NIEER
NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

EVALUATION OF WEST VIRGINIA UNIVERSAL PRE-K. FOURTH YEAR LONGITUDINAL OUTCOMES

June 2020

Milagros Nores, Ph.D., Kwanghee Jung, Ph.d., Elisa Valle, B.S and Carol Contreras, B.A. The National Institute for Early Education Research

Mindy Allenger, Ed.D., Marshall University

About the Authors

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

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

Elisa Valle, B.S. is Assistant Project Coordinator at The National Institute for Early Education Research (NIEER) at Rutgers University.

Carol Contreras, B.A. is Program Manager at The National Institute for Early Education Research (NIEER) at Rutgers University.

Mindy Allenger, Ed.D. is Assistant Professor at the College of Education & Professional Development, Marshall University. Dr. Allenger's research and teaching focuses on literacy education, diagnosis, and intervention.

Grateful acknowledgment is made to the West Virginia Department of Education's Office of Early Learning. In addition, the authors would like to thank the schools and school districts of Fayette, Greenbrier, Kanawha, Nicholas, Putnam, Roane and Wood, which opened their doors and classrooms to the research team.

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., Jung, K, Valle, E. Contreras, C., & M. Allenger (2020). Evaluation of West Virginia Universal Pre-K. Fourth Year Longitudinal Outcomes. New Brunswick, NJ: National Institute for Early Education Research.

Table of Contents

Introduction

The West Virginia (WV) Universal Pre-K system continues to serve a large number of its 4-year old children across all 55 counties in the state. In 32018-2019 it enrolled close to 14,000 children, equivalent to 59 percent of 4-year-olds in the state, ranking 8th in the nation in access to preschool for 4-year-olds (Friedman-Krauss et al., 2020). West Virginia has continuously invested in increased their quality standards in the last ten years. It meets all NIEER's minimum quality standards benchmarks and nine of the 10 most recent benchmarks (Friedman-Krauss et al., 2020). In addition, starting 2017-2018, the state implemented a new policy mandating at least 1,500 minutes (25 hours) of instruction per week and 48,000 minutes (800 hours) of instruction per year in preschool. In addition, in 2017, the West Virginia Legislature changed the enrollment date for pre-K and kindergarten to four and five before July 1. This was implemented in Pre-K during the 2018-19 school year and in 2019-2020 for the kindergarten cohort.

Given the expansion and investments in quality of the WV Universal Pre-K program, the West Virginia partnered with NIEER and Marshall University to study the quality and effectiveness of such programs, and how children progress through the early elementary years. A five-year longitudinal study was therefore commenced in the fall of 2015 with the goal of estimating the effects of the WV pre-K program, understand the extent to which initial benefits result in persistent educational advantages, and with an emphasis also in assessing the quality of the educational experiences of children through their P-3 progression. This report presents results for the fourth year of the study. It reports the results of a longitudinal cohort of children at the end of their second-grade year. The report examines results across various child developmental domains, and how results varied across children. In particular, the reports attempts to understand to what extent pre-K impacts are sustained over time and for whom.

In the previous kindergarten and first grade year (2016-2018) reports, we highlighted positive impacts of pre-K on children's learning and development as demonstrated in language and literacy that persist at kindergarten entry, as well as evidence that lower income children benefitted most from the program. Girls also showed stronger effects. However, positive impacts diminished by the end of kindergarten year and continued to fade out through end of first grade year, with non pre-K attending children starting to converge with pre-K attending children in the developmental outcomes measured. This report follows this end of first grade findings by assessing results at the end of second grade. The positive impacts of pre-k fade out at the end of second grade.

As found in the kindergarten classrooms, observations of second grade classroom quality following the longitudinal sample showed that classroom quality experienced by the longitudinal sample in early elementary years may be a strong contributor to this convergence. While second grade classrooms showed slightly higher quality than they experienced in first grade and kindergarten classrooms, and somewhat closer to what they experienced in preschool, the quality of instructional supports is still lower than what the children experienced in their preschool.

Limitations in this study are: a sample of children and classrooms from seven counties with lower enrollment rates in the WV Pre-K program (cor comparisons with a group of children that did not attend the program). This condition for inclusion also means these counties have been the slower ones in their pre-K program expansion. Thus, generalizability of the findings to the rest of the state is dependent on differences between these counties and those with larger enrollment rates.

