NIEER
NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

EVALUATION OF THE PHILADELPHIA PREK PROGRAM

Year 7 Report

Erin Harmeyer, PhD, Milagros Nores, PhD, Zijia Li, PhD, Carmen Espinosa, M.Ed. The National Institute for Early Education Research

About the Authors

Erin Harmeyer, Ph.D. Dr. Harmeyer is an Assistant Research Professor at the National Institute for Early Education Research (NIEER) at Rutgers University. Dr. Harmeyer conducts research at NIEER related to early childhood programs and evaluation, supporting projects in New Jersey, Rhode Island, and Pennsylvania.

Milagros Nores, Ph.D. Dr. Nores is Co-Director of Research and Associate Research Professor at The National Institute for Early Education Research (NIEER) at Rutgers University. Dr. Nores conducts research at NIEER on early childhood policy, programs, and evaluation, both nationally and internationally.

Zijia Li, Ph.D. Dr. Li is an Assistant Research Professor at the National Institute of Early Education Research (NIEER) at Rutgers University. Dr. Li is an experienced psychometrician and statistician. She has led and participated leading and conducting rigorous reliability and validity research studies for multiple subjects, including Peabody Picture Vocabulary Test III and IV, the Hawaii Early Learning Profile®, High/Scope Child Observation Record.

Carmen Espinosa, M.Ed. Espinosa is Project Coordinator I at the National Institute for Early Education Research (NIEER). She leads NIEER's field work on the PHLpreK Evaluation Study and related work in New Jersey. She has contributed to NIEER's research in Philadelphia and New Jersey.

Correspondence regarding this report should be addressed to Milagros Nores at the National Institute for Early Education Research. Email: mnores@nieer.org.

Suggested citation: Harmeyer, E., Nores, M., Li, Z., & Espinosa, C. E. (2023). *Evaluation of the Philadelphia PreK program. Year 7 report.* National Institute for Early Education Research.

Table of Contents

Introduction

Philadelphia's Preschool Program (PHLpreK) has recently concluded its seventh year of programming. The program was initiated after a May 2015 vote where city voters approved the creation of the Philadelphia Commission on Universal Pre-kindergarten. The commission was given the responsibility of proposing a universal pre-K program to provide high-quality, affordable, and accessible services to children in the city, ages 3 and 4. This is the seventh year that the National Institute for Early Education Research (NIEER) at Rutgers has conducted an evaluation assessing program components, program quality, and children's learning and development.

Previous reports for this evaluation have summarized the importance of high-quality preschool education to reduce persistent achievement gaps in kindergarten and throughout primary (e.g., Barnett et al., 2018; Nores et al., 2017, 2018, 2019, 2020, 2021, 2022). We have highlighted research that has shown that high-quality preschool education programs can produce lasting effects on school success and achievement and reduce achievement gaps at kindergarten entry and beyond (e.g., Barnett, 2008; Barnett & Jung, 2021; Barnett.& Nores, 2015; Ceci & Papierno, 2005; Duncan & Murnane, 2011; Gray-Lobe, 2023; Johnson et al., 2023). Strengthening and supporting preschool systems and supporting them to achieve and sustain high-quality requires continuous systems of improvement that include measurement and assessment, training and technical assistance and use of data to align system weaknesses and strengths with the initiative to increase quality over time (Barnett & Frede, 2017; Nores & Harmeyer, 2023). This includes understanding the quality of classroom processes and interactions, space, and use of time (Pianta & Hamre, 2009; Hamre et al., 2014).

This report summarizes the results of the year seven evaluation of Philadelphia's PreK Program (PHLpreK). The report provides a comprehensive overview of the environment and teaching interactions in these classrooms and summarizes the progress made by children in the program. In addition, this report also summarizes quality and the gains of children in a small group of feasibly comparable classrooms in the city. The present report is one of the various components of a longitudinal evaluation since 2017, with the goal of supporting a data-driven continuous improvement approach to support improvements in quality in the city's program, alongside its expansion.

Findings indicate that PHLpreK classrooms consistently demonstrate high to moderate levels of quality in the emotional support and classroom organization domains. Quality along these domains looks similar to the year prior, despite the large program expansion which incorporated many new classrooms and providers. In contrast, classroom scores for instructional supports are low and this trend has persisted over time (i.e., lower than 3.0, on average). There was even a decrease relative to last year. We explore quality separately for several subgroups of interest, including Star level, lead teacher credentials, area of study, PHLpreK partner agency, and new and returning sites. Small differences were found between subgroups, and are reported.

For the first time since 2018-19, we assessed children's developmental gains over the school year for a sample of children in a subset of 50 classrooms. We report overall gains and explored differences among subgroups of children. We also assessed how centers and teaching and learning characteristics relate to child gains. Overall, the sampled children exhibited higher gains on executive functions and literacy compared to the cohort assessed pre-pandemic, and lower gains on receptive vocabulary and mathematics.

Study Methods

The PHLpreK Evaluation is a multi-year, multi-site study encompassing several components to provide a comprehensive perspective of the program's design, its quality, and its impact on children over time. This report focuses on the seventh year of the study. Data collection included assessing child gains on a sub-sample of children (fall and spring assessments), and classroom observations across all providers to inform the following research questions:

1. What is the observed quality of children's classroom experiences and how does it compare relative to the prior years?
2. How does quality in PHLpreK classrooms compare to quality in other classrooms in the City of Philadelphia? And to other programs in the country?
3. What are the learning gains of children in vocabulary, literacy, math, and executive functions through 2022–23, and how did gains relate to classroom quality and children's background characteristics? How do these compare relative to prior years and to children in other classrooms in the City of Philadelphia?

The PHLpreK evaluation was designed to assess program progress and quality over time, with the goal of informing a continuous improvement approach to quality. In Year 1, the research team measured classroom quality. In Years 2 and 3, the research team assessed children's learning and development at the beginning and end of the school year and repeated the observations of classroom quality. In Year 4, the research team collected some classroom and child-level data, but study procedures were interrupted by the onset of the COVID-19 pandemic. In year 5 (2020-21), teachers completed a self-report measure of classroom quality, and directors participated in focus groups discussing the impact of the pandemic on their programming. In year 6, data was collected from all classrooms enrolled in the PHLpreK program in order to measure quality. This year, year 7 for the program and the evaluation, we again measured quality in all PHLpreK classrooms, along with quality in a sample of other classrooms in the City of Philadelphia. In addition, we collected data on children's gains across the school year in a subset of 50 PHLpreK classrooms along with a small number of non-PHLpreK classrooms in the City of Philadelphia. Procedures and measures are described in detail below. Children were assessed early in the Fall of 2022, and again at the end of the school year in the Spring of 2023. Classroom observations assess teacher-child interactions and quantify children's experiences during a typical learning day. Classroom observations took place between February and June 2023. As in previous years, quality was assessed using well-known observation protocols during one visit of about two and a half hours.

1. Sample

In the 2022–23 school year, classroom quality was assessed with one instrument: CLASS Second Edition (pre-K – 3rd). The CLASS was used in 283 PHLpreK classrooms (center and home-based). We also conducted classroom observations in 24 additional comparison classrooms in the city of Philadelphia.

In addition, NIEER assessed 153 children in 49 PHLpreK classrooms at both pre- and post-test. To recruit children, consent forms were distributed to families as part of the PHLpreK enrollment process. A total of 206 children were assessed at pre-test with family consent, and 176 children were assessed at post-test; we were able to assess 153 children at both pre- and post-test, and selected children to replace those lost to unenrolling from programs or being unavailable at post-test. We randomly selected approximately four children per classroom. The final sample of children enrolled in PHLpreK programs with data at both timeframes was 60.9% African American, 11.3% Hispanic, 13.2% White, and 14.6% other. This is closely comparable to the K-12 PHL school district demographics of 52% African American, 21% Latino, 14% White, and 13% other.[1]

2. Measures and Procedures

Classroom quality was captured using one observational instrument: *The Classroom Assessment Scoring System Pre-K- 3rd Second Edition* (*CLASS*; *Classroom Assessment Scoring System 2nd Edition,* 2022). The CLASS measures teacher-child interactions and classroom processes; this was the second year using the second edition of the observation tool. According to the measure developers, CLASS Second Edition was developed using more equitable and inclusive measures of effective interactions, and includes increased representation of children and teachers in training materials. The developers state it allows for consideration of possible variations in effective interactions due to context, and aims to help observers confront bias in their own observations (*Classroom Assessment Scoring System 2nd Edition, 2022*). Notably, the updated tool does not make any changes to the dimensions or domains that are scored; revisions focus on broadening the description of effective interactions. In addition, the revised tool covers the range of preschool through third grade classrooms.

Children were assessed with a measure of receptive language (the *Peabody Picture Vocabulary Test—Fourth Edition or PPVT-IV*; Dunn & Dunn, 2007), emerging literacy (the letter-word identification subtest from the *Woodcock-Johnson Psycho-Educational Battery—Fourth Edition or WJ-IV;* Schrank, Mather & McGrew, 2014) and mathematics (the applied problems subtest from the WJ-*IV*). In addition, children were assessed with one measure of executive functions, which captures children's inhibitory control, short term memory, and attention, the *Dimensional Change Card Sort Task* (DCCS; Zelazo, 2006). More detail on child measures is provided in Appendix A.

Observers were trained to reliability before conducting observations of classroom quality. CLASS observers were trained using the Teachstone® virtual training platform, completed the online reliability certification test required by Teachstone® and met their requirement (80%) for observer certification. Observers were also trained in practices and procedures for conduct and required to complete background checks, as well as training in human subjects research (human subject protections, ethical issues, etc.). In addition, observers were required to pass a calibration assessment about mid-point through the data collection period.

[1] https://dashboards.philasd.org/extensions/philadelphia/index.html#/

Results

Results are presented first for classroom observations followed by a comparison to a set of comparable classrooms and centers in the city of Philadelphia. The following section reports children's gains across child and center characteristics and in relation to observed classroom quality. We conclude with a summary of the findings and recommendations.

1. Classroom Observations

CLASS Pre-K Results

Average CLASS scores for PHLpreK classrooms across all domains and dimensions are reported in Table 1. Patterns are consistent with the field and previous years, with instructional support scoring lower than other domains. Emotional Support (ES) scores look very similar to those recorded in 2022 (5.86 in 2022 and 5.81 in 2023). This is also the case for Classroom Organization (CO) (5.40 in 2022 and 5.42 in 2023). However, scores on the Instructional Support domain are lower in 2023 than they were in 2022 (2.75 in 2022, and 2.45 in 2023). The only statistically significant difference in scores from 2022 to 2023 is for Instructional Support. Results for each domain are discussed further below.

Observed decreases in this seventh year in ES were minimal, of 0.06 SD (standard deviations),[2] and in CO these increases were also minimal, of 0.02 SD. In contrast, for IS there was a decrease of 0.28 SD.

[2] Standard deviation is a measure of variation in the data. 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. It also helps to understand change, by dividing change by the standard deviation of the previous year. This helps understand how much of a standard deviation a distribution has changed.

Table 1. PreK CLASS Dimension and Domain Means and Ranges.

