NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

EVALUATION OF THE PHILADELPHIA PREK PROGRAM

Year 4 Report

August 2020

Milagros Nores, PhD, Zijia Li, PhD, Mariel Acevedo, M.Ed. The National Institute for Early Education Research

About the Authors

Milagros Nores, Ph.D. Dr. Nores is a Co-Director of 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.

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.

Mariel Acevedo, **M.Ed.** Acevedo is a Project Coordinator II at the National Institute for Early Education Research (NIEER). She leads NIEER' field work on the PHLpreK Evaluation Study and related work in New Jersey. She has contributed to NIEER's research in Philadelphia, New Jersey, and West Virginia.

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

Suggested citation: 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.

Table of Contents

Introduction

In the school year of 2019-2020, Philadelphia's Preschool Program (PHLpreK) initiated its fourth year of programming. The program has its origins in a May 2015 vote, where city voters approved the creation of the Philadelphia Commission on Universal Pre-kindergarten. The commission was entrusted with proposing a universal pre-K program to provide high-quality, affordable, and accessible services to children in the city, ages 3 and 4. The National Institute for Early Education Research (NIEER) has been conducting a multi-year, multi-site evaluation assessing program components, program quality, and children's learning and development since 2016.

Previous reports have highlighted the importance of high-quality preschool education to reduce persistent achievement gaps in kindergarten and throughout primary (Nores, Francis & Barnett, 2017; Nores et al., 2018 & 2019). Research 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.[1] Consequently, this report represents an evaluation effort to strengthen and support the PHLpreK system through a continuous improvement system that includes understanding the quality of classroom processes, space, and use of time.[2]

This report summarizes the fourth year of the Philadelphia's PreK Program (PHLpreK) evaluation, conducted by the National Institute of Early Education Research (NIEER). The school year of 2019-2020 faced unprecedented challenges for school systems, teachers, children, and families due to the COVID-19 pandemic. In early March 2020, preschools in Philadelphia, much like schools in various parts of the northeast, closed and did not reopen throughout the 2019-2020 school year. Beyond the massive implications that closures had for programs, teachers, families, and children, the COVID-19 pandemic created interruptions for our evaluation of the program.

Accordingly, this report summarizes classroom quality for students in a limited sample of PHLpreK classrooms (observed before the COVID-19 school interruptions took place). It describes the environment and teaching practices in only a sample of classrooms. Also, it does not include information on children's gains in the program since we were not able to assess children in the Spring of 2020. The report does include information on children's performance at school entry (Fall of 2019) and in relation to a comparable group of providers in the city. The latter provides some initial information on the levels present across various domains on this cohort of children at the beginning of the school year, and the degree to which these children were comparable to those in other programs (such as those programs reported in the third-year report). The present report is one of the various components of this evaluation meant to support a data-driven continuous improvement approach to support improvements in the city's program.

Findings evidence PHLpreK classrooms are averaging moderate to high levels of quality in the emotional support and classroom organization domains. These are lower in the subsample assessed the year before the COVID-19 related closures. Instructional support scores are lower as well and, on average, show a consistent need for improvement. We explored quality separately for a few subgroups of interest, including Star level, lead PHLpreK partner agency, and new/old

[1] Ceci & Papierno, 2005; Barnett, 2008; Duncan & Murnane, 2011; Barnett & Nores, 2015; Camilli et al., 2010; Friedman-Krauss et al., 2016; Yoshikawa et al., 2013.

[2] Pianta & Hamre, 2009; Hamre et al. 2014.

sites. Small differences were found between subgroups and are reported. Higher rated classrooms evidenced higher CLASS scores, and classrooms were teachers were stable between the fall and the spring also evidenced higher scores.

In the 2019–20 school year, we assessed children's developmental levels at the beginning of the school year. We report overall levels and how they differ among subgroups of children, and in contrast to other programs in the city. Hispanic, Black, and DLL children and children with an IEP started the school year performing comparatively lower. This is similar to findings from the previous year. Children in the control group, for the most part, evidence minimally lower scores.

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, quality, and impact on children. This report presents limited findings in the fourth year of the study. Data collection included assessing children in the fall of 2019 and classroom observations for a subgroup of programs early in the spring of 2020. This report addresses the following research questions within the limitations imposed by the COVID-19 pandemic:

1. What is the observed quality of children's classroom experiences, and how does it compare to prior years?
2. How did the 2019-20 child cohort perform at preschool entry in vocabulary, literacy, math, executive functions, and social-emotional development? To what extent did these differ by children's background characteristics? How do these compare to children in other programs that are not part of PHLpreK and previous cohorts?

The PHLpreK evaluation was designed to assess the program's trajectory in its early years in terms of quality and children's learning and development. 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 classroom quality observations. Year 3 included a cohort of programs that were not part of PHLpreK to assess quality in other city programs. In Year 4, the plan was to measure child development over the school year, including classroom quality, and to measure child progress and program quality in other city preschool programs. The study had to adapt to the restrictions imposed by the COVID-19 pandemic. Procedures and measures are described in detail below. Children were assessed early in the Fall of 2019, and a limited sample of Classroom observations were conducted before programs were ordered to close in March 2020. The latter assess teacher-child interactions. Classroom observations took place between February and early March 2020. Like previous years, the quality was assessed using a well-known observation protocol during one visit of about two and a half hours.

1. Sample

NIEER assessed 837 children in 159 PHLpreK classrooms (14, which were home-based providers) in the fall of 2019. To recruit children, consent forms were distributed to families as part of the PHLpreK enrollment process. We randomly selected four consented children per classroom. The final sample of children was 55% African American, 18% Hispanic, 13% White, and 13% Asian, mixed-race, or other.[3]

In addition, we assessed children in other programs in the city that did not make part of PHLpreK. These programs were enrolled in the study through a randomized ordered list of programs within the zip codes in which PHLpreK operated.[4] We recruited 8 private programs, 20 Head Start programs, and 6 home-based programs into the study. We then randomly selected four consented children from each classroom in these programs. A total of 257 children in 47 control classrooms were assessed.[5]

Classroom quality was observed using the CLASS Pre-K. The CLASS was used to observe processes in 103 PHLpreK classrooms (center and home-based), just before the COVID-19 school closures of March 2020.

2. Measures and Procedures

Classroom quality was captured using *The Classroom Assessment Scoring System Pre-K* (*CLASS Pre-K*; Pianta, La Paro, & Hamre, 2008). The CLASS measures teacher-child interactions and classroom processes. Home-based providers were observed only using the CLASS. The protocol used required that at least four children were present, and at least half were of preschool age for the observation to be done. Given the smaller size of FCCs, we required at least three children present. More detail on the CLASS is provided in Appendix A.

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 two measures of executive functions, which capture children's inhibitory control, short-term memory, and attention. These are the *Dimensional Change Card Sort Task* (DCCS; Zelazo, 2006) and the *Peg Tapping Test* (PT; Diamond & Taylor, 1996). Socio-Emotional development was measured using the *Caregiver-Teacher Report Form* (C-TRF: Achenbach, 2009). 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 by a CLASS Affiliate Trainer from NIEER, completed the online reliability 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

[3] Comparable to the K-12 PHL school district demographics of 53% African American, 19% Latino, 14% White, and 13% other. https://dashboards.philasd.org/extensions/philadelphia/index.html#/

[4] Programs were ordered into random lists and programs were recruited following this list. We selected programs from three randomized lists within three categories: private, Head Start, and home-based providers.

[5] The target was 600 control children. However, many programs declined participation and/or families in the program did not consent to the study.

background checks and training in human subjects' research (human subject protections, ethical issues, etc.).

Results

Results are presented first for the CLASS (for the limited sample we were able to observe this year). The second section reports children's fall 2019 scores across child and center characteristics and in relation to children enrolled in other non-PHLpreK programs. We conclude with a discussion of the findings.

1. Classroom Observations

CLASS Pre-K Results

Average CLASS scores for PHLpreK classrooms for 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. CLASS scores for 102 classrooms observed before COVID-19 school closures were 5.74 for Emotional Support (ES) 5.26 for Classroom Organization (CO) and 2.30 for Instructional Support (IS).