Study Methods

In the fall of 2015, the National Institute for Early Education Research initiated a five-year longitudinal study of the WV pre-K program. This report includes the fourth year of the study in which the longitudinal cohort was assessed in the spring of their second-grade year. The current report estimates the differences between longitudinal children that attended pre-K with those that did not in various developmental areas. The following research questions were examined.

1. What is the impact of the prekindergarten program on children's language, math, literacy, and executive functions skill measures at the end of their second grade year?
2. Are there child subgroups (as defined by low income or child gender) that benefit more from the program than others?
3. What is the overall observed quality experienced by the longitudinal children from pre-K through to second grade?

Sample

In the fall of 2015, we randomly selected two groups of children as our initial sample: 599 children who were just beginning the WV pre-K program and 573 children who had attended the pre-K program the previous year and were beginning kindergarten. In the following 2016-2017 school year, pre-K group of children from the initial sample were followed into their kindergarten year and another group of children who did not attend WV pre-K was randomly selected from their same classrooms and schools, as a comparison group. That year the study sample consisted of 605 kindergarten children who attended the WV pre-K program the previous year[1] and another group of 366 kindergarten children who had not attended the WV pre-K. This report follows these two groups of children into second grade.

The second grade sample consists of 623[2] second grade children who attended WV pre-K program and another group of 333 second grade children who did not attend WV pre-K. Table 1 reports demographics for the sample of 956 children in the study. The sample is predominantly White (90.8%) and low income (51.8%), with a balanced gender composition (48.2% female). This sample does not differ significantly with the original sample. That is, children assessed in second grade were comparable to the average children in these districts in terms of gender and race. Control group of children were slightly less likely to be low income. We assessed between 1 and 15 children per second grade classroom, following the longitudinal children as the moved across the P-3 system.

In addition, classroom observations were conducted using the CLASS observation tool in 135 second grade classrooms and 125 pre-K classrooms. Classroom observation results for the preschool sample were reported in a separate report.

[1] Additional pre-K attenders were assessed in this round, who were originally identified as non-attenders, but then tracked in the West Virginia Education Identification System (WVEIS) as pre-K attenders.

[2] The n in this round increased because some children that had not been located in the K year were located this year. These children added to those assessed in the K round that were identified as attenders add to a larger n than assessed in pre-K.

Table 1. Child demographics for sample, N=956

Child Characteristics	Total sample N=956		Comparison K sample N=333		Treatment K sample N=623		WV school average for these districts*
	N	%	N	%	N	%	
Gender							
Male	495	51.8%	162	48.7%	333	53.5%	51.4%
Female	461	48.2%	171	51.3%	290	46.6%	49.1%
Low Income							
Low Income	499	52.2%	163	49.0%	336	53.9%	68.1%
Other	457	47.8%	170	51.0%	287	46.1%	31.9%
Race/Ethnicity							
White	868	90.8%	308	92.5%	560	89.9%	93.1%
Black	36	3.8%	7	2.1%	29	4.7%	3.5%
Other	52	5.4%	18	5.4%	34	5.4%	3.3%

*Source: WV Department of Education, https://zoomwv.k12.wv.us/Dashboard/portalHome.jsp.

Measures

Measures on Children

This evaluation assessed children's outcomes in receptive vocabulary (using the Peabody Picture Vocabulary Test), literacy (using the Woodcock-Johnson Tests of Achievement: Letter-Word subtest and Passage Comprehension subtest), and math (using the Woodcock-Johnson Tests of Achievement Applied Problems and Calculation subtest). Moreover, the research team also assessed executive functioning (EF) using the Minnesota Executive Function Scale (MEFS) and prosocial and behavior problems with the Strength and Difficulties Questionnaire (SDQ).

We follow with a description of these measures.

The *Peabody Picture Vocabulary Test—Fourth Edition (PPVT-IV;* Dunn & Dunn, 2007) is a test of receptive English vocabulary. The PPVT is predictive of general cognitive abilities and is a direct measure of vocabulary size. The test is reliable based on reported split-half reliabilities or test-retest reliabilities. 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). The test is normed at 100 with a standard deviation of 15.

The *Woodcock-Johnson Psycho-Educational Battery—Third Edition (WJ-III;* Woodcock, McGrew, Mather, & Schrank, 2001) includes multiple subtests. The *Applied Problem, Caculation, Letter-Word Identification, and Passage Comprehension* subtests were used in second grade. Correlations of the WJ-R 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; Wong, Cook, Barnett, & Jung, 2008; Barnett, et al., 2018). This test is also normed at 100 with a standard deviation of 15.