CLASS Dimensions and Domains	2017	2018	2019	2020	2022	2023
	Mean	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**
	(Range)	**(Range)**	**(Range)**	**(Range)**	**(Range)**	**(Range)**
	N=139	**N=137**	**N=147**	**N=102****	**N=270**	**N=283**
Emotional Support Domain (ES)	**5.85**	**5.64**[a]	**6.01**	**5.74**	**5.86**	**5.81**
	(2.85-6.90)	**(3.20-6.95)**	**(3.05-7.00)**	**(3.55-6.80)**	**(2.75-7.00)**	**(3.15-7.00)**
1. Positive Climate	5.90	5.73	6.13	5.77	5.95	6.07
	(1.60-7.00)	(3.20-7.00)	(2.40-7.00)	(3.20-7.00)	(2.40-7.00)	(3.20-7.00)
2. Negative Climate*	6.77	6.67	6.91	6.74	6.78	6.79
	(5.00-7.00)	(4.00-7.00)	(5.40-7.00)	(4.2-7)	(3.80-7.00)	(3.40-7.00)
3. Teacher Sensitivity	5.69	5.52	5.89	5.58	5.54	5.69
	(2.20-7.00)	(2.80-7.00)	(1.60-7.00)	(3.20-7.00)	(1.80-7.00)	(1.20-7.00)
4. Regard for Student Perspectives	5.03	4.65	5.11	4.88	5.19	4.70
	(2.00-6.80)	(2.40-7.00)	(1.60-7.00)	(2.8-6.8)	(2.00-7.00)	(1.00-7.00)
Classroom Organization Domain (CO)	**5.34**	**5.28**	**5.60**	**5.26**	**5.40**	**5.42**
	(1.87-6.93)	**(2.80-6.93)**	**(2.40-7.00)**	**(3.20-6.80)**	**(1.87-6.93)**	**(2.07-7.00)**
5. Behavior Management	5.49	5.48	5.81	5.54	5.44	5.67
	(1.60-7.00)	(2.80-7.00)	(2.40-7.00)	(3.00-7.00)	(2.00-7.00)	(1.80-7.00)
6. Productivity	5.76	5.65	5.72	5.54	5.76	5.66
	(1.80-7.00)	(2.80-7.00)	(2.40-7.00)	(3.40-7.00)	(1.20-7.00)	(1.80-7.00)
7. Instructional Learning Formats	4.77	4.72	5.27	4.68	5.00	4.93
	(1.60-7.00)	(1.80-6.80)	(2.00-7.00)	(2.40-6.60)	(1.80-7.00)	(1.20-7.00)
Instructional Support Domain (IS)	**2.41**	**2.05**[a]	**2.54**	**2.30**	**2.75**	**2.45**[a]
	(1.00-5.00)	**(1.00-4.60)**	**(1.00-5.33)**	**(1.33-4.13)**	**(1.00-5.80)**	**(1.00-6.40)**
8. Concept Development	2.09	1.84	2.27	2.10	2.50	2.10
	(1.00-4.80)	(1.00-4.00)	(1.00-5.60)	(1.00-4.00)	(1.00-6.60)	(1.00-6.60)
9. Quality of Feedback	2.23	1.91	2.53	2.10	2.65	2.44
	(1.00-5.00)	(1.00-4.40)	(1.00-5.20)	(1.00-4.20)	(1.00-6.00)	(1.00-6.60)
10. Language Modeling	2.91	2.41	2.80	2.70	3.11	2.80
	(1.00-5.20)	(1.00-5.60)	(1.00-5.80)	(1.40-4.40)	(1.00-6.00)	(1.00-6.00)

*The Negative Climate dimension is reverse scored so that a high score represents “good.” [a]Statistically significant difference between 2022 and 2023. **No scores recorded for 2021 due to the COVID-19 pandemic; which also limited data collection in 2020.

The changes in the distribution of ES, CO, and IS scores across the years are shown in Figures 1, 2, and 3, respectively. Some research appears to support thresholds for ES and CO above 5 and IS above 3 as necessary to evidence a relation between quality and children’s outcomes (other research defines these as slightly higher, at 5.5 and 3.5) (Burchinal et al., 2009;

Burchinal et al., 2014; Hatfield, et al., 2016). Emotional support scores have, on average, increased with a higher number of classrooms reaching scores of 6 and 7. The number of classrooms scoring at 5+ in ES was up in 2023 as compared to 2022 – from 86% in 2022 to 88% in 2023 (Figure 1). For CLASS CO, an improvement in classrooms scoring above the 5 threshold is also observed, with 67% of classrooms meeting this bar in 2022, compared to 77% in 2023 (Figure 2). The distribution for CLASS IS has shifted to the left in 2023, however, with 37% of classrooms scoring above 3 in 2022, compared to 23% of classrooms meeting this bar in 2023 (Figure 3).

Figure 1. Distribution of CLASS Emotional Support scores for 2017 – 2023.

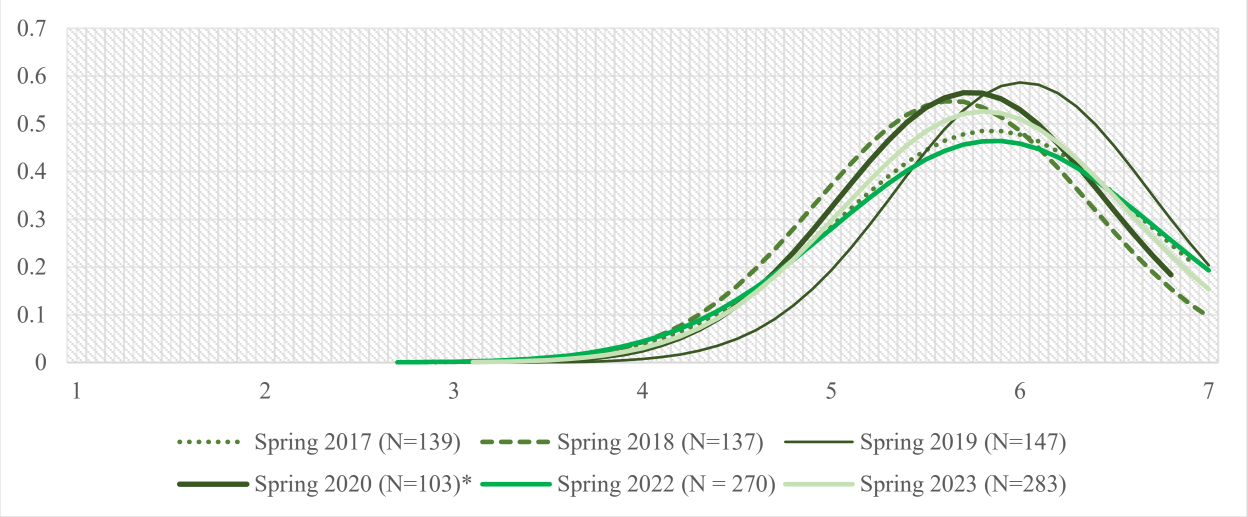


Figure 2. Distribution of CLASS Classroom Organization scores for 2017 – 2023.

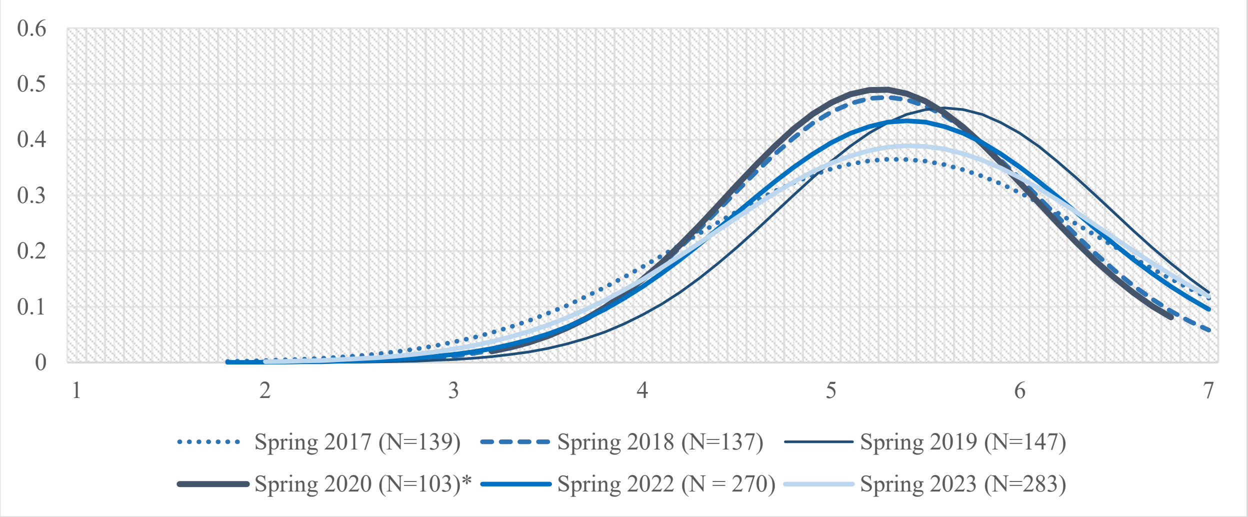
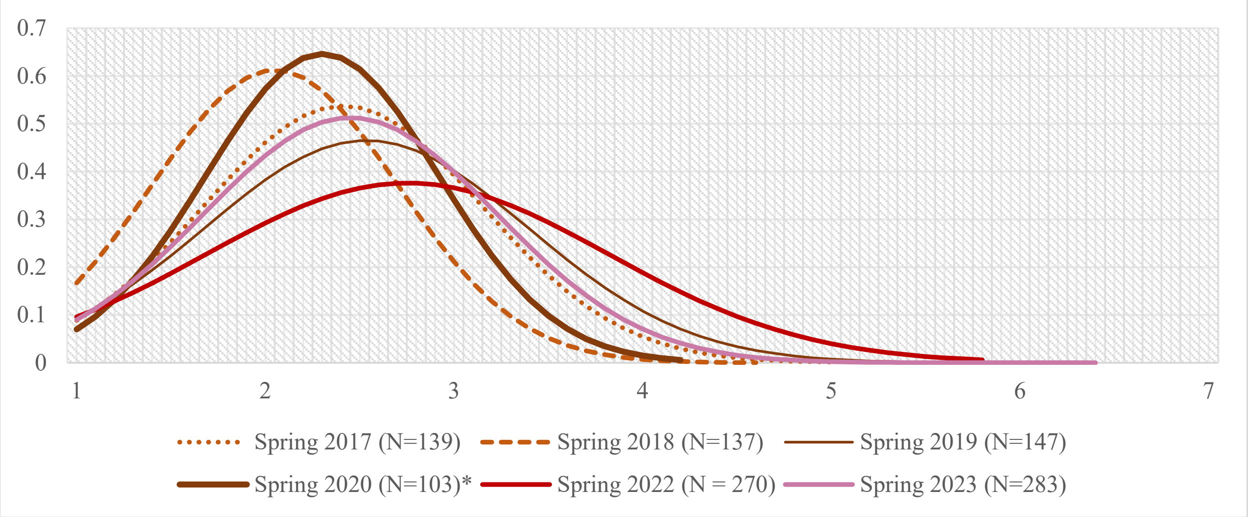


Figure 3. Distribution of CLASS Instructional Support scores for 2017 – 2023.



CLASS Pre-K – 3rd Domains

The Emotional Support (ES) domain focuses on teaching behaviors that support the development of supportive relationships and interactions between teachers and children, and that help children enjoy the learning process and feel comfortable in the classroom. The overall mean score for ES is 5.81 (SD 0.76), showing that on average, teachers are offering emotional support that is high quality. The minimum score is 3.15, which indicates that no classrooms in the PHLpreK program are offering poor-quality emotional support. The highest scoring dimension is Negative Climate (6.79, this dimension is reverse scored so a higher score represents “good”) indicating that on average classrooms exhibited few negative interactions between teachers and children and among children. The lowest scoring dimension is Regard for Student Perspectives (4.70), and this has consistently been the lowest-scoring dimension in the ES domain across the years. Increasing this dimension requires that teachers become flexible and follow the lead of children, provide choice in what children are doing, and encourage student responsibility. Additional opportunities for children to express their ideas and to be involved in activities that will allow them to be active would further increase this score.