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

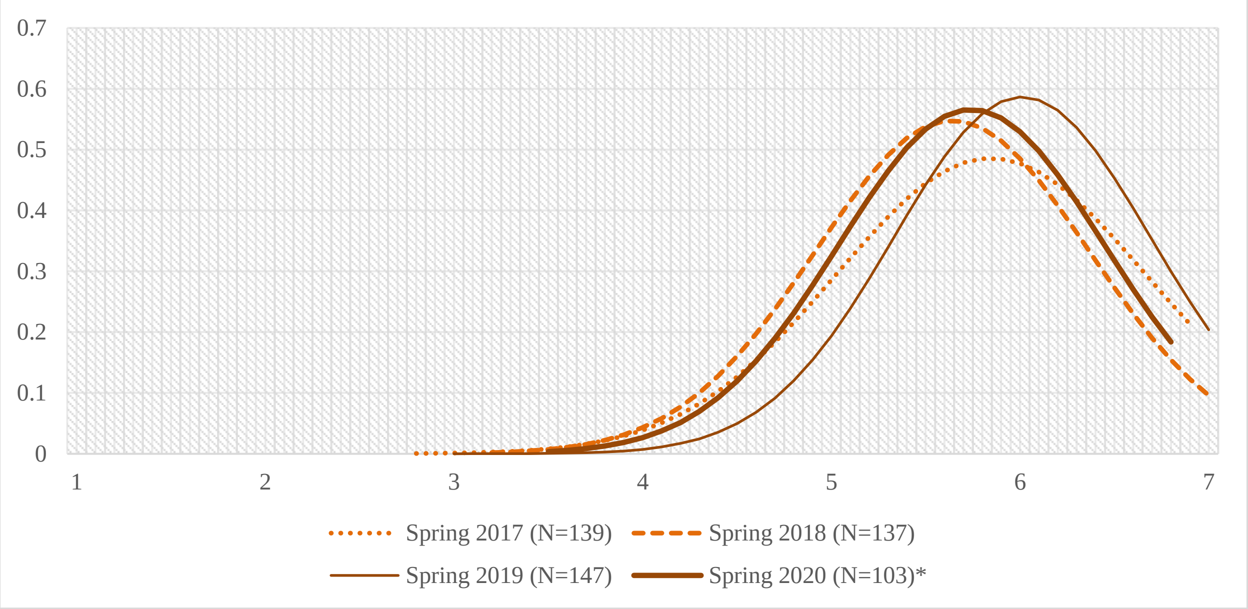
CLASS Dimensions and Domains	**2017 Mean (Range) N=139**	**2018 Mean (Range) N=137**	**2019 Mean (Range) N=147**	**2020 Mean (Range) N=103†**
Emotional Support Domain (ES)	**5.85 (2.85-6.90)**	**5.64[a] (3.20-6.95)**	**6.01[b] (3.05-7.00)**	**5.74[c] (3.55-6.80)**
1. Positive Climate	5.90 (1.60-7.00)	5.73 (3.20-7.00)	6.13 (2.40-7.00)	5.77 (3.20-7.00)
2. Negative Climate*	6.77 (5.00-7.00)	6.67 (4.00-7.00)	6.91 (5.40-7.00)	6.74 (4.2-7)
3. Teacher Sensitivity	5.69 (2.20-7.00)	5.52 (2.80-7.00)	5.89 (1.60-7.00)	5.58 (3.20-7.00)
4. Regard for Student Perspectives	5.03 (2.00-6.80)	4.65 (2.40-7.00)	5.11 (1.60-7.00)	4.88 (2.8-6.8)
Classroom Organization Domain (CO)	**5.34 (1.87-6.93)**	**5.28 (2.80-6.93)**	**5.60[b] (2.40-7.00)**	**5.26[c] (3.20-6.80)**
5. Behavior Management	5.49 (1.60-7.00)	5.48 (2.80-7.00)	5.81 (2.40-7.00)	5.54 (3.00-7.00)
6. Productivity	5.76 (1.80-7.00)	5.65 (2.80-7.00)	5.72 (2.40-7.00)	5.54 (3.40-7.00)
7. Instructional Learning Formats	4.77 (1.60-7.00)	4.72 (1.80-6.80)	5.27 (2.00-7.00)	4.68 (2.40-6.60)
Instructional Support Domain (IS)	**2.41 (1.00-5.00)**	**2.05[a] (1.00-4.60)**	**2.54[b] (1.00-5.33)**	**2.30[c] (1.33-4.13)**
8. Concept Development	2.09 (1.00-4.80)	1.84 (1.00-4.00)	2.27 (1.00-5.60)	2.10 (1.00-4.00)
9. Quality of Feedback	2.23 (1.00-5.00)	1.91 (1.00-4.40)	2.53 (1.00-5.20)	2.10 (1.00-4.20)
10. Language Modeling	2.91 (1.00-5.20)	2.41 (1.00-5.60)	2.80 (1.00-5.80)	2.70 (1.40-4.40)

*The Negative Climate dimension is reverse scored so that a high score represents “good.” [a]Statistically significant difference between 2017 and 2018. [b]Statistically significant difference between 2018 and 2019 distributions of scores. [c] [b]Statistically significant difference between 2019 and 2020 distributions of scores

† Does not include all classrooms in the program. Observation work was interrupted in March 2020 by mandated school closures due to COVID-19.

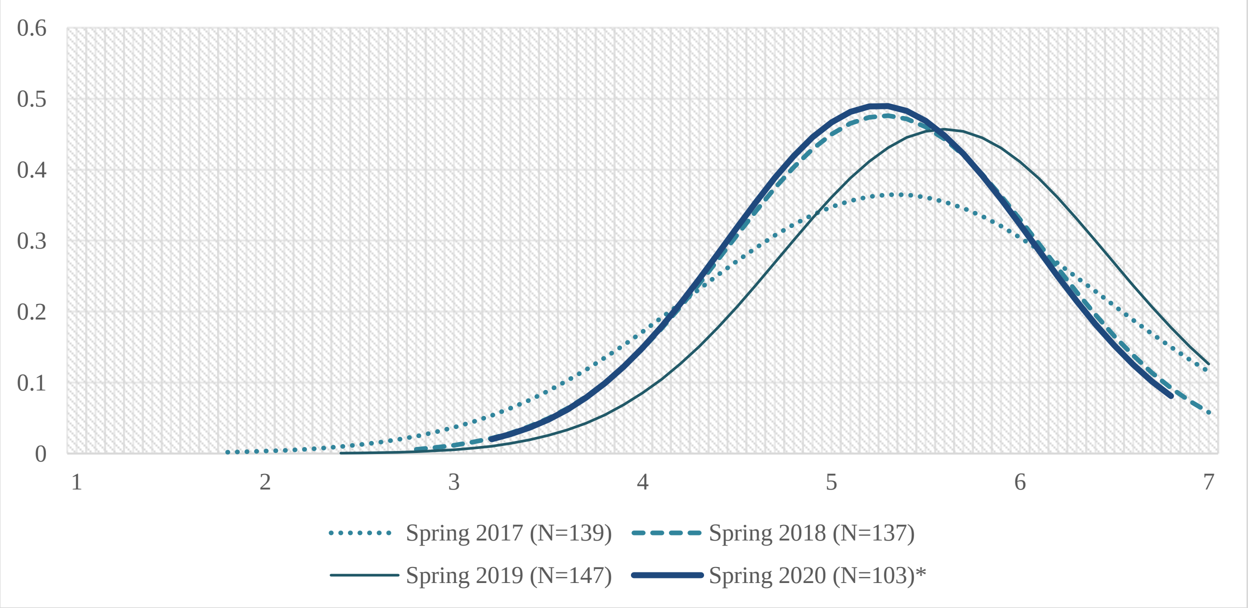
The differences in the distribution of scores on all three domains are observable in Figures 1, 2, and 3. Recent research appears to support thresholds for ES and CO above 5 and IS above 3 as necessary to evidence a relationship between quality and children’s outcomes (other research defines these as slightly higher, at 5.5 and 3.5) (Burchinal et al., 2009; Burchinal et al., 2014; Hatfield et al., 2016). Emotional support scores for this limited sample are lower than the previous year. A 90% of classrooms found to have ES levels above 5 (higher than in 2018 but lower than in 2019). For CLASS CO, average scores are lower as well, and 67% of classrooms having CO scores above 5 (similar to 2018 and lower than in 2019). For CLASS IS, the limited sample shows only 14% above the threshold of 3 in IS (closer to the results in 2018, and lower than in 2019).