The Strength and Difficulties Questionnaire (SDQ; Goodman, 1999) is a 25-item questionnaire for assessing prosocial behavior and psychopathology of children and youths. It consists of five subscales including emotional symptoms, conduct problems,

hyperactivity/inattention, peer relationship problems and prosocial behavior. Teachers completed 3-point Likert scale to indicate how far each attributes applies to target child. Validation study of the scale indicated satisfactory reliability and validity and proved it to be a useful brief measure of the adjustment of children and adolescent (Goodman, 2001).

Minnesota Executive Function Scale (MEFS; Carlson & Zelazo, 2014) is a standardized assessment of executive function skills for children ages 2 and up. Children were instructed to sort test cards into one of two boxes according to an increasingly complex set of rules. The MEFS has been used with over 30,000 children and has adequate test-retest reliability (0.86; Carlson, 2017).

Measures on Classrooms

The experiences of the children in the longitudinal cohort has been assessed in the elementary years using observational measure:

Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). The CLASS is an observational system that assesses classroom practices in preschool and early elementary classrooms by measuring the interactions between students and adults. Observations consist of four to five 20-minute cycles followed by 10-minute coding periods. Scores (codes) are assigned during various classroom activities, and then averaged across all cycles for an overall quality score. Interactions are measured through 3 domains, which are divided into 10 dimensions. CLASS uses a 7-point Likert-scale, for which a score of 1 or 2 indicates low quality and a score of 6 or 7 indicates high quality.

Since the quality of the experiences of the longitudinal cohort has been consistently assessed utilizing the CLASS from preschool through second grade, we are able to track how this quality has changed over time.

Procedures

The classroom sample was assessed through the collaborative efforts of WVDE, Marshall University. NIEER, and county coordinators. Observers were trained by a CLASS Affiliate Trainer from NIEER that met the Teachstone[3] reliability requirements for trainer certification. All observers were then required to fulfill the Teachstone reliability process. In addition, all observers had to pass an in-person inter-rater reliability. Additionally, data collectors took and passed Teachstone's online calibration test mid-way through data collection in order to reduce scoring drift.

Observations were conducted between January and April 2019. Marshall University observers called schools in advance to schedule appointments for observations, and teacher names were disclosed at that time. All observation score sheets were cleaned, entered and analyzed at NIEER.

[3] Teachstone is the company that sells CLASS products and manages/sells CLASS observer trainings, certifications etc. All training activity is monitored and reported to them. http://www.teachstone.com/about-teachstone/

Results

Below we address the research questions on the associations between having attended the pre-K program and children's kindergarten entry, end of kindergarten, end of first grade, and end of second grade developmental outcomes. Main results are presented below with additional analyses included in the appendices. In addition, the last question addresses the effect of the policy change on the number of mandatory minutes of preschool.

1. What is the impact of the prekindergarten program on children's language, math, literacy, and executive functions skill measures at kindergarten entry, at the end of kindergarten, at the end of first grade year, and at the end of second grade year?

We start by reporting the average results for the spring of 2019 for the whole sample, and then separated by treatment and comparison groups. This is the spring of the longitudinal children's second grade year. For comparison purpose, descriptive results from kindergarten entry, at the end of kindergarten year, and at the end of first grade year are also reported.

We follow this with estimations using multi-level regression analysis on the association between children having attended pre-K (or not) and children's language, literacy, math, executive function skills, and behavioral adjustment. From the WVEIS (WV education information system) we are able to capture children's race and/or ethnicity (White, African American or Other), their gender and low-income status. These are included as controls in the estimations (coded as dichotomous variables with values 0 or 1).

Table 2a reports summary statistics on children's outcomes for the overall sample, as well as for the treatment and comparison groups separately. Raw and standard scores are reported for the PPVT and WJ-III subtests. These are standardized with a mean of 100 and a standard deviation of 15. This means that children scoring above the mean of 100 are showing outcomes above average for their age. Similarly, children scoring below the mean of 100 are scoring below the norm for their age. Table 2b reports outcomes on executive functions and socio-emotional difficulties.

Fall kindergarten entry scores, spring end of kindergarten scores, spring end of first grade scores, and spring end of second grade are also reported. At kindergarten entry and end of kindergarten, children in treatment group showed higher standard scores in all of the measures except end of K PPVT scores. This trend is consistent at the end of first grade, by which children in treatment groups showed higher standard scores in in all of the measures except PPVT standard scores. However, at the end of second grade, children in treatment groups showed lower standard scores in all of the measures. For SDQ measures, higher scores mean more difficulties in social-emotional behaviors except prosocial behavior. In Conduct problems and Peer relationship problems, children in treatment groups showed higher scores, but the difference is minimal.