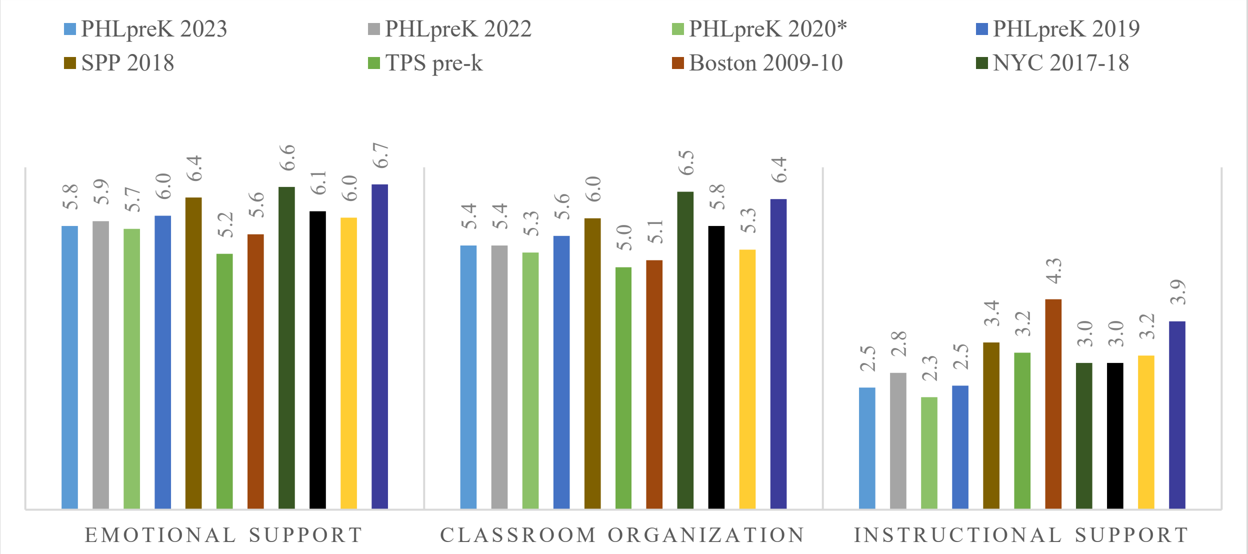
The Classroom Organization (CO) domain focuses on making the most of instructional time and routines, setting effective behavioral expectations, and providing activities that maximize children’s interests and engagement. The average mean score for the Classroom Organization Domain is 5.42 (SD 1.03). This high score indicates that in general, teachers demonstrate effective methods to both prevent and redirect misbehavior, and students for the most part are compliant, demonstrating little aggression and defiance. High scores also indicate that teachers plan ahead, maximize learning time, and focus students’ attention to the learning objectives. Within this domain, Instructional Learning Formats scored lower than the other two dimensions (4.93), which is also consistent with previous years’ scores. Increasing this dimension requires consistent use of interesting and creative materials, actively facilitating and maintaining interest in the lessons and activities, and ensuring the learning objectives are clear. To further increase this domain, teacher involvement in learning activities and exposure to opportunities that allow children to use different modalities, including hands-on activities, is required.

Instructional Support measures the ways in which teachers encourage analysis and reasoning, prompt children to think more deeply through high-quality feedback, and encourage and advance language development. This domain has consistently scored lower across preschool evaluations and systems. However, it is a critically central domain to further children's learning and development. The average IS score is 2.45 (SD 0.78) with averages ranging from 1 to 6.40 on a 7-point scale. Consistent with prior years' evaluations, Concept Development and Quality of Feedback were the lowest scoring dimensions, both in this domain and on the tool as a whole, with scores of 2.10 and 2.44, respectively. Concept Development measures how well teachers use activities with students that encourage discussion and reasoning, the opportunities they provide children for creating and brainstorming, how well they integrate current content with previous lessons, and the connections they make to children's lives. Quality of Feedback measures how well teachers provide additional information during lessons, engage in back-and-forth exchanges with students, prompt thought processes, and provide hints and assistance when concepts are difficult to understand. Scores in the Language Modeling dimension are higher but still average below the threshold of 3. Consistent and intentional use of strategies is critical to increasing scores in this dimension, particularly as they remain below the threshold for high quality.

CLASS Pre-K Comparison of Programs

Figure 4 reports score patterns for PHLpreK in relation to those of other cities and states in which the CLASS has been used. The PHLpreK CLASS scores from 2019, 2020, 2022, and 2023 are reported by domain together with scores from various other programs in the U.S. This includes high-quality city programs such as the Seattle Preschool Program (SPP), the NYC Pre-K for All program and Boston's program.

Figure 4. Comparison of PHLpreK CLASS scores with other programs.



Note: SPP is Seattle's preschool program, reported in Nores, et. al (2019); Boston results are reported in Weiland, et. al (2013); TPS is Tulsa's preschool program, reported in Phillips et. al (2009); NYC is reported in NYC Department of Education (2018).

CLASS Pre-K Domains for Selected Center Characteristics

Table 2 shows CLASS domain scores for selected program-level characteristics. Classrooms with lower star levels (e.g., 1-3) score lower on all domains, although these differences were not significant, and the majority of sites in the PHLpreK program were rated STAR 4. In terms of teacher credentials, classrooms with teachers with a master's degree scored highest on the ES and CO domains. Similarly, all classrooms with a lead teacher with at least a 2-year degree scored higher than teachers without a degree or missing information on IS, although these differences were slight. Concerning partner agency, classrooms in sites in collaboration with PHMC scored lower on the ES and IS domains than did school district programs, but as with STAR level, the sample size of school district programs is much smaller (15 classrooms total) than PHMC affiliated programs, and so any differences should also be interpreted with caution. We also analyzed scores for child care centers and family/group child care homes (combining both types of home-based programs). These differences were also slight, with family child care providers (FCCs) scoring slightly higher on all domains, although these differences were not significant and the sample size of FCCs/group FCCs was quite small. Finally, there were no statistically significant differences noted between new and returning programs, with similar scores across all domains for all sites, although returning programs scored higher on CO and IS.

Table 2. CLASS domains scores by subgroups, N = 283.

		CLASS Mean Scores		
		Emotional Support	Classroom Organization	Instructional Support
STAR Level	1-3 (n=30)	5.79	5.33	2.34
	4 (n=253)	5.81	5.43	2.46
Lead Teacher Credential	No Degree/Some College (n=50)	5.74	5.32	2.37
	AA (n=80)	5.82	5.42	2.53
	BA (n=69)	5.82	5.46	2.51
	Master's Degree (n=50)	6.01	5.75 [a]	2.48
	Missing (n=34)	5.58	5.02 [a]	2.20
New or Returning Site	New (n=33)	5.80	5.24	2.34
	Returning (n=250)	5.81	5.45	2.46
PHLpreK Partner Agency	PHMC (n=268)	5.79	5.41	2.43
	SDP (n=15)	6.18	5.72	2.85
Program Type	FCC/Group FCC (n=23)	5.84	5.53	2.60
	Child Care Center (n=260)	5.81	5.41	2.44

[a]Differences are statistically significant for teachers with a Master's Degree and those with missing information on the Classroom Organization dimension.

CLASS Pre-K Domains for PHLpreK and Comparison classrooms

This year we assessed classroom quality using the CLASS Second Edition in 24 additional, randomly selected classrooms.[3] It is important to take into account that only such a small group

[3] A list of providers for the City of Philadelphia was used. Providers were categorized as Head Start, Center-based non-Head Start, and home providers. A random list was created for a target of N=50 classrooms, with a percentage of each of these groups targeted that resembled the current PHLpreK sites distribution across these three groups. Participation was very low and we ultimately were only able to recruit 24 programs that would allow us to complete the observations and child assessments.

of programs accepted participating when interpreting the comparisons below, and caution is warranted. The sample of control classrooms is 83% in centers, and 17% are home-based providers.

A comparison of overall scores for PHLpreK providers and the sample of comparison providers is summarized in Table 3. On average, PHLpreK classrooms exhibited higher scores across all CLASS domains, and these differences were significant for all three domains. PHLpreK settings on average evidence higher variation, particularly for the Instructional Support domain. Statistically significant differences are marked with an asterisk. This is the case for all dimensions under each domain, with the exception of teacher sensitivity.

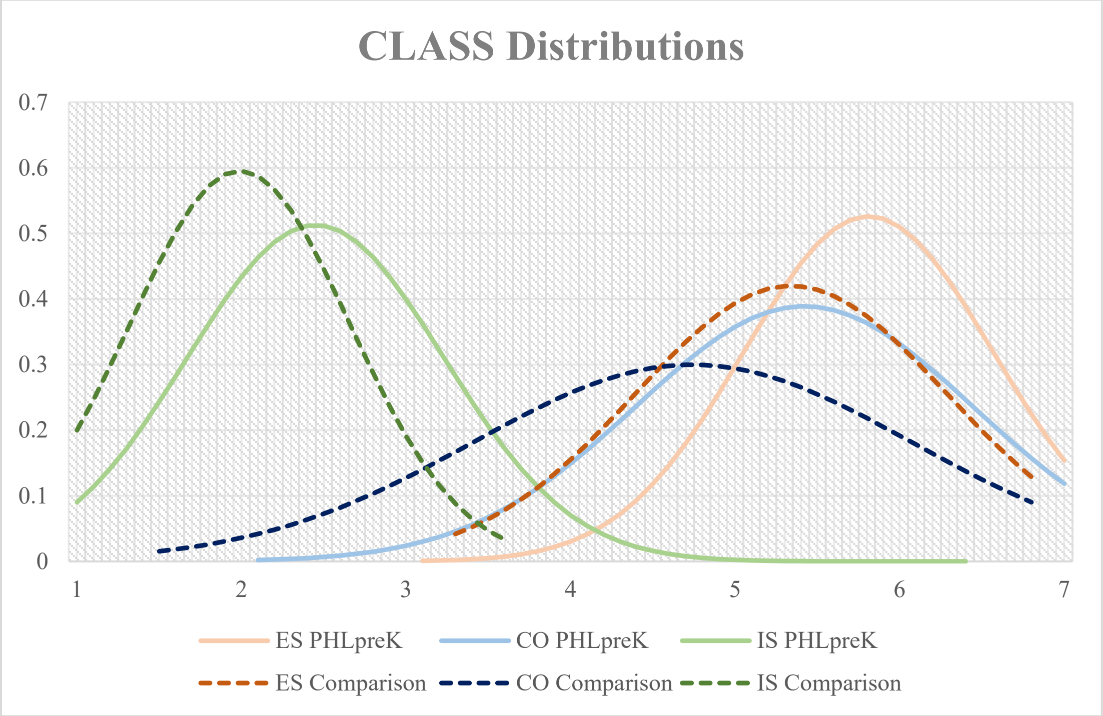
Table 3. CLASS domains & dimension scores for PHLpreK and comparison providers in PHL.

Domains and Dimensions	PHLpreK 2023 (N=283)				Comparison group 2023 (N=24)			
	Mean	(SD)	Min.	Max.	Mean	(SD)	Min.	Max.
Emotional Support *	**5.81**	**(0.76)**	**3.15**	**7.00**	**5.34**	**(0.95)**	**3.35**	**7.00**
1. Positive Climate*	6.07	(0.90)	3.20	7.00	5.54	(1.13)	2.80	7.00
2. Negative Climate*[a]	6.79	(0.44)	3.40	7.00	6.60	(0.40)	5.80	7.00
3. Teacher Sensitivity	5.69	(1.10)	1.20	7.00	5.08	(1.48)	2.00	7.00
4. Regard for Student Perspectives*	4.70	(1.16)	1.00	7.00	4.14	(1.21)	1.40	6.40
Classroom Organization *	**5.42**	**(1.03)**	**2.07**	**7.00**	**4.74**	**(1.33)**	**1.53**	**6.80**
5. Behavior Management*	5.67	(1.09)	1.80	7.00	5.05	(1.32)	2.00	7.00
6. Productivity*	5.66	(1.09)	1.80	7.00	5.03	(1.42)	1.00	6.80
7. Instructional Learning Formats*	4.93	(1.18)	1.20	7.00	4.15	(1.55)	1.00	6.80
Instructional Support*	**2.45**	**(0.78)**	**1.00**	**6.40**	**1.99**	**(0.67)**	**1.00**	**3.53**
8. Concept Development*	2.10	(0.77)	1.00	6.60	1.77	(0.77)	1.00	4.00
9. Quality of Feedback*	2.44	(1.00)	1.00	6.60	2.00	(0.70)	1.00	3.60
10. Language Modeling*	2.80	(0.97)	1.00	6.00	2.20	(0.85)	1.00	3.80

[a]Inversely coded for ease of interpretation.

Figure 5 illustrates the distributions for PHLpreK classrooms (solid line) and comparison sites (dotted line). PHLpreK distributions are further to the right (exhibiting a larger fraction of classrooms at higher quality levels), particularly for IS and CO. All domains of the CLASS were significantly higher for PHLpreK classrooms.