Figure 1. Distribution of CLASS Emotional Support scores for 2017, 2018, 2019 & 2020.



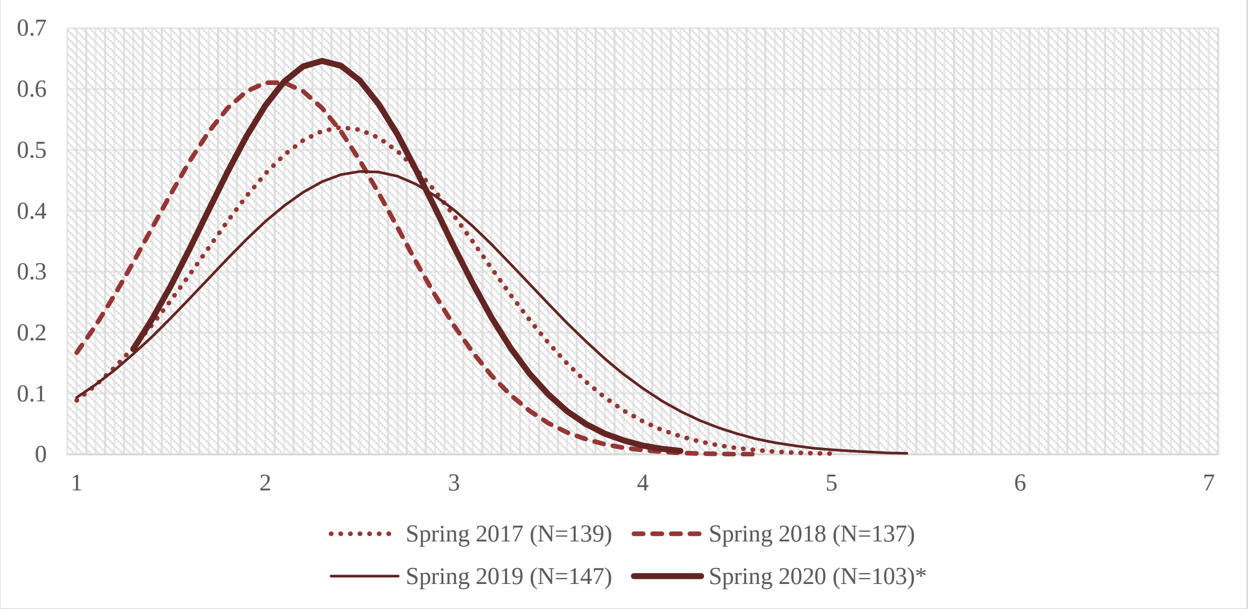
*Smaller sample due to COVID-19 schooling interruptions.

Figure 2. Distribution of CLASS Classroom Organization scores for 2017, 2018, 2019 & 2020.



*Smaller sample due to COVID-19 schooling interruptions.

Figure 3. Distribution of CLASS Instructional Support scores for 2017, 2018, 2019 & 2020.



*Smaller sample due to COVID-19 schooling interruptions.

CLASS Pre-K Domains

The Emotional Support (ES) domain is focused on strengthening supportive relationships between teachers and children, and that help children enjoy the learning process and their comfort in the classroom. The overall mean score for ES of 5.74 is in the high-quality range. The minimum score is 3.55, up from 3.05 the previous year. This indicates there is some compression of the distribution, and classrooms have moderate to high levels of emotional support (above 3). The highest scoring dimension is Negative Climate (6.74), indicating that, on average, classrooms exhibited few negative interactions between teachers and children or among children. The lowest scoring dimension is Regard for Student Perspectives (4.88). These patterns resemble those of previous years. Improving the quality of interactions under this dimension requires teacher flexibility and an emphasis on following children's lead, providing choices to children, and providing ample opportunities for children to express their ideas.

The Classroom Organization (CO) domain is centered on using effective methods to manage instructional time and routines, and behavior expectations. In addition, it includes the provision of activities that maximize children's interests and engagement. The average mean score for the Classroom Organization Domain is 5.26. The scores above 5 denote effective methods to prevent and redirect misbehavior, organized and planned teaching, clarity of instructions, and minimization of time on managerial tasks. The lowest score observed was of 3.20, which was higher than in the previous year, which implies all classrooms scored moderate to good levels. In this domain, the lowest scoring dimension was Instructional Learning Formats scored (4.68). Improving this dimension requires explicitly orienting children towards learning objectives, using effective questioning that expands children's involvement, consistent use of interesting and creative materials, teacher engagement with learning activities, and opportunities that allow children to use different modalities, including hands-on activities.

The Instructional Support domain captures interactions that foster and facilitate higher-order thinking skills, promote language development, and expand children's understanding and learning. While critical for children's learning and development, this domain consistently scores lower across all preschool evaluations and systems. The average score this year is 2.30, ranging from 1.33 to 4.33. Concept Development and Quality of Feedback score lower under this domain (both averaging 2.10). Concept Development refers to facilitating children's individual thought processes. Effective concept development strategies include brainstorming, experimentation, problem-solving, and linking concepts and ideas to children's lives and the real world. Quality of Feedback captures teachers' scaffolding, including back-and-forth exchanges, metacognitive approaches to expand on children's thinking processes, and the use of follow-up questions to enhance information. Consistent and intentional use of language modeling strategies is also critical to improve this domain.

CLASS Pre-K Domains for selected center characteristics

Table 2 reports CLASS domain scores for selected program-level characteristics. Classrooms with higher star levels score higher on all CLASS domains. Concerning partner agency, classrooms in sites in collaboration with PHMC and 1199c evidence higher scores across all domains. Classrooms in sites that started in PHLpreK only this last year perform lower only in CLASS IS. Classrooms where the teacher was replaced between the fall data collection and the spring data collection evidenced lower CLASS domain scores.

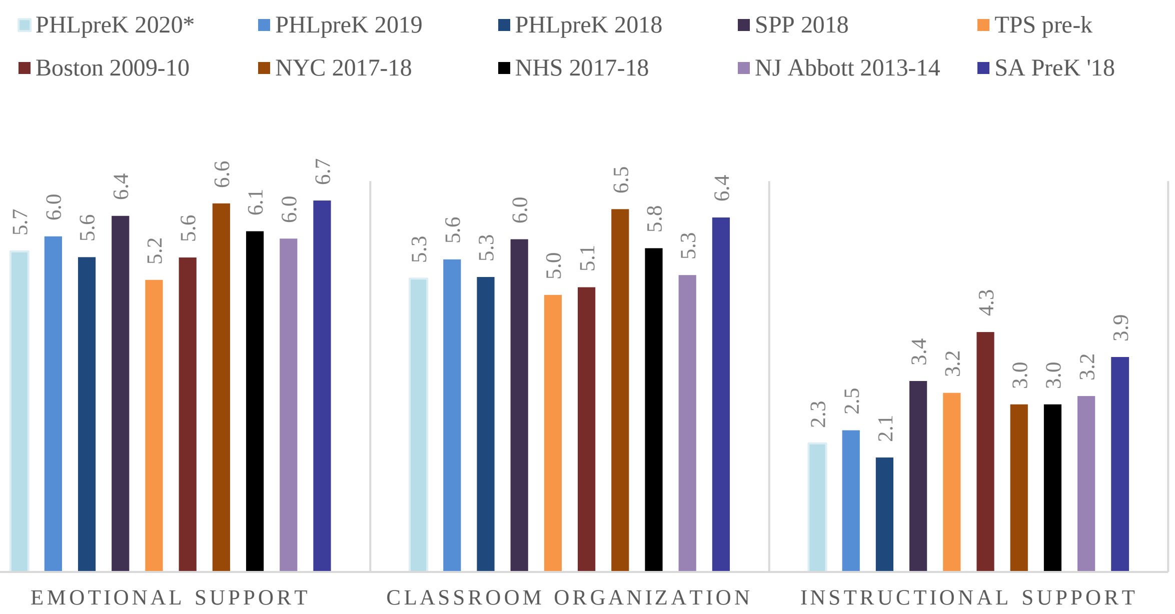
Table 2. CLASS domains scores by subgroups, N = 103.

