	(1.40)	(1.29)	(1.44)	(0.07)	(0.81)
Hispanic	0.042	-1.258	1.591	0.038	0.886
	(1.52)	(1.43)	(1.59)	(0.08)	(0.89)
Other	-2.436	0.591	1.151	0.005	2.477**
	(1.50)	(1.36)	(1.51)	(0.08)	(0.85)
DLL	-0.436	0.820	0.938	0.028	-0.620
	(1.15)	(1.01)	(1.11)	(0.06)	(0.63)
Agency Selected	2.564*	-0.177	0.386	-0.009	-0.489
	(1.06)	(1.00)	(1.06)	(0.06)	(0.60)
HH Income<20k	-1.917	-3.374	2.322	-0.148	-0.575
	(2.64)	(2.55)	(2.80)	(0.15)	(1.59)
HH Income 21-40k	-1.802	-1.482	-1.227	-0.189	-1.375
	(2.07)	(1.88)	(2.09)	(0.11)	(1.18)
HH Income 41-60k	-0.432	-1.258	0.205	-0.095	-0.550
	(2.00)	(1.86)	(2.04)	(0.11)	(1.16)
HH Income 61-80k	0.481	-2.112	0.134	-0.059	-0.008
	(1.99)	(1.81)	(1.99)	(0.11)	(1.13)
FPL < 100	0.423	2.166	-4.144	0.087	0.807
	(2.61)	(2.57)	(2.79)	(0.15)	(1.58)
FPL 100 to 300	-0.510	-1.320	-1.631	0.012	-0.283
	(1.83)	(1.68)	(1.84)	(0.10)	(1.04)
High Scope	0.416	0.144	-0.402	-0.030	0.245
	(1.08)	(1.03)	(1.06)	(0.06)	(0.60)
Class Size	0.398	-0.019	-0.357	-0.002	-0.088
	(0.27)	(0.25)	(0.26)	(0.01)	(0.15)
Teacher Qual. Exceeds	3.066*	1.858	1.333	-0.022	-0.673
	(1.46)	(1.46)	(1.50)	(0.08)	(0.86)
Teacher Qual. Meets	2.844*	-0.664	0.915	0.001	0.387

	(1.28)	(1.21)	(1.26)	(0.07)	(0.71)
Teacher Black	4.995**	1.666	-1.095	-0.078	-0.679
	(1.65)	(1.63)	(1.68)	(0.09)	(0.96)
Teacher Hispanic	1.901	-0.565	-1.166	-0.130	-1.029
	(1.54)	(1.50)	(1.54)	(0.08)	(0.88)
Teacher Asian	4.595*	2.055	-0.048	-0.090	-1.926
	(1.87)	(1.80)	(1.86)	(0.10)	(1.06)
Teacher Other	1.977	3.357**	0.873	-0.003	-0.551
	(1.30)	(1.30)	(1.33)	(0.07)	(0.76)
ECERS-3 above 3.0	-0.348	0.889	0.427	-0.007	0.535
	(1.40)	(1.37)	(1.42)	(0.07)	(0.80)
N	702	573	573	571	574

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualification and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.5. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and CLASS dimensions' thresholds, including FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.387	2.615**	0.571	-0.139*	-1.847**
	(1.07)	(0.97)	(1.05)	(0.06)	(0.62)
Returning Status	-2.480	0.850	-0.752	-0.011	0.017
	(1.37)	(1.32)	(1.43)	(0.08)	(0.81)
Asian	-2.817*	0.020	0.789	-0.061	1.829*
	(1.35)	(1.24)	(1.35)	(0.07)	(0.75)
Black	-0.918	0.184	-0.880	-0.155*	-0.393
	(1.37)	(1.26)	(1.39)	(0.07)	(0.78)
Hispanic	-0.518	-0.881	1.728	0.006	0.805
	(1.48)	(1.40)	(1.54)	(0.08)	(0.86)
Other	-2.553	1.234	1.662	-0.009	2.552**
	(1.48)	(1.35)	(1.47)	(0.08)	(0.83)
DLL	-0.569	0.549	0.486	0.044	-0.590
	(1.11)	(0.97)	(1.06)	(0.06)	(0.60)
Agency Selected	2.338*	-0.264	0.380	-0.011	-0.550
	(1.05)	(1.00)	(1.04)	(0.06)	(0.59)
HH Income<20k	-1.518	-2.799	1.272	-0.205	-0.877
	(2.58)	(2.46)	(2.67)	(0.14)	(1.52)
HH Income 21-40k	-1.355	-1.306	-0.621	-0.184	-1.308
	(2.05)	(1.87)	(2.05)	(0.11)	(1.16)
HH Income 41-60k	-0.505	-1.278	0.402	-0.093	-0.406
	(1.98)	(1.85)	(2.01)	(0.11)	(1.13)
HH Income 61-80k	0.288	-1.918	0.184	-0.076	0.083
	(1.99)	(1.82)	(1.97)	(0.11)	(1.12)
FPL < 100	-0.640	1.228	-3.307	0.114	1.032
	(2.55)	(2.47)	(2.66)	(0.14)	(1.50)
FPL 100 to 300	-0.673	-1.464	-1.886	-0.011	-0.328
	(1.82)	(1.68)	(1.83)	(0.10)	(1.03)
FCC	-0.489	-5.807	-2.451	-0.393*	-1.777
	(3.31)	(3.03)	(3.11)	(0.16)	(1.76)
High Scope	-0.371	-0.299	-0.870	-0.037	0.132