Figure 5. CLASS domain distributions for PHLpreK and comparison classrooms.



Note: For PHLpreK n= 283 for comparison n=24.

2. Children's gains in the PHLpreK program, 2022-2023

This evaluation measured gains in 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). Also included is an evaluation of executive functioning (EF) using the Dimensional Change Card Sort Game (DCCS), for a randomly selected subsample of children in 50 classrooms.

Child gains for the 2022–23 school year for the overall sample and for selected subgroups of interest are shown below and reported in detail in Appendix B. Included in the measured gains are only scores for children assessed in both fall and spring of the school year. Figures 16-18 report gains in standardized scores for the PPVT (receptive vocabulary) and Woodcock-Johnson (literacy and math) assessments which allow comparing results for children in the program in relation to growth due to age maturation (that is, in relation to growth due to children's natural average growth as captured in the norming sample for each measure). These measures are standardized at the mean score of 100 and with a standard deviation of 15. Positive gains in standard scores point to gains that are larger than those of other children, after adjusting for age. It should be noted this last school year, we randomly selected a group of 50 classrooms in the program to assess children's growth, while in previous years we were able to include a much larger sample of children.

The subsample of children assessed this past 2022-23 school year exhibited slight positive gains relative to peers their age (of 0.31) in receptive vocabulary. As compared to the last cohort of children in the program assessed pre-pandemic, children scored slightly higher in the fall on the PPVT and the WJ AP, and slightly lower on the WJ LW. Growth patterns are lower on the PPVT than they were in the past (compared to 4.17 in 2018-19). That is, children started the year with higher average standard scores in 2022-23 (surprisingly, given the pandemic) but demonstrated lower average standard gains than children in the 2018-19 cohort.

In terms of WJ AP scores, average gains were also positive but also lower than they were in the cohort assessed pre-pandemic. Average standard scores on the WJ AP (capturing emerging math) were slightly higher in the fall of 2022-23 than for the 2018-19 cohort, but gains were lower. Average gains on the WJ AP in terms of standard scores were 1.22, compared to 4.37 in 2018-19.

Finally, a positive trend was seen in terms of average growth on the WJ LW (literacy). In 2018-19, standard score gains were 0.56. In 2022-23, gains were 1.17. Put another way, students in the fall of 2018-19 started slightly higher on the WJ LW but ended lower than students in the 2022-23 cohort; growth in this year's cohort in literacy thus exceeded and was just over two times greater than the growth of the prior cohort of children (assessed pre-pandemic).

In terms of other trends (Figures 6-8), we documented that standard score gains in receptive vocabulary were greatest for children who identified as Black and Hispanic, with White and children who identified as other not demonstrating positive standard gains relative to the norm (although the sample sizes for these children were both relatively small). Receptive vocabulary average gains were also greater for 4-year-olds than for 3-year-olds, and females demonstrated larger average gains than males.

Overall, children's average standard scores increased on two of three academic skills measures in relation to the norm, and increased on one of three of these measures in relation to the 2018-19 cohort. On the two academic skills measures in which there were not higher gains (PPVT and WJ AP), variance this year was higher than in the 2018-19 cohort, meaning children showed more differences within the sample in terms of gains.

Other trends observed included:

(1) Black and Hispanic children made larger gains on the WJ LW in 2022-23 than in 2018-19. Although the sample as a whole made larger gains in this cohort on this measure (1.17) as compared to the previous cohort (0.56), this was particularly pronounced for Black children (1.88) whose gains were 0.15 in the previous cohort. This was also the case for Hispanic children although they accounted for a small group in the sample.

(2) Almost all subgroups of children in this cohort had slightly higher gains overall in the DCCS as compared to the previously measured cohort. This was particularly so for children who identified as Black and children with an IEP. Children who identified as White made smaller gains this year relative to the prior cohort.

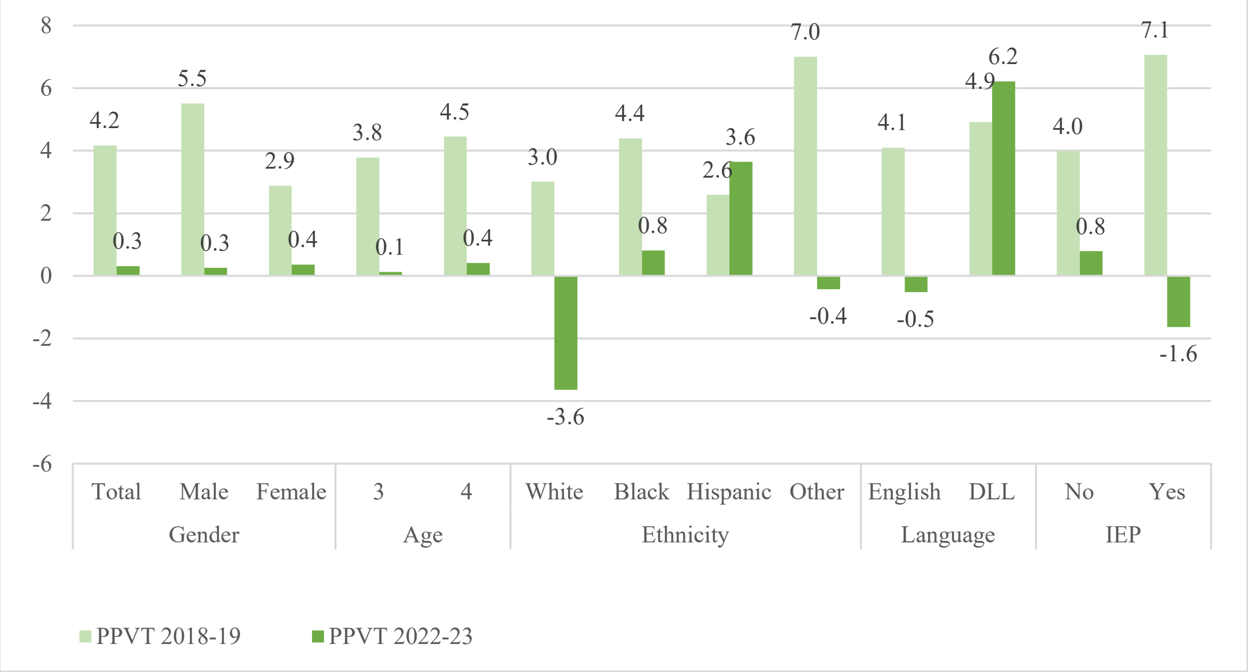
(3) Higher gains are observed in receptive vocabulary for dual-language learners (DLLs) than for children who are native English speakers. DLLs made greater gains on receptive vocabulary this year than they did in the previous cohort.

As comparison, it is useful to assess gains for lower income and minority children in other evaluations of preschool programs. For example, one-year gains for children in this year's sample was of 0.31 standard points on the PPVT, which are smaller gains than those reported for 3 and 4-year-olds in the FACES study of children enrolled in Head Start (Aikens et al., 2013; Aikens, et al., 2017).[4] One-year gains in LW identification were 1.21 standard points, which is

[4] FACES is The Head Start Family and Child Experiences Survey. This is an ongoing national longitudinal study of the cognitive, social, emotional, and physical development of Head Start children. The 2014-15 cohort of FACES

smaller than the gains made in FACES, in which gains were 5.8 in 2014-15 (although for the WJ-III). Lastly, gains in WJ AP (math) were on average 1.17 standard points, which are just slightly smaller than the gains in FACES (for the WJ-III). Similar to PHLpreK children, Head Start children in the FACES study also scored well below average before and after a year in the program (Aikens, et al., 2017).[5]

Figure 6. Standard score gains for children in the PPVT 2018-19 and 2022-23 cohorts.

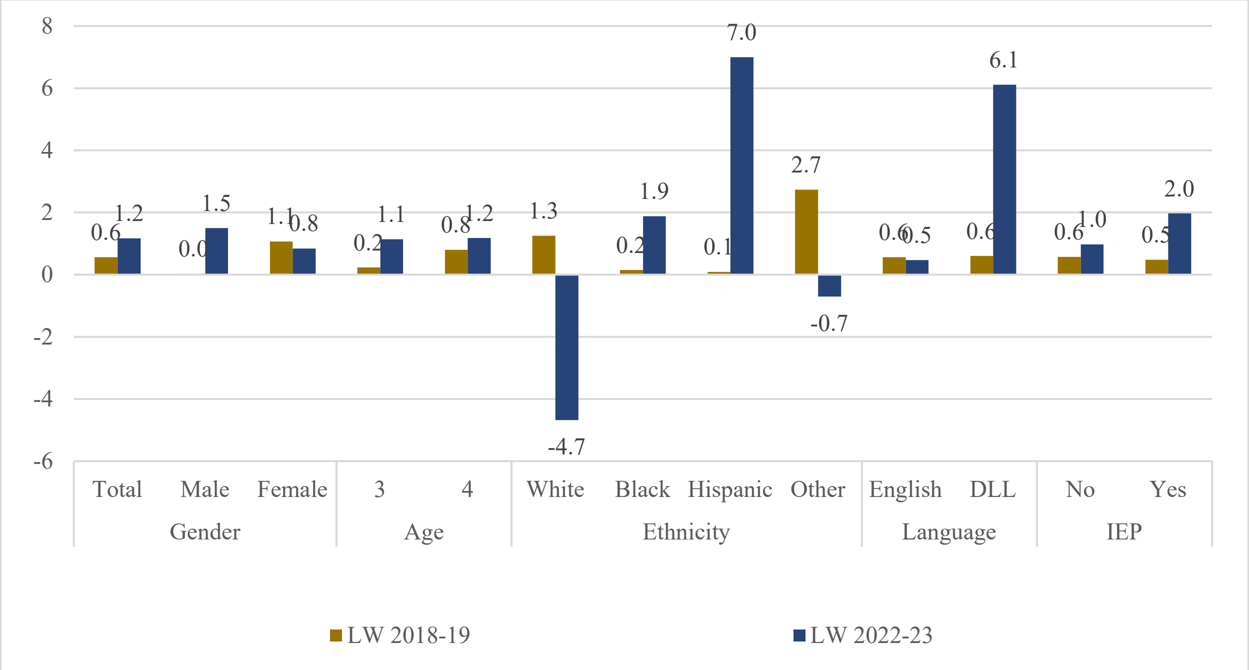


Note: For 2018-19 n= 585 for the PPVT; for 2022-23 n= 153.

evidence lower gains in these three measures than the 2009-10 cohort which. The 2014-15 gained 1.6 in the PPVT, 2.1 in letter word and 2.0 in applied problems. The 2014-15 gains for Head Start children were 3.4 in the PPVT, 5.8 standard points in LW ID and 2.0 in applied problems.