		CLASS Mean Scores		
		Emotional Support	**Classroom Organization**	**Instructional Support**
STAR Level	< or equal to 3 (n=54)	5.66	5.09	2.21
	4 (n=49)	5.83	5.44	2.40
PHLpreK Partner Agency	UAC (n=26)	5.64	5.08	2.33
	PHMC (n=71)	5.76	5.29	2.29
	1199c (n=2)	6.23	6.30	2.67
	SDP (n=4)	5.75	5.23	2.10
New Site	Yes (n=42)	5.81	5.32	2.24
	No (n=61)	5.69	5.17	2.31
Lead Teacher replaced between fall and spring	Yes (n=16)	5.56	4.75	2.06
	No (n=87)	5.78	5.35	2.35

CLASS Pre-K comparison to other programs

Patterns for the CLASS scores for PHLpreK over the year in relation to those of other cities and states are reported. Figure 4 illustrates the PHLpreK CLASS scores from 2020, 2019, 2018 and 2017 by domain in relation to CLASS scores from various other U.S. preschool programs. This figure includes high-quality city programs. It is worth highlighting that CLASS IS scores for the PHLpreK program lag those of other programs illustrated.

Figure 4. CLASS comparisons across years



*Smaller sample due to COVID-19 schooling interruptions.

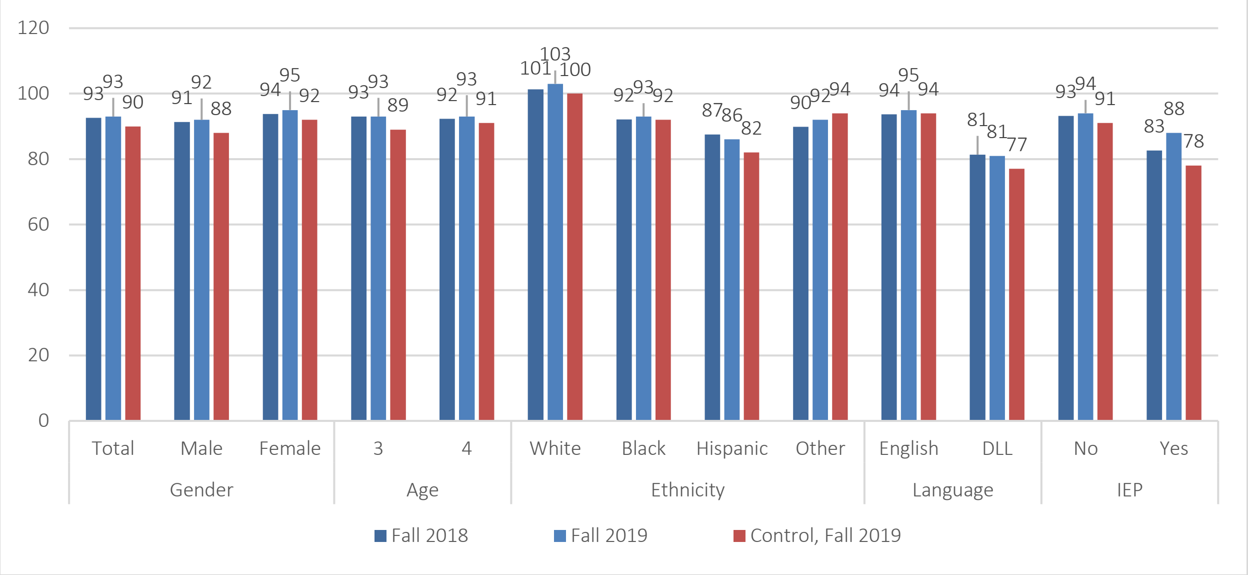
2. Children's fall scores in PHLpreK, Fall of 2019

This evaluation measured fall child levels 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). Moreover, it evaluated executive functioning (EF) using two measures: the Dimensional Change Card Sort Game (DCCS) and the Peg Tapping task (PT). Socio-emotional development was measured with the ASEBA teacher reported form (C-TRF).

Child fall scores for the 2019-20 entering child cohort for the overall sample, for selected subgroups of interest and children in other programs not in PHLpreK are shown below. Figure 4 reports Fall 2019 scores in comparison to Fall 2018 scores and in comparison to a control group of children recruited for the study in the Fall of 2019. These are reported in detail in Appendix B. Figures 5-7 report Fall scores in standardized scores for the PPVT (vocabulary) and Woodcock-Johnson (literacy and math) assessments which allow comparing the cohort of children in the program in this school year in relation to average children their age These measures are standardized at the mean score of 100 and with a standard deviation of 15. Standard scores under 100 points signify scores below average for children of this age.

For the overall sample, the Fall 2019 PHLpreK group performed similarly to the cohort of children enrolled in the previous year at school entry on the PPVT. Control group children scored slightly lower, though (See Figure 16). This pattern remained the same for children ages 3 and 4 children, White, Black, Hispanic, DLL, and native English speaking children. Control group children with an IEP scored the lowest.

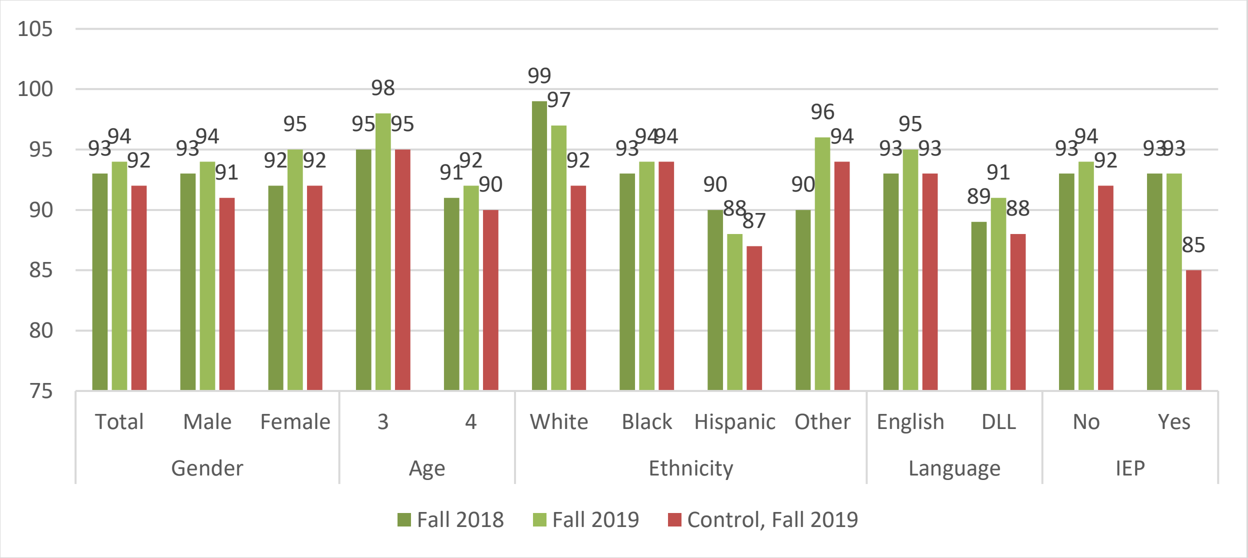
Figure 5. PPVT Standard scores for children, fall of 2018 and 2019.



Note: For Fall 2018 n=585 for the PPVT; for Fall 2019 PHLpreK group n= 572 and Control group n=257.