	(1.03)	(0.99)	(1.00)	(0.05)	(0.57)
Class Size	0.320	-0.090	-0.240	-0.007	-0.122
	(0.25)	(0.24)	(0.24)	(0.01)	(0.14)
Teacher Qual. Exceeds	2.882	0.834	0.644	-0.078	-0.646
	(1.48)	(1.50)	(1.51)	(0.08)	(0.86)
Teacher Qual. Meets	3.122*	-1.319	-0.039	-0.029	0.231
	(1.35)	(1.30)	(1.31)	(0.07)	(0.74)
Teacher Black	3.219*	0.755	-0.480	-0.122	-0.847
	(1.61)	(1.59)	(1.60)	(0.09)	(0.91)
Teacher Hispanic	1.344	-1.050	-1.210	-0.168*	-0.706
	(1.58)	(1.56)	(1.56)	(0.08)	(0.88)
Teacher Asian	3.506	1.410	0.557	-0.128	-2.135*
	(1.86)	(1.79)	(1.81)	(0.10)	(1.02)
Teacher Other	2.145	2.990*	0.573	-0.029	-0.525
	(1.33)	(1.35)	(1.35)	(0.07)	(0.76)
CLASS ES above 5.5	1.435	-1.930	-2.354	-0.144	0.458
	(2.31)	(2.28)	(2.32)	(0.12)	(1.31)
CLASS CO above 5.5	-0.848	2.068	2.389	0.080	1.408
	(1.29)	(1.26)	(1.27)	(0.07)	(0.72)
CLASS IS above 3.0	-0.281	-0.012	0.545	0.012	-1.080
	(1.01)	(0.99)	(0.99)	(0.05)	(0.56)
N	735	606	606	604	607

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualification and race. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.6. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and overall ECERS-3 with Agency Fixed Effects, excluding FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.471	3.628***	0.739	-0.113~	-1.386*
	(1.16)	(1.04)	(1.14)	(0.06)	(0.68)
Returning Status	-2.299	1.732	0.356	0.017	0.280
	(1.40)	(1.34)	(1.48)	(0.08)	(0.85)
Asian	-2.856*	-0.937	0.354	-0.032	1.563*
	(1.37)	(1.23)	(1.37)	(0.07)	(0.77)
Black	-0.868	-0.255	-1.050	-0.094	-0.355
	(1.40)	(1.29)	(1.44)	(0.08)	(0.81)
Hispanic	0.309	-1.267	1.138	0.046	0.908
	(1.51)	(1.42)	(1.58)	(0.08)	(0.89)
Other	-2.353	0.341	1.416	0.016	2.396**
	(1.50)	(1.35)	(1.50)	(0.08)	(0.85)
DLL	-0.681	0.330	0.795	0.010	-0.720
	(1.16)	(1.01)	(1.12)	(0.06)	(0.64)
Agency Selected	3.914**	1.432	1.801	0.063	-0.145
	(1.35)	(1.24)	(1.37)	(0.07)	(0.78)
HH Income<20k	-1.913	-3.396	2.575	-0.163	-0.561
	(2.67)	(2.55)	(2.80)	(0.15)	(1.61)
HH Income 21-40k	-2.490	-1.714	-1.502	-0.221*	-1.424