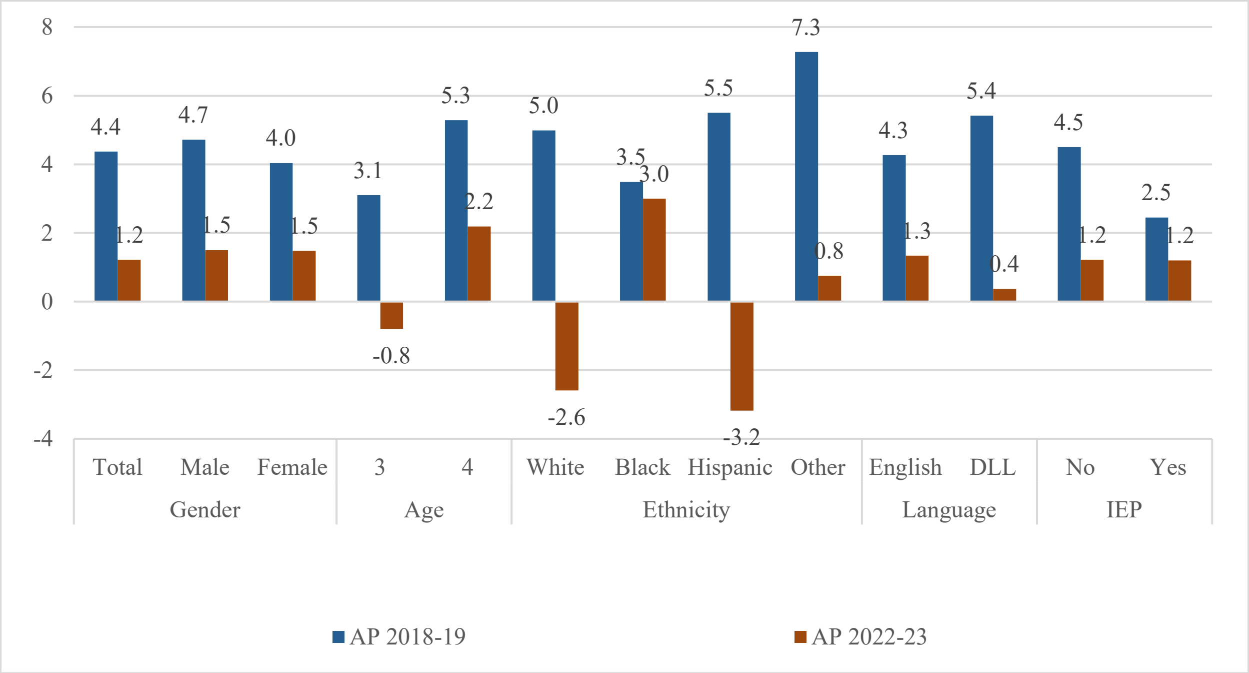
[5] Head Start children in the FACES study were 27% White, 23% African American and 42% Hispanic. A total of 40% spoke a language other than English at home. Almost a third of mothers of children in the 2014 study had less than a high school diploma (26%), another third reported a GED (33%) and another third some college or a vocational/technical degree (33%). Only 8% reported a Bachelor's or higher degree. Over 90% of the sample reported incomes below 200% of the federal poverty threshold. PPVT average fall scores for children in the FACES were 89.9 standard points, which is lower than those in the PHLpreK sample.

Figure 7. Standard score gains for children in the WJ LW identification for the 2018-19 and 2022-23 cohorts.



Note: For 2018-19 n= 585 for the WJ LW; for 2022-23 n= 153.

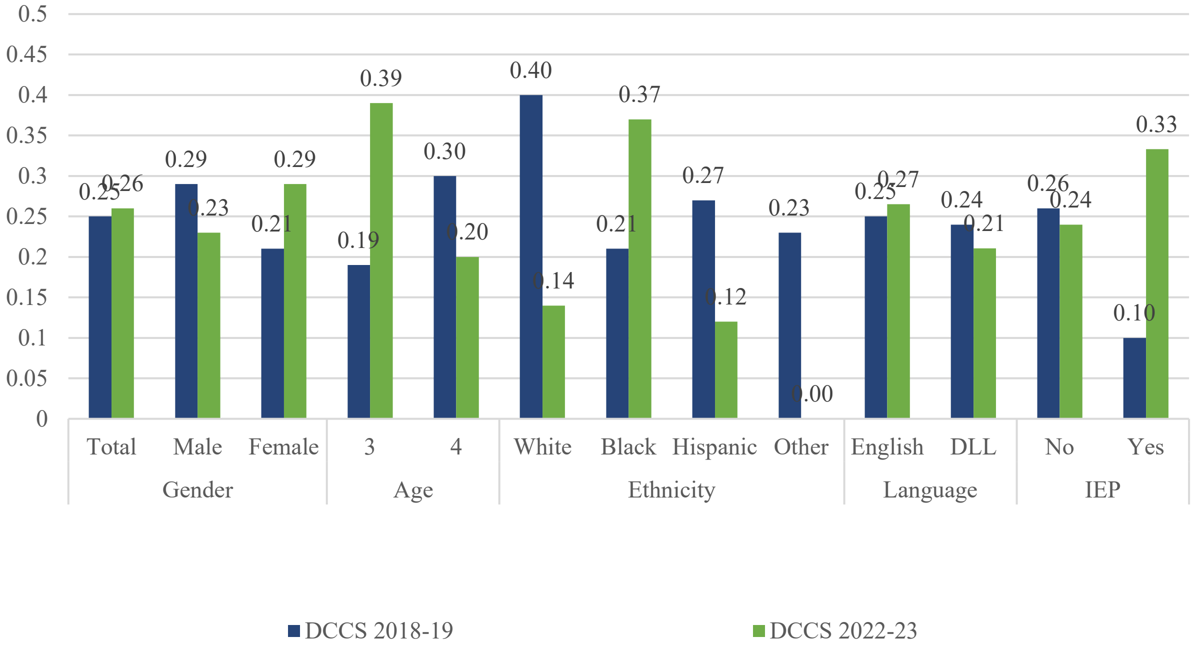
Figure 8. Standard score gains for children in the WJ applied problems for the 2018-19 and 2022-23 cohorts.



Note: For 2018-19 n= 585 for the WJ AP; for 2022-23 n= 153.

Figure 9 shows gains in the DCCS. As reference, the Learning-Related Cognitive Self-Regulation School Readiness Measures for Preschool Children Study (aka the Self-Regulation Measurement Study; Meador, et al., 2013) reports average DCCS scores of 1.42 at 51–53 months of age and 1.62 at 57–59 months. This is an average difference of 0.20 between these two ages. Children gained in executive functions (DCCS) at a higher level than children in the self-regulation study, with overall gains being 0.26, which is just slightly higher than measured gains in the last PHLpreK child cohort measured in 2018-19 (0.25). We recorded stronger gains in children who identified as Black (0.37), and smaller gains in children who identified as White (0.14) and Hispanic (0.12). The Self-Regulation Measurement Study also reports significant differences on the measure between 3-and-4-year-olds; consistent with this, we found gains were greatest for 3-year-olds (0.39), and that they started lower than 4-year-olds in the fall (1.01 as compared to 1.30).

Figure 9. DCCS gains in children for the 2018-19 and 2022-23 cohorts.



Note: For 2018-19 n= 585 for the DCCS; for 2022-23 n= 151.

Finally, we were able to assess a small group of children in comparison classrooms in the City of Philadelphia. The sample of centers that accepted participation was low, despite having invited approximately 1,600 centers and family child care homes to the study. Although the sample was small relative to the PHLpreK sample, understanding the growth and development of a comparison group at the same point in time (i.e., towards the end of the COVID-19 pandemic) may provide helpful as a baseline for future evaluations. In terms of growth over the year, children in PHLpreK classrooms made greater gains than children in the comparison sample on two of the four measures: the PPVT and the WJ LW. Children in the comparison group slightly outperformed PHLpreK children on the DCCS (.35 for the comparison group; .26 for the PHLpreK sample) and the WJ AP (1.58 for the comparison group; 1.22 for the PHLpreK sample). The comparison group started with higher scores on three of the four measures (and had identical scores on the DCCS), yet demonstrated lower growth on two of the measures.

Descriptive analyses of developmental gains for children do not take into account the intersectionality of varied inter-relationships of social identities and interacting social processes that compound in the production of inequities (Bécares & Priest, 2015). Estimations that account for varied socio-demographic identities allow understanding inequities between groups accounting for inter-group differences. Therefore, we next examine the association between children's end of year learning outcomes, their various demographic characteristics, and program features, as well as teacher qualifications when available, using multi-level estimations. We include information on children's start of year outcome on each assessment (fall scores), gender, race and ethnicity, home language, and IEP. Program features for PHLpreK include star ratings, teacher qualifications when available, teacher ethnicity, and classroom quality. The analyses also consider that scores of children who are in classrooms together cannot be considered independent of each other (that is, clustering of children within classrooms, as they experience the same program and teacher). Multivariate analyses account for how children are grouped, their background and their preschool experience. That is, this allows understanding how children's gains differ among children, and what aspects of centers and teaching and learning, contribute to those gains.

We present analyses including the CLASS. Results are shown in appendix C and summarized here. Table C.1. – C.4. shows these for estimations for models that just include children's characteristics, subsequently for models including children and teacher characteristics, and lastly for models include CLASS ES, CO and IS domain scores. Table C.5. includes estimations when the comparison group is included.

Estimations show that children's gains do not generally differ across race and ethnicity on receptive vocabulary and executive functions, with the exception of children who identified as "other" in terms of their race and ethnicity. This was also the case for children with IEPs on math and receptive vocabulary. Although as previously discussed there are some differences in overall gains as a function of child characteristics, these differences are generally not statistically significant when taking into account different child and classroom aspects.

In terms of center and classroom characteristics, estimations show that CLASS CO was positively associated with receptive vocabulary (which is consistent with our findings in 2017-18 and 2018-19) and with math. Finally, we found positive associations between CLASS IS scores and executive function and math scores. Although in past evaluations we have found a positive association between 3- and 4- star rated programs and children's outcomes (i.e., in 2018-19), we do not find that association this year. Of note, the vast majority of PHLpreK sites (253 of 283 observed classrooms) were 4-star programs, so the lack of variability in program ratings likely explains this finding.

The main patterns that emerge from the multivariate estimations are: (a) positive associations between CLASS IS scores and child performance, specifically in math and executive functions; (b) positive associations between CLASS CO scores and children's gains in receptive vocabulary and math; and (c) negative associations between IEP and children's gains in receptive vocabulary and math, such that children with IEPs made significantly lower gains in receptive vocabulary relative to children without an IEP. This differs from the findings of the evaluation in 2018-19 (pre-pandemic), in which African American children showed lower growth relative to their White peers on most measures of academic skills. Such inequities were smaller in the most recent year of data collection (2018-19) than in prior years, and even more so this last year. That is, results show that the program seems to be supporting children's growth more equitably, with the exception of children with IEPs, who appear to need a stronger set of supports across the system. In addition, it seems that strengthening CLASS scores, in particular CLASS CO and CLASS IS, could further support children's gains in the classroom, particularly in math. The low variation in IS scores likely explains the lack of an association with children's development across all measures.

Discussion of Findings

This report summarizes the findings for the 2022-23 school year for Philadelphia's preschool program. The program has concluded its seventh year of operations and continues to grow since its inception through solidifying partnerships with community-based providers across the city. The purpose of this component of the evaluation is to provide information that allows identifying strengths and weaknesses in the program through its expansion period in order to inform professional development and technical assistance efforts. This information also serves to inform continuous improvement strategies to support the program's maturation.

Pre-K classrooms in these programs are averaging high to moderate levels of quality as measured by the CLASS Emotional Support and Classroom Organization domains. Higher quality classrooms in these domains show associations with some of the measured child outcomes and therefore supporting lower quality classrooms in these domains would further support child development. The Instructional Support domain is still low across classrooms, and scores on this domain were lower than last year and continue to evidence a need for strong supports. In summary, classrooms on average are nurturing and safe environments for children and are adequately structured and organized. Areas to strengthen include teachers' use of strategies to scaffold children's learning, incorporating conversational feedback loops that support children's understanding of concepts, increasing conversations to encourage children to use advanced language, questioning that supports the development of analytical thinking skills, linking concepts across activities so that children learn to apply their knowledge to the real world, providing opportunities to engage in problem-solving activities, and planning and production processes that incorporate and build upon children and their initiatives.

Encouragingly, children made greater gains in literacy and executive functions as compared to children in the cohort we last previously assessed in 2018-19. However, child gains in receptive vocabulary were quite small, and these gains were also smaller than those demonstrated by children in the PHLpreK program as assessed in 2018-19, and 2017-18. Gains in math were also much smaller than for the most recently assessed PHLpreK cohort.

The 2022-23 scores in Emotional Support and Classroom Organization demonstrate that providers are building on previous year strengths in terms of developing a warm classroom climate, fostering positive relationships amongst children, and setting and maintaining high behavioral expectations. Supports for teachers on classroom quality should ensure this trend persists in future years, and emphasize efforts to address these two domains in low scoring classrooms. However, a focus on increasing classroom quality on Instructional Support specifically within PHLpreK will require particular focus around strengthening instructional supports (concept development, quality of feedback, language modeling, metacognition), and providing teachers with targeted coaching and supports for doing so across the whole system.