Table 2a. Average child scores across the different measures for the total sample, treatment and comparison group, language, literacy and math outcomes

	Total sample		Comparison		Treatment	
	Mean	SD	Mean	SD	Mean	SD
PPVT Raw Fall K Score	99.79	19.70	99.81	20.29	99.78	19.35
PPVT Raw Spring K Score	110.20	18.32	111.84	18.66	109.20	18.06
PPVT Raw Spring 1st Score	124.49	17.56	125.60	17.58	123.82	17.53
PPVT Raw Spring 2nd Score	137.05	19.42	139.60	18.29	135.69	19.89
PPVT SS Fall K Score	105.02	14.77	104.42	15.35	105.39	14.41
PPVT SS Spring K Score	106.32	13.46	107.05	13.50	105.88	13.43
PPVT SS Spring 1st Score	105.18	12.92	105.67	13.34	104.89	12.67
PPVT SS Spring 2nd Score	105.32	14.57	106.68	14.23	104.59	14.7
WJ-AP Raw Fall K Score	16.80	4.19	16.95	4.31	16.72	4.12
WJ-AP Raw Spring K Score	20.62	3.67	20.64	3.69	20.62	3.67
WJ-AP Raw Spring 1st Score	24.64	4.31	24.78	4.02	24.55	4.48
WJ-AP Raw Spring 2nd Score	28.19	5.34	28.82	4.76	27.86	5.60
WJ-AP SS Fall K Score	102.41	12.79	102.00	13.44	102.66	12.38
WJ-AP SS Spring K Score	105.40	12.33	104.55	12.98	105.91	11.90
WJ-AP SS Spring 1st Score	103.04	15.07	102.87	14.62	103.14	15.35
WJ-AP SS Spring 2nd Score	99.70	16.70	101.05	15.54	98.98	17.26
WJ-LW Raw Fall K Score	12.96	5.44	13.09	5.86	12.88	5.17
WJ-LW Raw Spring K Score	23.56	6.89	23.68	6.87	23.48	6.90
WJ-LW Raw Spring 1st Score	34.11	9.29	34.24	9.38	34.03	9.24
WJ-LW Raw Spring 2nd Score	41.76	10.09	42.22	9.94	41.51	10.18
WJ-LW SS Fall K Score	97.06	12.29	96.25	13.61	97.55	11.41
WJ-LW SS Spring K Score	107.98	13.12	107.23	14.31	108.44	12.34
WJ-LW SS Spring 1st Score	105.90	14.54	105.41	15.71	106.20	13.79
WJ-LW SS Spring 2nd Score	102.02	15.33	102.09	15.30	101.98	15.36
WJ-PC Raw Fall K Score	5.39	2.14	5.48	2.16	5.33	2.12
WJ-PC Raw Spring K Score	9.30	4.16	9.49	4.02	9.18	4.25
WJ-PC Raw Spring 1st Score	17.04	5.94	17.17	5.98	16.97	5.93
WJ-PC Raw Spring 2nd Score	22.71	6.65	23.20	6.47	22.44	6.74
WJ-PC SS Fall K Score	95.45	10.05	94.66	10.83	95.93	9.53
WJ-PC SS Spring K Score	98.92	13.60	98.37	14.54	99.25	13.01
WJ-PC SS Spring 1st Score	96.02	14.40	95.59	15.63	96.28	13.62
WJ-PC SS Spring 2nd Score	94.72	15.46	95.44	15.13	94.34	15.64
WJ-C Raw Spring 2nd Score	11.33	3.22	11.58	3.23	11.21	3.21
WJ-C SS Spring 2nd Score	97.10	18.26	97.63	18.48	96.82	18.16

Note: Kindergarten N=971, First Grade N = 827, Second Grade N = 956

Table 2b. Average child scores across the different measures for the total sample, treatment and comparison group, executive functions and socio-emotional outcomes

	Total sample		Comparison		Treatment	
	Mean	SD	Mean	SD	Mean	SD
MEFS Raw Spring 2nd Score	67.79	16.15	68.63	16.21	67.34	16.12
MEFS SS Spring 2nd Score	101.38	11.07	101.97	11.16	101.06	11.01
SDQ Spring 2nd Score						
Emotional Symptoms	1.53	2.02	1.69	2.18	1.44	1.93
Conduct Problems	2.37	1.15	2.35	1.07	2.37	1.19
Hyperactivity/Inattention	4.58	1.44	4.62	1.40	4.56	1.46
Peer Relationship Problems	4.21	1.08	4.17	1.06	4.25	1.08
Prosocial Behavior	8.30	2.22	8.54	2.10	8.18	2.27
Difficulty	12.70	3.18	12.83	3.25	12.62	3.13