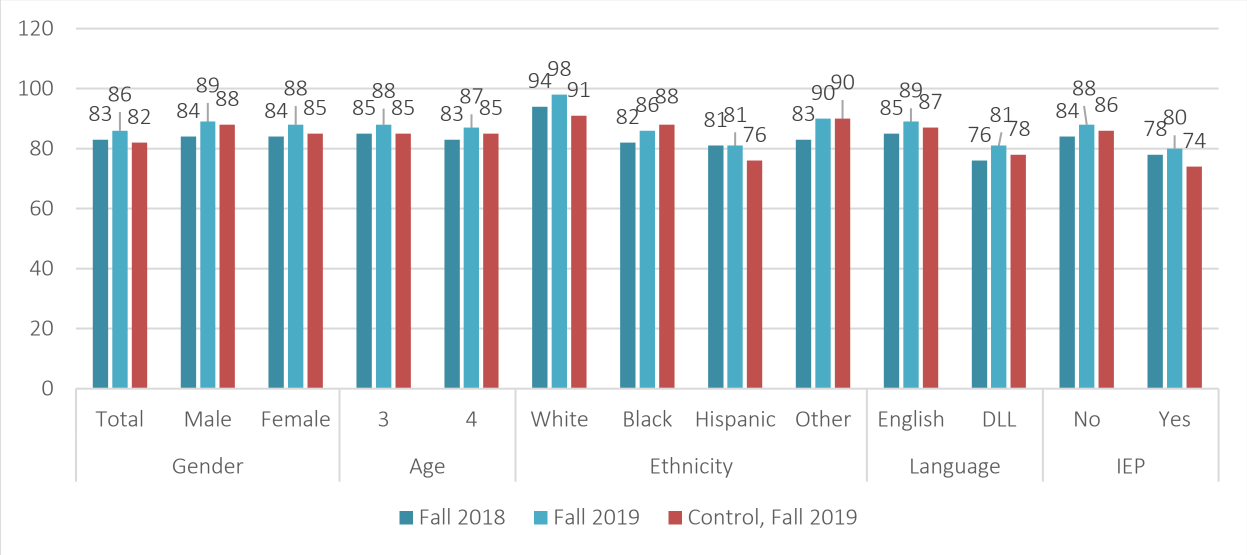
Similar patterns emerge for Letter-word (LW) and Applied Problems (AP) and most subpopulations of interest. Important patterns are lower fall scores for Hispanics and Whites, and for DLL and IEP children relative to their PHLpreK counterparts.

Figure 6. LW Standard scores for children, fall of 2018 and 2019.



Note: For Fall 2018 n=585 for the letter word identification subset; for Fall 2019 PHLpreK group n= 572 and Control group n=257.

Figure 7. AP Standard scores for children, fall of 2018 and 2019.

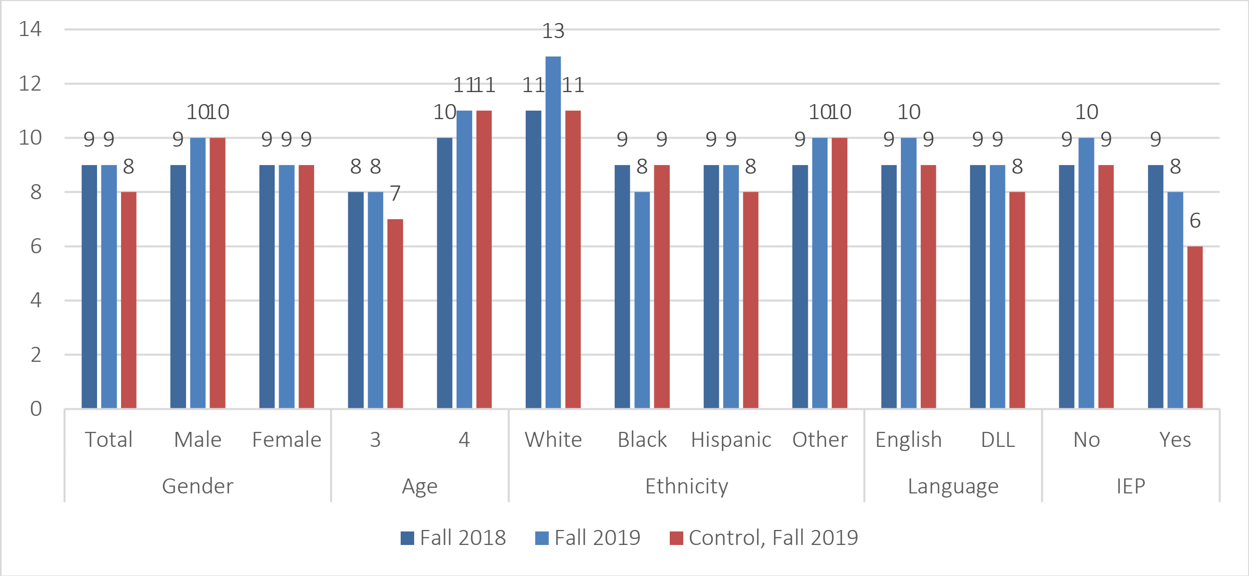


Note: For Fall 2018 n=585 for the applied problem subset; for Fall 2019 PHLpreK group n= 565 and Control group n=252.

Figures 8-11 show fall scores in the DCCS and Peg Tapping (executive function) and the C-TRF (socio-emotional). 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) reports average PT scores of 6.02 at 51–53 months and 8.80 at 57–59 months.

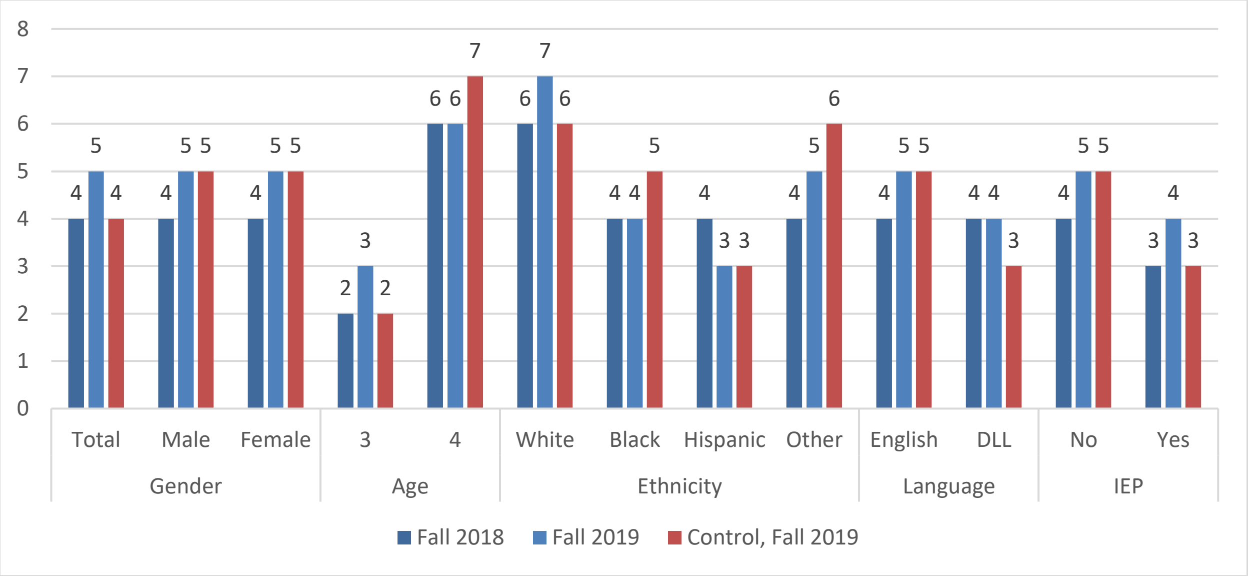
The Fall 2019 control group show the lowest scores on the DCCS. Differences are not large between groups within groups. Overall patterns are: White children start in the fall with higher scores, and control group children tend to have lower scores for any subgroup. These patterns are also found for PT scores. Younger children expectedly show lower scores. Children with an IEP also evidence lower average scores.

Figure 8. DCCS scores in children, fall of 2018-19 and 2019-20.



Note: For Fall 2018 n=585 for DCCS; for Fall 2019 PHLpreK group n= 572 and Control group n=257.