Acknowledgments

We are grateful for the partnership with the City of Philadelphia and PHMC, whose funding and support made this work possible. This partnership is instrumental in supporting the research and facilitating collaboration with program providers throughout the City across research components. Finally, we are thankful for all the programs and teachers that opened their doors for the work to be conducted.

References

Aikens, N., Kopack Klein, A., Knas, E., Hartog, J., Manley, M., Malone, L., Tarullo, L., & Lukashanets, S. (2017). *Child and family outcomes during the Head Start year: FACES 2014–2015 data tables and study design.* OPRE Report 2017-100. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Aikens, N., Kopack Klein, A., 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, Evaluation, Administration for Children, Families, U.S., Department of Health, Human Services

Barnett, W. S. (2008). Preschool education and its lasting effects: Research and policy implications. Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved September 20, 2017, from http://nepc.colorado.edu/files/PB-Barnett-EARLY-ED_FINAL.pdf

Barnett, W. S., & Frede, E. C. (2017). Long-term effects of a system of high-quality universal preschool education in the United States. *Childcare, early education and social inequality: An international perspective*, 152-172.

Barnett, W. S., & Jung, K. (2021). Effects of New Jersey's Abbott preschool program on children's achievement, grade retention, and special education through tenth grade. *Early Childhood Research Quarterly, 56,* 248–259.

Barnett, W. S., & Nores, M. (2015). Investment and productivity arguments for ECCE. Investing against Evidence, 73.

Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J.T., Howes, C. & Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open*, 4(2).

Bécares, L., & Priest, N. (2015). Understanding the influence of race/ethnicity, gender, and class on inequalities in academic and non-academic outcomes among eighth-grade students: Findings from an intersectionality approach. *PLOS One*, 10(10), e0141363.

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., (2009). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. Early Childhood Research Quarterly. 25, 166-176.

Burchinal, M., Vernon-Feagans, L., Vitiello, V., & Greenberg, M. (2014). Thresholds in the Association between child care quality and child outcomes in rural preschool children. *Early Childhood Research Quarterl*y, 29, 41–51. doi:10.1016/j.ecresq.2013.09.004

Ceci, S. J., & Papierno, P. B. (2005). The Rhetoric and Reality of Gap Closing: When the "Have-Nots" Gain but the "Haves" Gain Even More. American Psychologist, 60(2), 149-160.

Classroom Assessment Scoring System 2nd Edition (2022). Teachstone.

Duncan, G, & Murnane, R. (2011). Introduction: The American dream, then and now. In eds. Greg J. Duncan, Richard J. Murnane, (Eds.), *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances* (pp. 3–26). New York: Russell Sage Foundation and Spencer Foundation.

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.

Frede, E. C., Jung, K., Barnett, W. S., Lamy, C. E., & Figueras, A. (2009). The APPLES blossom: Abbott Preschool Program Longitudinal Effects Study (APPLES): Preliminary effects through second grade. National Institute for Early Education Research: New Brunswick, NJ.

Gormley, W. T. (2008). The effects of Oklahoma's pre-K program on Hispanic children. *Social Science Quarterly, 89(4),* 916-936.

Graham, G. (2013, March 11). Tulsa's preschool programs seen as national model. Tulsa World: Boulder, Tulsa. Retrieved from: http://www.tulsaworld.com/news/local/tulsa-s-preschool-programs-seen-as-national-model/article_a932f79f-ec5c-5619-a89a-0ef88d271830.html

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2023). The long-term effects of universal preschool in Boston. *The Quarterly Journal of Economics*, *138*(1), 363-411.

Hamre, B., Hatfield B. E., Pianta R. C., & Jamil F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development, 85,* 1257–1274.

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. doi:10.1016/j.ecresq.2015.09.005

Johnson, A. D., Partika, A., Martin, A., Horm, D., Phillips, D. A., & Tulsa SEED Study Team. (2023). A deeper dive, a wider pool: Preschool benefits sustain to first grade on a broader set of outcomes. *Child Development, 94,* 1298–1318. https://doi.org/10.1111/cdev.13928

Jung, K., Barnett, W. S., Hustedt, J. T., & Francis, J. (2013). Longitudinal effects of the Arkansas Better Chance Program: Findings from first grade through fourth grade. National Institute for Early Education Research: New Brunswick, NJ

Ludwig, J., & Phillips, D. A. (2008). Long-term effects of Head Start on low-income children. Annals of the New York Academy of Sciences, 1136(1), 257-268.

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

New York City Department of Education (2018). *Pre-K program assessments Classroom Assessment Scoring System (CLASS) and Early Childhood Environmental Rating Scale – Revised (ECERS-R) release*. Available at https://infohub.nyced.org/docs/default-source/default-document-library/2017-18--program-assessment-results-summary-update.pdf

Nores, M., Barnett, W.S., & Acevedo, M. (2018) Evaluation of the Philadelphia Prek Program. Year 2 Report. New Brunswick, NJ: National Institute for Early Education. Submitted report.

Nores, M., Barnett, W.S., Jung, K., Joseph, G. & Bachman, L. (2019). Year 4 report: Seattle Preschool Program evaluation. New Brunswick, NJ: National Institute for Early Education Research & Seattle, WA: Cultivate Learning, 66 pp. http://nieer.org/research-report/seattle-pre-k-program-evaluation

Nores, M., Barnett, W.S., Li, Z., Acevedo, M., & C. Whitman (2019). Evaluation of the Philadelphia PreK Program. Year 3 Report. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., Francis, J. & Barnett, W.S. (2017). Evaluation of the Philadelphia Pre-K program. Classroom quality report. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., & Harmeyer, E. (2023). Quality in New Jersey's Abbott Preschool Program: A closer look across the years. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., Harmeyer, E., Li, Z., & Acevedo, M. (2021). Evaluation of the Philadelphia PreK Program. Year 5 Report. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., Harmeyer, E., Li, Z., & Espinosa, C. (2022). Evaluation of the Philadelphia PreK Program. Year 6 Report. New Brunswick, NJ: National Institute for Early Education Research.

Nores, M., Li, Z., & Acevedo, M. (2020). Evaluation of the Philadelphia PreK Program. Year 4 Report. New Brunswick, NJ: National Institute for Early Education Research.

Office of Head Start, U.S. Department of Health and Human Services, Administration for Children and Families. (2015). A National Overview of Grantee CLASS Scores in 2015. Washington, DC.

Peisner-Feinberg, E. S., LaForett, D. R., Schaaf, J. M., Hildebrandt, L. M., Sideris, J., & Pan, Y. (2014). Children's outcomes and program quality in the North Carolina Pre-kindergarten Program: 2012–2013 Statewide evaluation. Frank Porter Graham Child Development Institute: Chapel Hill, NC.

Philadelphia Commission on Universal Pre-kindergarten. (2016). *Final Recommendations Report*. Retrieved September 20, 2017, from http://www.phila.gov/universalprek/Documents/Recommendations%20Report.pdf

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38 (2009), pp. 109-119, 10.3102/0013189X09332374

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*.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). Woodcock-Johnson IV tests of achievement. Riverside Publishing.

Weiland, C. & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84(6), 2112-2130.

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.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature protocols*, 1(1), 297.

Appendix A. Measures

Classroom Observation Measures

Classroom Assessment Scoring System 2.0 (CLASS; Classroom Assessment Scoring System 2nd Edition, 2022)

The Classroom Assessment Scoring System (CLASS) 2.0 is an observational system that assesses classroom practices by measuring the interactions between students and teachers. CLASS measures interactions along ten distinct dimensions, which are grouped into three overarching domains. The Emotional Support (ES) domain is measured by four dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives. The Classroom Organization (CO) domain is measured by three dimensions: Productivity, Behavior Management, and Instructional Learning Formats. The Instructional Support (IS) domain is measured by three dimensions: Concept Development, Quality of Feedback, and Language Modeling. Observations consist of five 20-minute cycles, with 10-minute coding periods between each cycle. Scores (codes) are assigned during various classroom activities and then averaged across all cycles for overall scores in three domains. Each dimension is scored on 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.

Table A.1. 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.

Child Measures

The *Peabody Picture Vocabulary Test—Fourth Edition (PPVT-IV;* Dunn & Dunn, 2007) is an adaptive test comprised of 228-items measuring receptive vocabulary in standard English. The PPVT is predictive of general cognitive abilities and is a direct measure of vocabulary size. That is adaptive means that a portion of the test is used with rules for establishing a floor, below which the child is assumed to know all the answers and a ceiling above which the child is assumed to know none of the answers. It is designed for use with population ages 2.5 and above. The PPVT has shown concurrent validity (e.g., Qi et al., 2006) and the results of these tests are found to be strongly correlated with school success (Blair & Razza, 2007; Early, et al., 2007). This instrument has been used in various preschool studies (e.g., Barnett, et al., 2018; Frede, et al., 2009; Gormley, 2008; Jung et al., 2013; Ludwig & Phillips, 2008; Peisner-Feinberg, et al., 2014; Weiland & Yoshikawa, 2013) and capture large gains for low income, dual-language and non-white children. In the Faces study (Aikens, et al., 2017) Cronbach's alpha reliability for the PPVT-4 was 0.97.

The *Woodcock-Johnson Psycho-Educational Battery—Fourth Edition (WJ- IV;* Woodcock, McGrew, Mather, & Schrank, 2001) includes multiple subtests. Only the *Applied Problems* and *Letter-Word Identification* subtests were used. WJ- IV is normed on a stratified random sample of 6,359 English-speaking subjects in the United States. The WJ is also an adaptive test, used with populations above age 3. Correlations of the WJ with other tests of cognitive ability and achievement are reported to range from 0.60 to 0.70. This measure has been used in numerous large-scale preschool studies (e.g., Early, et al., 2007; Gormley, 2008; Graham, 2013; Peisner-Feinberg, et al., 2014; Weiland & Yoshikawa, 2013; Wong, et al., 2008). In the Faces study (Aikens, et al., 2017) Cronbach's alpha reliability for the WJ-LW III was 0.90 and for the WJ-AP III was 0.88.

The *Dimensional Change Card Sort Task* (DCCS; Zelazo, 2006) is an executive function task requires children to sort a set of cards based on different sorting criteria given by the examiner. The test assesses attention-shifting and short-term memory combined. Scores on the DCCS reflect a pass/fail system on each of three levels of increasing difficulty. Raw scores range between 0 and 3, where a score of 0 means a child did not pass the first level, which includes a color sorting task. In addition, full scores reflect the level of total passes. In the first level, children are tasked with sorting two objects by a color rule, in a second level by a shape rule, and in the advanced level, children are asked to ignore color or shape by adding a border to cards to indicate which attribute to sort by. 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).

Appendix B. Outcomes.

Table B.1. PPVT raw score means and gains by child characteristics

		Valid N	PPVT Raw F22		PPVT Raw S23		PPVT Raw Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		152	58.37	24.76	71.79	25.38	11.58	15.44
Gender	Female	76	58.29	24.38	74.55	22.1	12.67	15.19
	Male	76	58.45	25.25	68.8	28.34	10.49	15.72
Age	3	50	46.57	22.4	64.69	21.85	13.86	15
	4	102	64.98	23.6	75.31	26.34	10.46	15.61
Ethnicity	Black	93	56.08	21.77	69.42	22.77	11.83	15.92
	Hispanic	17	43.2	27.86	57.58	32.52	11.94	14.26
	Other	20	57.42	24.32	75.5	25.09	13.8	18.44
	White	22	82.64	19.3	90.87	20.69	8.23	11.16
Language	DLL	19	41.12	29.55	59.9	29.8	13.84	18.1
	English	133	60.75	23.12	73.41	24.38	11.26	15.08
IEP	Yes	30	49.44	26.02	58.62	29.41	7.87	17.11
	No	122	60.13	24.19	74.96	23.33	12.49	14.94
Comparison								
Total		31	62.49	28.2	69.21	32.14	7.9	16.53
Gender	Female	19	62.39	27.65	72.33	26.31	10.63	15.22
	Male	12	62.63	29.46	65.35	38.67	3.58	18.24
Age	3	11	46.82	18.03	51	27.65	12.55	15.14
	4	20	75.1	28.74	81.09	29.62	5.35	17.07
Ethnicity	Black	13	55.66	24.47	55.78	23.83	9.77	14.37
	Hispanic	6	48.77	22.41	41.14	7.73	0.33	19.97
	Other	1	60.83	20.77	108	-	10	-
	White	11	91.33	27.84	102.5	20.55	9.64	18.16
Language	DLL	4	33.83	10.74	42.6	8.38	12.25	4.79
	English	27	65.01	27.89	73.24	32.53	7.26	17.59
IEP	Yes	6	76.67	38.55	69.63	37.41	8	15.86
	No	25	59.95	26.05	69.1	31.31	7.88	17

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 206 and 175, respectively. For the comparison group, the valid N of F22 and S23 are 74 and 38, respectively.