	(2.07)	(1.88)	(2.09)	(0.11)	(1.19)
HH Income 41-60k	-1.094	-1.407	-0.080	-0.128	-0.753
	(1.99)	(1.84)	(2.04)	(0.11)	(1.16)
HH Income 61-80k	0.094	-1.975	0.231	-0.084	-0.107
	(1.99)	(1.80)	(1.99)	(0.11)	(1.14)
FPL < 100	-0.209	1.752	-4.728	0.078	0.839
	(2.66)	(2.57)	(2.79)	(0.15)	(1.60)
FPL 100 to 300	-0.061	-1.682	-1.710	0.015	-0.284
	(1.83)	(1.67)	(1.84)	(0.10)	(1.05)
High Scope	1.268	11.346	-10.407	0.486	-1.157
	(10.51)	(8.64)	(9.52)	(0.50)	(5.45)
Class Size	0.112	0.090	0.444	0.019	-0.017
	(0.39)	(0.35)	(0.38)	(0.02)	(0.22)
Teacher Qual. Exceeds	-4.322	-13.185	8.327	-0.601	1.237
	(10.99)	(9.13)	(10.06)	(0.53)	(5.75)
Teacher Qual. Meets	1.505	-0.868	1.522	-0.013	0.547
	(1.69)	(1.46)	(1.61)	(0.08)	(0.92)
Teacher Black	3.368	-0.103	-1.156	0.033	-0.140
	(2.16)	(1.98)	(2.18)	(0.11)	(1.25)
Teacher Hispanic	1.758	-0.467	-1.463	-0.110	-0.997
	(1.66)	(1.48)	(1.63)	(0.09)	(0.93)
Teacher Asian	3.972	-2.884	-3.638	-0.319*	-4.250**
	(2.81)	(2.45)	(2.72)	(0.14)	(1.55)
Teacher Other	1.875	3.965**	1.750	0.047	-0.424
	(1.36)	(1.26)	(1.38)	(0.07)	(0.79)
ECERS-3	0.322	-1.345	-1.157	-0.096*	-0.887
	(0.86)	(0.79)	(0.87)	(0.05)	(0.50)
N	702	573	573	571	574

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualifications, race and agency fixed effects. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table C.7. Multivariate analyses of children's 2017–18 standard score gains in relation to child and site or classroom characteristics and overall CLASS dimensions with Agency Fixed Effects, including FCCs

Variables	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Math (WJ/WM-AP)	Executive Function	
				DCCS	PT
3-year-olds	-1.702	2.410*	0.476	-0.114	-1.597*
	(1.11)	(1.00)	(1.09)	(0.06)	(0.64)
Returning Status	-2.481	1.029	0.229	0.017	0.184
	(1.39)	(1.33)	(1.45)	(0.08)	(0.83)
Asian	-2.925*	-0.564	0.384	-0.039	1.673*
	(1.36)	(1.22)	(1.35)	(0.07)	(0.76)
Black	-0.850	0.128	-0.693	-0.127	-0.252
	(1.37)	(1.27)	(1.39)	(0.07)	(0.79)
Hispanic	-0.175	-0.998	1.346	0.022	0.904
	(1.48)	(1.39)	(1.53)	(0.08)	(0.87)
Other	-2.346	0.846	1.770	-0.001	2.375**
	(1.47)	(1.33)	(1.46)	(0.08)	(0.83)
DLL	-0.523	0.068	0.501	0.025	-0.706