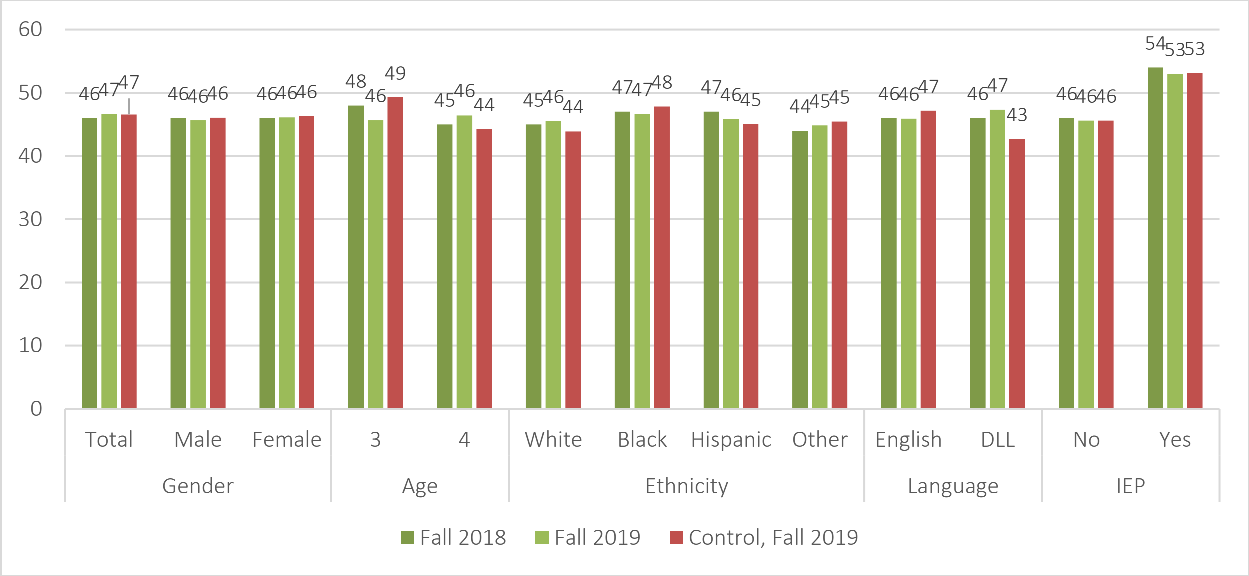
Figure 9. Peg-Tapping scores in children, fall of 2018-19 and 2019-20



Note: For Fall 2018 n=585 for Peg-Tapping; for Fall 2019 PHLpreK group n= 572 and Control group n=257.

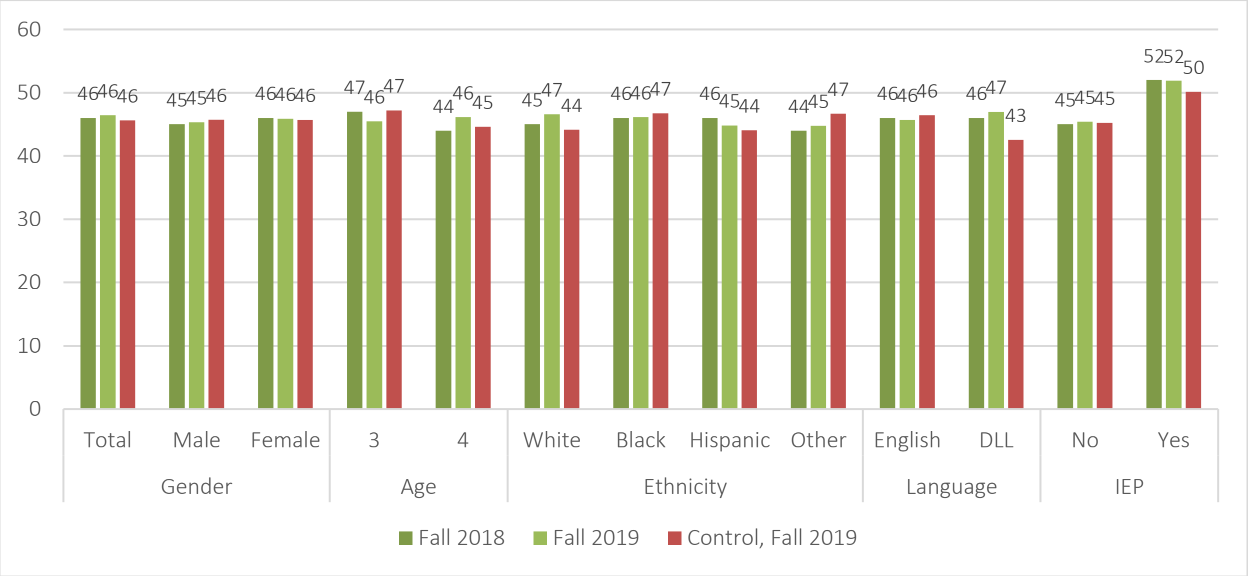
In relation to children's socio-emotional development, the differences between the three groups are small. T scores reported reflect how a child's score on each scale compares with the scores of the normative sample of peers. The C-TRF is inversely coded, with higher scores indicating higher levels of problem behaviors. As shown in Figure 10, children with IEP scored highest in relation to general education peers. This pattern remains the same for both internalizing and externalizing problems (Figures 11 and 12). Other differences across groups and within groups between treatment and control children are minimal.

Figure 10. C-TRF scores in children (socio-emotional problems), fall of 2018-19 and 2019-20



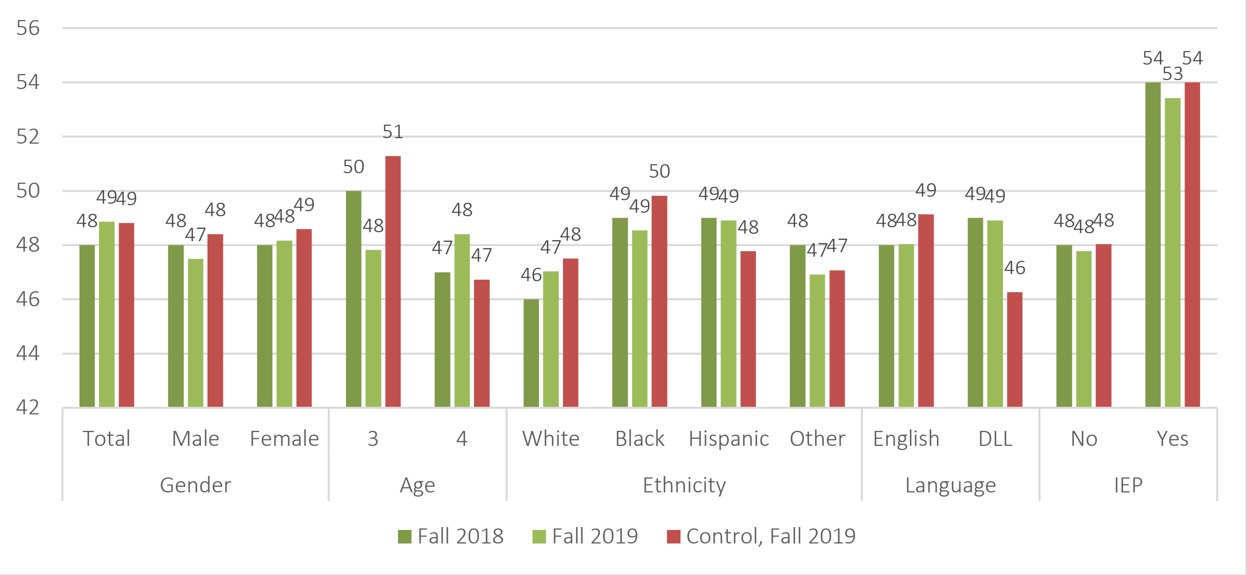
Note: For Fall 2018 n=542 for C-TRF; for Fall 2019 PHLpreK group n= 536 and Control group n=202.

Figure 11. C-TRF scores in internalizing behaviors, fall of 2018-19 and 2019-20.



Note: For Fall 2018 n=542 for C-TRF internalizing problem domain; for Fall 2019 PHLpreK group n= 536 and Control group n=202.

Figure 12. C-TRF scores in externalizing behaviors, fall of 2018-19 and 2019-20.



Note: For Fall 2018 n=542 for C-TRF externalizing problem domain; for Fall 2019 PHLpreK group n= 536 and Control group n=202.