Table B.2. PPVT standard score means and gains by child characteristics

		Valid N	PPVT SS F18		PPVT SS S19		PPVT SS Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		153	93.47	19.35	95.06	20.22	0.31	14.34
Gender	Female	77	94.97	18.49	97.9	18.71	0.36	14.46
	Male	76	91.97	20.16	91.95	21.44	0.26	14.32
Age	3	51	93.18	20.76	97.08	20.11	0.12	16.43
	4	102	93.64	18.6	94.04	20.28	0.41	13.27
Ethnicity	Black	93	92.27	16.13	94.1	18.39	0.81	13.08
	Hispanic	17	80.32	22.98	84.47	23.1	3.65	13.36
	Other	21	90.5	20.65	92.96	24.13	-0.43	23
	White	22	113.29	14.31	110.65	12.62	-3.64	8.52
Language	DLL	19	76.08	25.36	83.81	20.75	6.21	17.48
	English	134	95.87	17.12	96.59	19.73	-0.52	13.72
IEP	Yes	30	83.18	23.28	84.21	23.88	-1.63	16.08
	No	123	95.51	17.86	97.66	18.4	0.79	13.92
Comparison								
Total		31	98.66	18.09	95.53	23.85	-1.77	12.74
Gender	Female	19	97.32	18.02	98.43	16.79	0.26	11.37
	Male	12	100.63	18.32	91.94	30.63	-5	14.57
Age	3	11	96.42	13.93	89.2	26.73	0.45	13.02
	4	20	100.46	20.84	99.65	21.38	-3	12.75
Ethnicity	Black	13	96.61	16.42	87.67	23.44	-1.15	12.33
	Hispanic	6	86	14.64	78.57	9.36	-6.17	15.61
	Other	1	96.83	10.05	114	-	-2	-
	White	11	115.13	17.86	115.67	16.09	-0.09	12.93
Language	DLL	4	76.67	12.6	81.8	8.41	2.75	4.65
	English	27	100.6	17.25	97.61	24.8	-2.44	13.46
IEP	Yes	6	100.56	24.73	87.38	31.96	-1.17	10.96
	No	25	98.22	17.28	97.7	21.36	-1.92	13.33

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 206 and 175, respectively. For the comparison group, the valid N of F22 and S23 are 74 and 38, respectively.

Table B.3. WJ-LW Raw score means and gains by child characteristics

		Valid N	LWIDNT Raw F22		LWIDNT Raw S23		LWIDNT Raw Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		153	6.50	6.60	10.40	7.50	3.58	5.15
Gender	Female	77	6.10	5.49	9.98	6.08	3.40	4.19
	Male	76	6.91	7.58	10.86	8.79	3.76	6.00
Age	3	50	5.00	6.30	8.41	6.05	2.78	3.64
	4	103	7.34	6.64	11.37	7.95	3.97	5.72
Ethnicity	Black	94	6.61	7.25	10.77	8.77	4.03	5.61
	Hispanic	17	3.75	4.08	8.65	4.57	4.00	3.64
	Other	20	6.37	4.05	10.37	4.47	3.40	2.84
	White	22	8.46	6.60	10.17	5.10	1.50	5.45
Language	DLL	19	4.40	4.10	9.48	4.58	4.53	4.02
	English	134	6.79	6.83	10.53	7.81	3.45	5.29
IEP	Yes	30	6.67	5.94	10.38	8.18	3.93	5.82
	No	123	6.47	6.74	10.41	7.36	3.50	5.00
Comparison								
Total		31	7.15	6.44	9.89	9.38	3.23	5.54
Gender	Female	19	7.73	6.33	12.48	11.53	4.26	6.57
	Male	12	6.30	6.62	6.71	4.19	1.58	2.84
Age	3	11	6.73	8.78	9.6	12.99	3.82	5.58
	4	20	7.49	3.73	10.09	6.36	2.9	5.63
Ethnicity	Black	13	6.63	6.71	9.11	11.88	2.31	5.82
	Hispanic	6	7.15	9.03	7.29	3.09	3.5	2.74
	Other	1	6.67	5.05	12	-	9	-
	White	11	8.87	3.66	12.42	7.73	3.64	6.55
Language	DLL	4	3.50	1.87	5.8	2.17	2	1.41
	English	27	7.47	6.61	10.52	9.9	3.41	5.9
IEP	Yes	6	11.78	10.69	12.5	14.82	3.33	5.24
	No	25	6.42	5.44	9.2	7.54	3.2	5.71

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 204 and 177, respectively. For the comparison group, the valid N of F22 and S23 are 74 and 38, respectively.

Table B.4. WJ-LW standard score means and gains by child characteristics

		Valid N	LWIDNT SS F222		LWIDNT SS S23		LWIDNT SS Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		153	92.74	16.67	94.64	14.8	1.17	12.43
Gender	Female	77	93.63	15.36	95.28	13.92	0.84	9.52
	Male	76	91.83	17.93	93.95	15.75	1.5	14.88
Age	3	50	96.14	18.21	99.22	15.23	1.14	13.6
	4	103	90.85	15.49	92.41	14.12	1.18	11.9
Ethnicity	Black	94	92.96	17.69	95.19	16.35	1.88	12.2
	Hispanic	17	83.5	14.35	91.6	10.32	7	12.78
	Other	20	92.79	10.73	93.33	12.26	-0.7	8.01
	White	22	99.89	14.94	96.04	12.82	-4.68	14.46
Language	DLL	19	85.2	13.81	90.81	11.3	6.11	13.71
	English	134	93.79	16.79	95.16	15.16	0.47	12.14
IEP	Yes	30	89.97	20.43	93.41	17.81	1.97	12.47
	No	123	93.27	15.86	94.94	14.05	0.98	12.47
Comparison								
Total		31	96.82	15.53	94.53	18.92	-1.90	11.67
Gender	Female	19	97.64	15.00	99.29	21.32	-0.53	12.82
	Male	12	95.63	16.46	88.65	13.90	-4.08	9.71
Age	3	11	102.7	17.55	99.87	23.87	0.00	11.83
	4	20	92.1	11.93	91.04	14.40	-2.95	11.75
Ethnicity	Black	13	98.21	14.63	92.78	24.53	-6.23	13.10
	Hispanic	6	93.62	19.02	95.71	7.27	3.33	8.91
	Other	1	94.33	23.64	92.00	-	18.00	-
	White	11	97.53	12.34	96.67	15.29	-1.45	9.08
Language	DLL	4	87.67	6.77	94.00	8.03	0.25	8.62
	English	27	97.63	15.85	94.61	20.14	-2.22	12.16
IEP	Yes	6	98.56	23.49	90.25	26.31	-2.17	12.95
	No	25	96.59	14.45	95.67	16.83	-1.84	11.63

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 204 and 177, respectively. For the comparison group, the valid N of F22 and S23 are 74 and 38, respectively.

Table B.5. WJ-AP raw score means and gains by child characteristics

		Valid N	APPROB Raw F22		APPROB Raw S23		APPROB Raw Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		153	6.75	4.21	9.75	4.39	2.57	3.35
Gender	Female	77	6.88	4.18	10.12	3.89	2.65	3.06
	Male	76	6.62	4.26	9.35	4.88	2.49	3.63
Age	3	50	4.8	3.53	7.86	4.01	2.34	3.47
	4	103	7.85	4.18	10.67	4.29	2.68	3.29
Ethnicity	Black	94	6.09	3.65	9.36	3.79	2.91	3.3
	White	17	5	4.57	7.85	5.25	1.71	3.06
	Other	20	7.33	4.91	10.33	5.23	2.2	3.12
	Hispanic	22	10.64	3.4	12.65	4.18	2.09	3.91
Language	DLL	19	4.76	4.13	8.24	5.67	2.16	2.77
	English	134	7.03	4.16	9.96	4.17	2.63	3.42
IEP	Yes	30	5.47	4.35	8.09	4.97	2.53	2.85
	No	123	7.01	4.15	10.15	4.17	2.58	3.47
Comparison								
Total		31	6.90	4.71	9.13	5.77	2.32	2.64
Gender	Female	19	7.12	4.68	9.00	4.86	2.21	2.68
	Male	12	6.60	4.81	9.29	6.89	2.50	2.68
Age	3	11	3.88	2.96	5.67	4.24	2.55	2.21
	4	20	9.40	4.43	11.39	5.57	2.20	2.89
Ethnicity	Black	13	5.46	3.72	6.00	4.20	2.46	2.60
	White	6	5.38	3.71	6.43	2.76	3.33	2.58
	Other	1	6.33	4.18	12.00	-	-1.00	-
	Hispanic	11	11.87	5.04	15.17	4.41	1.91	2.74
Language	DLL	4	3.83	2.48	5.40	2.61	3.00	2.16
	English	27	7.18	4.77	9.70	5.93	2.22	2.72
IEP	Yes	6	9.56	5.79	9.25	6.94	2.50	3.62
	No	25	6.43	4.42	9.10	5.55	2.28	2.44

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 205 and 177, respectively. For the comparison group, the valid N of F22 and S23 are 73 and 38, respectively.

Table B.6. WJ-AP standard score means and gains by child characteristics

		Valid N	APPROB SS F22		APPROB SS S23		APPROB SS Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		153	83.99	18.34	86.53	18.13	1.22	15.40
Gender	Female	77	86.21	17.75	89.17	15.85	1.48	13.47
	Male	76	81.75	18.74	83.67	20.01	0.95	17.22
Age	3	50	85.99	16.52	88.21	19.11	-0.8	17.73
	4	103	82.86	19.27	85.71	17.66	2.19	14.12
Ethnicity	Black	94	81.79	16.83	85.86	15.77	3.00	15.31
	Hispanic	17	76.54	18.3	76.85	21.09	-3.18	15.58
	Other	20	83.08	21.48	85.88	22.64	0.75	12.39
	White	22	101	12.56	98.83	15.19	-2.59	17.58
Language	DLL	19	72.68	19.37	76.05	22.79	0.37	14.36
	English	134	85.56	17.69	87.94	17.01	1.34	15.59
IEP	Yes	30	74.74	23.25	78.44	22.16	1.20	14.26
	No	123	85.83	16.68	88.45	16.55	1.22	15.72
Comparison								
Total		31	86.22	16.14	86.42	21.35	1.58	12.39
Gender	Female	19	85.88	16.31	86.29	17.47	2.05	13.45
	Male	12	86.7	16.15	86.59	25.93	0.83	11.03
Age	3	11	83.73	15.15	81.2	21.71	1.91	11.44
	4	20	88.28	16.82	89.83	20.87	1.4	13.17
Ethnicity	Black	13	83.89	15.42	76.28	21.19	1.08	13.48
	Hispanic	6	77.31	12.17	80.86	10.38	8.33	14
	Other	1	84.83	7.44	88.00	-	-9.00	-
	White	11	99.67	16.89	104.75	15.02	-0.55	10.04
Language	DLL	4	72.33	9.33	79	12.06	7.75	13.6
	English	27	87.46	16.07	87.55	22.33	0.67	12.21
IEP	Yes	6	85.78	22.53	78.63	26.90	3.83	18.24
	No	25	86.14	15.35	88.5	19.63	1.04	11

Note. The Valid N column shows the valid number of frequencies for the gain scores. For the *PHLpreK* group, the valid N of F22 and S23 are 205 and 177, respectively. For the comparison group, the valid N of F22 and S23 are 73 and 38, respectively.