	(1.13)	(0.98)	(1.08)	(0.06)	(0.61)
Agency Selected	3.892**	0.979	1.491	0.009	-0.633
	(1.31)	(1.22)	(1.33)	(0.07)	(0.76)
HH Income<20k	-1.379	-2.406	1.811	-0.212	-0.909
	(2.60)	(2.48)	(2.71)	(0.15)	(1.55)
HH Income 21-40k	-1.965	-1.282	-0.747	-0.221*	-1.387
	(2.04)	(1.87)	(2.06)	(0.11)	(1.17)
HH Income 41-60k	-1.145	-0.936	0.221	-0.127	-0.632
	(1.98)	(1.84)	(2.02)	(0.11)	(1.15)
HH Income 61-80k	-0.052	-1.436	0.458	-0.109	-0.182
	(1.98)	(1.81)	(1.98)	(0.11)	(1.13)
FPL < 100	-1.040	0.586	-4.061	0.099	1.039
	(2.58)	(2.48)	(2.68)	(0.14)	(1.53)
FPL 100 to 300	-0.051	-1.839	-2.122	0.001	-0.335
	(1.82)	(1.68)	(1.83)	(0.10)	(1.04)
FCC	1.240	-12.659*	9.340	-0.320	-1.382
	(6.64)	(5.63)	(6.14)	(0.33)	(3.51)
High Scope	-9.193	3.303	-8.112	0.101	-1.049
	(5.63)	(4.67)	(5.10)	(0.27)	(2.91)
Class Size	-0.035	-0.168	0.325	-0.007	-0.173
	(0.36)	(0.32)	(0.35)	(0.02)	(0.20)
Teacher Qual. Exceeds	4.726	-7.581	5.307	-0.303	0.947
	(6.52)	(5.56)	(6.06)	(0.32)	(3.46)
Teacher Qual. Meets	1.721	-2.296	0.426	-0.102	-0.181
	(1.59)	(1.39)	(1.53)	(0.08)	(0.87)
Teacher Black	1.430	-1.572	-1.344	-0.097	-0.866
	(1.99)	(1.81)	(1.98)	(0.11)	(1.13)
Teacher Hispanic	1.138	-0.837	-1.431	-0.144	-1.166
	(1.66)	(1.47)	(1.61)	(0.09)	(0.92)
Teacher Asian	3.234	-2.971	-2.797	-0.303*	-3.997*
	(2.91)	(2.54)	(2.79)	(0.15)	(1.59)
Teacher Other	1.695	3.123*	1.037	0.028	-0.457
	(1.39)	(1.28)	(1.40)	(0.07)	(0.80)
CLASS ES average	0.471	-1.616	-1.809	-0.062	-0.417
	(1.26)	(1.17)	(1.28)	(0.07)	(0.73)
CLASS CO average	-0.072	0.316	0.764	0.100	0.706
	(1.11)	(1.03)	(1.12)	(0.06)	(0.64)
CLASS IS average	-0.097	1.115*	0.649	-0.054	-0.585
	(0.54)	(0.49)	(0.53)	(0.03)	(0.30)
N	735	606	606	604	607

* $p<0.05$; ** $p<0.01$; *** $p<0.001$. Note: Reference groups omitted from the estimation are Males, White, English, FPL 300%+, Income>80 thousand, and Creative Curriculum. Other controls are pre-test, age in months, days between tests and an indicator for missing language, income, race, FPL, and teacher qualifications, race and agency fixed effects. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Appendix D. P-values for tests of differences in means.

Table D.1. P-values for T-tests comparing distributions

P(T<=t) two-tail	16' vs. 17'	17' vs. 18'	17' vs. 18' *including FCCs*
ECERS-3	**0.049**	0.442	n/a
CLASS ES	0.346	0.444	0.894
CLASS CO	0.620	**0.021**	**0.064**
CLASS IS	0.107	**0.099**	**0.062**

Table D.2. P-values for T-tests for comparisons of CLASS means between classrooms and FCCs

Domains and Dimensions	P-value
Emotional Support	**0.113**
1. Positive Climate	0.429
2. Negative Climate*	0.874
3. Teacher Sensitivity	0.145
4. Regard for Student Perspectives	0.426
Classroom Organization	**0.105**
5. Behavior Management	0.153
6. Productivity	0.341
7. Instructional Learning Formats	0.206
8. Facilitation of Learning & Dev.	n/a
Instructional Support	**0.733**
9. Concept Development	0.811
10. Quality of Feedback	0.824
11. Language Modeling	0.510

Table D.3. P-values for T-tests and Bonferroni tests comparing quality across children subgroups, includes FCCs

	Ethnicity	Gender	FPL	DLL
	Bonferroni Prob>chi2	T-Test Pr(|T| > |t|)	Bonferroni Prob>chi2	Bonferroni Prob>chi2
ECERS-3	0.116	0.604	0.459	0.028
CLASS ES	**0.000**	0.794	**0.000**	**0.000**
CLASS CO	**0.000**	0.384	0.401	0.886
CLASS IS	**0.006**	0.152	**0.077**	**0.000**