Discussion of Findings

This report summarizes findings for the 2019-20 school year for Philadelphia's preschool evaluation. The PHLpreK program continued to grow since its inception through private-public partnerships across the city. This report provides information on program quality and children's differences at the beginning of the school year.

Pre-K classrooms in the PHLpreK program averaged moderate to high levels of quality as measured by the CLASS Emotional Support and Classroom Organization domains. The Instructional Support domain continues to score low. Scores for the sample of programs we were able to assess before COVID-19 related school closures are lower than average scores the year before. This could be due to the large expansion in the number of programs included as part of PHLpreK this last academic year.

In essence, classrooms are on average nurturing and safe environments for children and adequately structured and organized. This last year's lower average scores imply that it is important to continue to address all areas of quality with teachers and providers. Similar to previous years, some areas that require attention are: 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's 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, opportunities to engage in problem-solving activities, and planning and production processes that incorporate and build upon children and their initiatives.

We also assessed children's fall scores at the beginning of the school and reported these different subgroups of children. Findings show no real differences by gender. Children in the 2018-19 cohort started the school year at a slightly higher average level than children in the 2017-18 cohort. Some small differences emerged between groups. In particular, we find lower scores in vocabulary, language and literacy, and math for Hispanic, DLL, and children with an IEP. Control children appear to have started the year with lower scores in the measures included, although differences are, for the most part, quite small. Younger children evidence lower levels of executive functions as expected, and teachers reported higher levels of behaviors for children with an IEP.

The 2019-20 lower scores in quality, as measured by the CLASS, needs to be incorporated into future strategies to grow and increase quality in the program, in order for the program to have long-term impacts on children´s development. Strengthened and growing supports for teachers on classroom quality that also consider the expansion of the program would ensure future positive trends in quality. Consistency in increasing classroom quality year-to-year requires a critical emphasis on the technical assistance and professional development needs of programs, particularly on strengthening instructional supports (concept development, quality of feedback, language modeling, metacognition) needs to be central in any future program supports. Findings show that classrooms where teachers were stable between the fall and the spring had, on average, higher CLASS scores. This could reflect organizational strength in those sites, and/or reflect the importance of teacher stability for the program to realize benefits from any technical assistance, professional development or coaching efforts.

Acknowledgments

We are grateful to the William Penn Foundation who funded the work of NIEER that made this research possible (grant No. 16-17). We are also grateful for the partnership with the City of Philadelphia and PHMC. This partnership is instrumental in supporting the research and facilitate 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

Achenbach, T. M. (2009). The Achenbach System of Empirically Based Assessment (ASEBA): Development, Findings, Theory, and Applications. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Aikens, N., A. Kopack Klein, E. Knas, J. Hartog, M. Manley, L. Malone, L. Tarullo, and S. Lukashanets. (2017) Child and Family Outcomes During the Head Start Year: FACES 2014–2015 Data Tables and Study Design. OPRE Report 2017-100. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and 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., & 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).

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

Camilli, G., Vargas, S., Ryan, S., & Barnett, W.S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. Teachers College Record, 112, 579-620.

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.

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.

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.

Friedman-Krauss, A., Barnett, S., & Nores, M. (2016). How much can high-quality universal pre-K reduce achievement gaps? Washington, DC: Center for American Progress and National Institute for Early Education Research. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2016/04/01115656/NIEER-AchievementGaps-report.pdf

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

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

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

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., 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. Submitted report.

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

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). CLASS: Classroom assessment scoring system manual preschool (Pre-K) version.

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.

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.

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

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 (CLASS; Pianta, La Paro, & Hamre, 2008; Pianta & Hamre, 2009; Hamre et al., 2014)

The Classroom Assessment Scoring System (CLASS) 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, Kaiser, Milan, & Hancock, 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, Cook, Barnett & Jung, 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).

The *Peg Tapping Test* (PT; Diamond & Taylor, 1996) requires children to follow directions to tap a peg twice when the experimenter taps once and vice versa. It requires children to inhibit a natural tendency to mimic the experimenter while remembering the rule for the correct response, tapping into inhibitory control, attention, and short-term memory. Sixteen trials are conducted with eight one-tap and eight two-tap trials in a random sequence. The final score for Peg Tapping is a sum of all the 16 items that comprise the test. While there are no standard score equivalents, in a study of test-retest reliability, means by age for children age 48 months or younger was 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).

The *Caregiver-Teacher Report Form* (C-TRF: Achenbach, 2009) ages 1½–5 is a short questionnaire for obtaining teachers' reports of their child's competencies and problems. It is normed based on 1,192 children. It has also been tested in 14 societies with 9,389 children. Teachers were instructed to rate the child's behavior early in the fall and again late in the spring. It consists of a 99-item list of behaviors to which the teacher gives a response of 0, 1, or 2 (not true, somewhat true, or very true). Scores included in this report are for total behavior problems.

Appendix B. Outcomes.

Table B.3. PPVT score means by child characteristics, Fall 2019

			PPVT Raw Score			PPVT Standard Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Missing	1	32.00	.	63.00	.
		Male	280	58.54	23.98	92.17	17.12
		Female	292	60.60	23.68	94.65	17.07
	Age	3	235	48.42	18.95	93.27	15.11
		4	338	67.28	23.85	93.46	18.47
	Ethnicity	White	81	74.70	25.83	103.46	17.51
		Black	348	57.32	21.71	92.64	15.05
		Hispanic	69	52.99	22.15	86.32	18.11
		Other	75	59.51	26.38	92.45	20.49
	Language	English	490	61.67	23.56	95.45	16.00
		DLL	83	47.00	21.61	81.16	18.71
	IEP	No	534	59.76	23.83	93.75	17.10
		Yes	39	56.54	24.03	88.31	17.41
Control	Gender	Male	121	51.64	24.06	88.37	17.33
		Female	136	58.75	22.33	91.57	15.24
	Age	3	107	43.50	18.53	88.99	15.52
		4	150	63.90	22.82	90.83	16.85
	Ethnicity	Missing	2	57.00	0.00	90.50	7.78
		White	31	67.48	21.65	100.19	13.63
		Black	108	57.23	22.75	92.45	14.49
		Hispanic	82	44.93	20.83	81.52	15.93
		Other	34	63.76	24.12	93.82	16.66
	Language	English	201	59.79	22.51	93.80	14.29
		DLL	56	39.66	19.53	76.68	16.17
	IEP	No	235	56.86	22.78	91.17	15.82
		Yes	22	39.82	24.69	78.32	17.14

Table B.4. WJ-LW score means by child characteristics, Fall 2019

			LW Raw Score			LW Standard Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Missing	1	10.00	.	93.00	.
		Male	281	6.88	6.82	94.11	15.73
		Female	292	6.49	4.97	94.62	13.70
	Age	3	235	5.20	4.70	98.43	13.42
		4	339	7.71	6.48	91.57	14.92
	Ethnicity	White	81	8.44	8.34	97.37	14.76
		Black	349	6.44	5.47	94.47	13.96
		Hispanic	69	5.06	3.66	88.12	15.12
		Other	75	7.41	6.23	96.43	16.19
	Language	English	491	6.76	5.91	94.99	14.17
		DLL	83	6.27	6.14	90.73	17.18
	IEP	No	535	6.64	5.92	94.45	14.53
		Yes	39	7.38	6.26	93.28	17.16
Control	Gender	Male	121	5.24	4.54	90.90	13.89
		Female	136	6.18	4.07	92.26	13.11
	Age	3	107	4.16	3.77	94.59	13.61
		4	150	6.86	4.34	89.51	13.01
	Ethnicity	Missing	2	7.50	3.54	100.00	1.41
		White	31	5.26	3.61	91.87	10.91
		Black	108	6.40	4.73	94.25	13.50
		Hispanic	82	4.52	3.93	86.68	14.14
		Other	34	6.88	3.85	94.47	11.08
	Language	English	201	5.99	4.36	92.64	12.85
		DLL	56	4.84	4.04	87.96	15.08
	IEP	No	235	5.97	4.34	92.25	13.43
		Yes	22	3.27	3.12	84.95	12.32