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

		Valid N	DCCS Final F22		DCCS Final S23		DCCS Final Gain	
			Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
PHLpreK								
Total		151	1.20	0.56	1.48	0.63	0.26	0.66
Gender	Female	76	1.23	0.56	1.58	0.58	0.29	0.67
	Male	75	1.16	0.56	1.38	0.67	0.23	0.65
Age	3	49	1.01	0.48	1.40	0.68	0.39	0.67
	4	102	1.30	0.58	1.52	0.61	0.20	0.65
Ethnicity	Black	92	1.12	0.50	1.50	0.60	0.37	0.61
	Hispanic	17	1.04	0.62	1.30	0.73	0.12	0.70
	Other	20	1.25	0.61	1.38	0.65	0.00	0.73
	White	22	1.61	0.57	1.65	0.65	0.14	0.71
Language	DLL	19	1.04	0.68	1.24	0.70	0.21	0.71
	English	132	1.22	0.54	1.52	0.62	0.27	0.65
IEP	Yes	30	1.06	0.60	1.44	0.66	0.33	0.66
	No	121	1.22	0.55	1.49	0.63	0.24	0.66
Comparison								
Total		31	1.20	0.64	1.58	0.83	0.35	0.75
Gender	Female	19	1.25	0.65	1.67	0.73	0.32	0.67
	Male	12	1.13	0.63	1.47	0.94	0.42	0.90
Age	3	11	0.88	0.60	1.20	0.77	0.45	0.82
	4	20	1.46	0.55	1.83	0.78	0.30	0.73
Ethnicity	Black	13	1.03	0.59	1.17	0.79	0.23	0.83
	Hispanic	6	1.08	0.64	1.29	0.49	0.17	0.75
	Other	1	1.33	0.52	2.00	-	0.00	-
	White	11	1.73	0.59	2.33	0.49	0.64	0.67
Language	DLL	4	1.17	0.75	1.20	0.45	0.00	0.82
	English	27	1.21	0.64	1.64	0.86	0.41	0.75
IEP	Yes	6	1.56	0.73	1.38	1.06	0.17	0.75
	No	25	1.16	0.62	1.63	0.76	0.40	0.76

Note. The Valid N column shows the valid number of frequencies for the gain scores. For *PHLpreK* group, the valid N of F22 and S23 are 204 and 176, respectively. For the comparison group, the valid N of F22 and S23 are 74 and 38, respectively.

Appendix C. Child Estimations.

Table C.1a. Multivariate analyses of children's 2022-23 posttest (spring) PPVT standard score in relation to child and site or classroom characteristics including the CLASS.

	Child	Child + Teacher	Child + CLASS	Child+Teacher + CLASS
Female	1.295	-0.649	-1.669	-1.676
	(3.51)	(3.91)	(3.69)	(3.95)
Black	-0.992	-3.508	1.809	-1.930
	(5.84)	(6.14)	(5.81)	(6.27)
Hisp.	1.215	-0.917	6.776	2.765
	(8.04)	(8.58)	(8.23)	(8.70)
Other Race/Ethn.	-11.749	**-14.314~**	-8.359	**-14.848~**
	(7.52)	**(7.78)**	(7.51)	**(7.85)**
DLL	1.405	-0.779	0.032	0.711
	(6.75)	(7.35)	(6.79)	(7.85)
IEP	**-13.903*****	**-15.088*****	**-14.549****	**-16.667*****
	(4.86)	**(5.62)**	**5.13**	**(5.80)**
Stars 1-3		-2.558		0.276
		(6.01)		(6.40)
LT Asian		**-19.548***		-12.588
		(9.14)		(9.85)
LT Black		3.522		7.252
		(5.87)		(6.13)
LT Missing		2.119		4.174
		(7.99)		(8.99)
LT Master's		8.471		3.003
		(7.96)		(8.59)
LT Associate's		-1.811		-2.338
		(5.63)		(6.06)
LT Bachelor's		-0.959		-2.363
		(5.43)		(6.04)
CLASS ES			-3.192	-4.908
			(3.88)	(4.55)
CLASS CO			**5.713***	4.278
			(2.73)	(3.16)
CLASS IS			2.561	5.498
			(3.89)	(4.39)
N	152	152		152

~p<0.1; * p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, Non-IEP, Star Level 4, Lead Teacher White, Lead Teacher No Degree/Some College. Other controls are pre-test, lead teacher college information missing, and age in months. Standard scores are used. Errors are clustered by site. Fall pre-test scores were imputed from spring post-test scores for children with only one timepoint of data.

Table C.1b. Multivariate analyses of children’s 2022-23 posttest (spring) WJ -LW standard score in relation to child and site or classroom characteristics including the CLASS.

	Child	Child + Teacher	Child + CLASS	Child+Teacher + CLASS
Female	0.394	0.528	0.023	0.262
	(1.75)	(2.00)	(1.82)	(2.03)
Black	4.389	4.791	**5.155~**	5.319
	(3.07)	(3.26)	**(2.98)**	(3.26)
Hisp.	4.903	5.961	6.409	6.700
	(4.87)	(4.96)	(5.01)	(5.11)
Other Race/Ethn.	0.552	2.484	0.976	2.861
	(3.90)	(3.91)	(3.84)	(3.90)
DLL	2.678	1.911	2.883	2.130
	(3.77)	(4.11)	(3.82)	(4.11)
IEP	-0.311	0.170	-0.230	0.385
	(2.26)	(2.76)	(2.40)	(2.76)
Stars 1-3		-0.926		-0.527
		(3.01)		(3.04)
LT Asian		-3.652		-2.677
		(4.32)		(4.70)
LT Black		-3.175		-3.009
		(2.83)		(2.99)
LT Missing		0.370		0.096
		(4.13)		(4.49)
LT Master’s		2.662		3.372
		(4.52)		(4.94)
LT Associate’s		-0.676		0.028
		(2.96)		(3.01)
LT Bachelor’s		0.779		1.416
		(3.01)		(3.16)
CLASS ES			-1.066	-0.490
			(2.21)	(2.57)
CLASS CO			2.282	1.428
			(1.53)	(1.74)
CLASS IS			-0.518	-1.449
			(2.03)	(2.36)
N	152	152		152

~p<0.1; * p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, Non-IEP, Star Level 4, Lead Teacher White, Lead Teacher No Degree/Some College. Other controls are pre-test, lead teacher college information missing, and age in months. Standard scores are used. Errors are clustered by site. Fall pre-test scores were imputed from spring post-test scores for children with only one timepoint of data.

Table C.1c. Multivariate analyses of children's 2022-23 posttest (spring) WJ -AP standard score in relation to child and site or classroom characteristics including the CLASS.

	Child	Child + Teacher	Child + CLASS	Child+Teacher + CLASS
Female	2.555	0.636	-0.699	-0.083
	(3.12)	(3.44)	(3.18)	(3.44)
Black	-2.387	-3.913	0.911	-3.031
	(4.93)	(5.42)	(4.92)	(5.47)
Hisp.	-2.754	-7.138	3.721	-3.156
	(4.93)	(7.36)	(7.00)	(7.39)
Other Race/Ethn.	-7.729	-9.367	-4.935	-10.708
	(6.49)	(6.79)	(6.38)	(6.78)
DLL	-4.588	-4.034	-4.785	-1.877
	(5.85)	(6.31)	(5.78)	(6.39)
IEP	**-8.972***	**-10.424***	**-10.175***	**-11.518***
	(4.25)	**(4.86)**	**(4.38)**	**(4.98)**
Stars 1-3		-4.166		-1.516
		(5.20)		(4.98)
LT Asian		-13.278		-5.840
		(7.98)		(8.54)
LT Black		2.090		6.665
		(5.10)		(5.30)
LT Missing		-2.440		1.089
		(6.98)		(7.81)
LT Master's		10.090		3.696
		(6.96)		(7.43)
LT Associate's		2.557		1.573
		(4.92)		(5.23)
LT Bachelor's		0.066		-1.853
		(4.73)		(5.20)
CLASS ES			-2.339	-4.497
			(3.35)	(3.92)
CLASS CO			**4.044~**	3.364
			(2.37)	(2.75)
CLASS IS			4.933	**7.067~**
			(3.34)	**(3.80)**
N	152	152		152

~p<0.1; * p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, Non-IEP, Star Level 4, Lead Teacher White, Lead Teacher No Degree/Some College. Other controls are pre-test, lead teacher college information missing, and age in months. Standard scores are used. Errors are clustered by site. Fall pre-test scores were imputed from spring post-test scores for children with only one timepoint of data.

Table C.1d. Multivariate analyses of children's 2022-23 posttest (spring) DCCS standard score in relation to child and site or classroom characteristics including the CLASS.

	Child	Child + Teacher	Child + CLASS	Child+Teacher + CLASS
Female	1.012	0.859	0.892	1.016
	(0.93)	(1.08)	(0.97)	(1.08)
Black	0.898	0.289	1.065	0.140
	(1.54)	(1.69)	(1.54)	(1.71)
Hisp.	0.736	0.076	1.736	0.831
	(2.10)	(2.30)	(1.97)	(2.33)
Other Race/Ethn.	-2.182	-3.000	-1.962	**-3.651~**
	(1.97)	(2.13)	(1.97)	**(2.13)**
DLL	-0.635	-0.545	-0.870	0.053
	(1.75)	(1.97)	(1.76)	(2.00)
IEP	-0.390	-1.610	-0.886	-1.600
	(1.27)	(1.51)	(1.33)	(1.54)
Stars 1-3		-1.178		-0.861
		(1.67)		(1.73)
LT Asian		0.027		1.208
		(2.55)		(2.69)
LT Black		2.074		3.246
		(1.63)		(1.67)
LT Missing		0.595		1.910
		(1.63)		(2.46)
LT Master's		**3.928~**		2.175
		(2.24)		(2.36)
LT Associate's		1.394		0.795
		(1.57)		(1.64)
LT Bachelor's		1.576		0.965
		(1.51)		(1.64)
CLASS ES			0.489	-0.372
			(1.09)	(1.24)
CLASS CO			-0.639	-0.321
			(0.77)	(0.86)
CLASS IS			**2.113~**	**2.336~**
			(1.09)	**(1.20)**
N				

~p<0.1; * p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, White, English, Non-IEP, Star Level 4, Lead Teacher White, Lead Teacher No Degree/Some College. Other controls are pre-test, lead teacher college information missing, and age in months. Standard scores are used. Errors are clustered by site. Fall pre-test scores were imputed from spring post-test scores for children with only one timepoint of data.

Table C.3. Multivariate analyses of children's 2022-23 post (spring) standard score in relation to child characteristics, including PHLpreK and comparison group children.

	Receptive Vocabulary	Literacy	Math	DCCS Final
PHLpreK	**13.333**** **(5.06)**	-2.019 (2.478)	**11.441*** **(4.21)**	-0.148 (0.13)
Female	0.349 (3.35)	0.293 (1.94)	-0.055 (2.92)	**0.170~** **(0.09)**
Black	1.221 (5.54)	3.812 (2.80)	0.935 (4.62)	-0.109 (0.13)
Hisp.	3.514 (7.34)	6.925 (4.27)	4.405 (6.21)	-0.253 (0.18)
Other Race/Ethnicity	-5.864 (7.10)	0.165 (3.76)	-1.591 (6.01)	**-0.395*** **(0.18)**
DLL	-2.496 (6.34)	1.216 (3.56)	-6.669 (5.39)	-0.032 (0.16)
IEP	**-11.925*** **(4.74)**	0.681 (2.31)	**-7.637~** **(4.07)**	-0.067 (0.11)
CLASS ES	-1.676 (4.16)	-0.289 (2.16)	-0.904 (3.45)	0.014 (0.10)
CLASS CO	2.193 (2.82)	1.461 (1.44)	1.402 (2.35)	-0.082 (0.07)
CLASS IS	2.099 (4.13)	-0.616 (1.89)	3.668 (3.42)	**0.240*** **(0.10)**

Note: ~p<0.1; * p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are the comparison group, Males, White, English, and Non-IEP. Other controls are pre-test and age in months. Standard scores are used. Errors are clustered by site.