Table B.5. WJ-AP score means by child characteristics, Fall 2019

			Applied Problems Raw Score			Applied Problems Standard Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Missing	1	4	.	60	.
		Male	281	7.14	4.35	86.05	17.10
		Female	292	7.63	4.14	89.06	15.31
	Age	3	235	5.43	3.50	88.17	15.94
		4	339	8.73	4.20	87.09	16.55
	Ethnicity	White	81	10.33	3.61	98.01	11.46
		Black	349	6.73	4.06	85.82	15.69
		Hispanic	69	6.35	3.91	81.39	18.13
		Other	75	8.16	4.54	89.59	16.86
	Language	English	491	7.56	4.22	88.54	15.72
		DLL	83	6.33	4.28	81.38	18.38
	IEP	No	535	7.43	4.22	88.07	15.93
		Yes	39	6.77	4.62	80.38	19.48
Control	Gender	Male	121	5.86	4.28	82.14	17.61
		Female	136	7.68	4.14	87.57	15.62
	Age	3	107	4.62	3.41	84.50	16.48
		4	150	8.39	4.17	85.41	17.02
	Ethnicity	Missing	2	6.50	0.71	85.00	14.14
		White	31	8.10	4.02	90.65	14.96
		Black	108	7.30	4.13	88.21	14.99
		Hispanic	82	4.95	4.23	76.13	18.42
		Other	34	8.68	3.82	90.32	11.73
	Language	English	201	7.22	4.13	86.95	15.14
		DLL	56	5.39	4.60	77.81	20.42
	IEP	No	235	7.04	4.28	86.06	16.36
		Yes	22	4.50	3.75	74.23	17.54

Table B.6. DCCS score means by child characteristics, Fall 2019

			DCCS Raw Score			DCCS Metric Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Missing	1	6.00	.	1.00	.
		Male	281	9.25	5.45	1.22	0.60
		Female	291	9.74	5.77	1.26	0.61
	Age	3	235	7.67	4.18	1.09	0.49
		4	338	10.75	6.11	1.35	0.65
	Ethnicity	White	81	13.15	6.24	1.54	0.67
		Black	349	8.49	4.92	1.16	0.54
		Hispanic	68	9.25	6.07	1.19	0.74
		Other	75	10.41	5.79	1.36	0.56
	Language	English	491	9.53	5.59	1.25	0.60
		DLL	82	9.24	5.76	1.20	0.66
	IEP	Yes	534	9.59	5.64	1.26	0.60
		No	39	8.10	5.10	1.05	0.65
Control	Gender	Male	121	8.30	4.83	1.12	0.52
		Female	136	9.78	5.63	1.27	0.56
	Age	3	107	6.85	3.56	0.99	0.42
		4	150	10.67	5.77	1.35	0.58
	Ethnicity	Missing	2	7.00	1.41	1.00	0.00
		White	31	10.81	5.87	1.35	0.55
		Black	108	8.83	5.16	1.15	0.54
		Hispanic	82	8.28	4.94	1.18	0.52
		Other	34	10.35	5.88	1.26	0.62
	Language	English	201	9.26	5.39	1.20	0.55
		DLL	56	8.45	5.03	1.18	0.54
	IEP	No	235	9.33	5.38	1.22	0.55
		Yes	22	6.41	3.59	0.95	0.49

Table B.7. C-TRF Total Problems score by child characteristics, Fall 2019

			C-TRF TP Raw Score			C-TRF TP T Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Male	261	19.35	23.01	46.63	12.38
		Female	275	12.59	15.02	45.63	9.78
	Age	3	221	14.56	17.86	45.67	10.56
		4	315	16.81	20.72	46.43	11.51
	Ethnicity	White	78	14.83	17.48	45.53	10.64
		Black	323	16.85	20.94	46.59	11.55
		Hispanic	67	16.31	19.64	45.85	11.41
		Other	68	12.09	14.47	44.82	9.24
	Language	English	462	15.72	19.98	45.92	11.22
		DLL	74	16.92	17.16	47.32	10.49
	IEP	No	500	14.91	18.79	45.62	10.98
		Yes	36	29.44	25.23	52.97	10.92
Control	Gender	Male	93	19.61	24.80	46.58	13.25
		Female	110	14.85	20.45	46.06	11.19
	Age	3	84	20.89	23.53	49.27	11.53
		4	119	14.31	21.64	44.24	12.19
	Ethnicity	Missing	2	18.50	14.85	52.00	8.49
		White	20	14.10	25.05	43.9	12.13
		Black	91	19.57	24.48	47.83	12.80
		Hispanic	60	14.25	18.74	45.05	11.07
		Other	30	16.77	22.98	45.43	12.44
	Language	English	165	18.68	24.24	47.14	12.60
		DLL	38	9.87	11.15	42.68	9.26
	IEP	No	184	15.54	21.31	45.60	11.83
		Yes	19	31.47	29.69	53.11	13.44

Table B.8. C-TRF Internalizing Problems raw score means by child characteristics, Fall 2019

			C-TRF IP Raw Score			C-TRF IP T Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Male	261	5.11	6.46	46.44	10.74
		Female	275	4.09	5.28	45.33	9.81
	Age	3	221	4.19	5.50	45.50	9.80
		4	315	4.87	6.16	46.13	10.60
	Ethnicity	White	78	5.33	6.67	46.6	11.24
		Black	323	4.72	6.10	46.15	10.38
		Hispanic	67	4.06	5.24	44.82	9.92
		Other	68	3.63	4.36	44.76	8.95
	Language	English	462	4.55	6.00	45.70	10.38
		DLL	74	4.85	5.30	46.95	9.60
	IEP	No	500	4.33	5.67	45.44	10.10
		Yes	36	8.25	7.69	51.92	10.94
Control	Gender	Male	93	4.75	6.33	45.62	10.59
		Female	110	4.64	6.60	45.74	10.52
	Age	3	84	5.32	6.55	47.20	10.12
		4	119	4.24	6.39	44.63	10.72
	Ethnicity	Missing	2	3.50	2.12	47.50	4.95
		White	20	3.50	5.00	44.15	9.12
		Black	91	5.42	7.00	46.76	11.05
		Hispanic	60	3.62	5.34	44.05	9.42
		Other	30	5.50	7.63	46.67	12.06
	Language	English	165	5.18	6.88	46.42	10.94
		DLL	38	2.58	3.48	42.53	7.89
	IEP	No	184	4.36	6.11	45.23	10.27
		Yes	19	7.89	8.76	50.11	12.21

Table B.9. C-TRF Externalizing Problems score means by child characteristics, Fall 2019

			C-TRF EP Raw Score			C-TRF EP T Score	
			Valid N	Mean	St.Dev.	Mean	St.Dev.
PHLpreK	Gender	Male	261	9.72	12.34	48.86	10.58
		Female	275	5.28	7.41	47.49	8.41
	Age	3	221	6.79	9.27	47.82	8.87
		4	315	7.90	11.03	48.40	9.99
	Ethnicity	White	78	5.97	8.17	47.03	8.50
		Black	323	7.94	10.99	48.54	9.89
		Hispanic	67	8.43	10.92	48.91	10.01
		Other	68	5.76	8.51	46.91	8.43
	Language	English	462	7.34	10.40	48.04	9.51
		DLL	74	8.08	10.00	48.91	9.76
	IEP	No	500	6.96	9.92	47.78	9.36
		Yes	36	14.06	13.58	53.42	10.57
Control	Gender	Male	93	9.83	12.91	48.82	11.27
		Female	110	6.55	9.74	48.40	9.44
	Age	3	84	10.05	11.30	51.28	9.78
		4	119	6.65	11.29	46.72	10.28
	Ethnicity	Missing	2	9.00	11.31	52.00	11.31
		White	20	7.35	14.21	47.50	11.34
		Black	91	9.30	12.21	49.82	10.91
		Hispanic	60	6.70	8.82	47.77	8.81
		Other	30	7.40	11.68	47.07	10.66
	Language	English	165	8.80	12.22	49.13	10.81
		DLL	38	4.82	5.78	46.26	7.43
	IEP	No	184	7.40	10.92	48.03	10.00
		Yes	19	14.42	14.06	54.00	11.87