Note: Kindergarten N=971, First Grade N = 827, Second Grade N = 956

Multivariate analyses examine the association between children's learning and program features while simultaneously controlling for children's characteristics. Estimation models include information on the age of children, gender, race and ethnicity, school days by date of assessment, low income, and IEP status. Teacher characteristics include teacher educational attainment, teaching experiences, and early childhood education certification. Program features include class size, number of children with disabilities per classroom, and classroom quality represented by the CLASS domains.

Results are represented in estimation results and effects sizes. Effect sizes are the estimated effect (or β) expressed in terms of standard deviations of the control group (children who did not go to pre-K). To facilitate interpretation, the current gap at kindergarten entry, between the lowest income quintile and the highest income quintile is about one standard deviation nationally.

Multivariate estimations are presented longitudinally, first for kindergarten entry, then for the spring of K, followed by the spring of 1st grade and finally at the end of second grade.

Kindergarten Entry

Table 3 presents the estimates of the associations between pre-K program participation and child learning and development, along with associations for child characteristics. Children who participated in WV pre-K program evidence higher literacy and language development at kindergarten entry. This results were reported in Nores, et al. (2019).Overall, there are small but positive effects in literacy and language observed at kindergarten entry.

Low income children evidence statistically significant lower scores across all outcome measures. Children with an IEP (Individualized Education Program) showed significantly lower scores in literacy and language. Girls showed significant benefits of pre-K program participation in executive functions as measured by Peg Tapping. Children identified as White evidence slightly higher pre-K benefits on executive functions as measured by Peg Tapping. Statistically significant effects are bolded. Full estimations and sensitivity analyses are reported Nores, et al. (2019).

Table 3. Multivariate analyses of children's 2016 fall standard score in relation to child characteristics

	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Language (WJ/WM-PC)	Math (WJ/WM-AP)	DCCS	PT
Treatment	0.098	**0.144***	**0.144***	0.065	0.070	-0.108
	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)	(0.07)
Age	**-0.020***	**-0.069*****	**-0.089*****	**-0.058*****	0.011	**0.020***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.006	0.096	-0.002	0.038	0.117	**0.154***
	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)
White	0.088	-0.074	-0.034	0.151	0.000	**0.001***
	(0.13)	(0.11)	(0.08)	(0.12)	(0.00)	(0.00)
Low Income	**-0.400*****	**-0.425*****	**-0.197****	**-0.378*****	**-0.222****	**-0.235****
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
IEP	-0.042	**-0.356*****	**-0.370*****	-0.027	**-0.328****	**-0.463*****
	(0.22)	(0.07)	(0.08)	(0.26)	(0.10)	(0.11)
N	967	967	967	965	968	968

* p<0.05; ** p<0.01; *** p<0.001. Note: Reference groups omitted from the estimation are Males, Non-White, middle to high income. Schools included as control. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered at the classroom level.

End of Kindergarten

The estimates of the associations between the pre-K program, kindergarten program features and child characteristics with children's development at the end of kindergarten year is reported in Table 4. The positive impact of pre-K participation on literacy and language is no longer significant by the end of kindergarten. Children that attended preschool (Treatment) show small positive differences in language, literacy and math relative to children that did not experience the program, but these differences are not significant. In this estimation classroom features and quality are included to accounts for the impact of children's kindergarten experiences on their development level. This results were reported in detail in Nores, et al. (2019). A negative association is found between low income status, as well as IEP status, and all child outcomes. Gender and Race/Ethnicity do not show associations with child outcomes. Statistically significant effects are bolded. Full estimations and sensitivity analyses are shown reported in Nores, et al. (2019).

Table 4. Multivariate analyses of children's 2017 spring standard score in relation to child and classroom characteristics and CLASS domains

	Rec. Vocabulary (PPVT/TVIP)	Literacy (WJ/WM-LW)	Language (WJ/WM-PC)	Math (WJ/WM-AP)	DCCS	PT
Treatment	-0.064	0.079	0.034	0.077	0.026	-0.045
	(0.08)	(0.06)	(0.06)	(0.07)	(0.08)	(0.08)
Age	**-0.022***	**-0.085*****	**-0.104*****	**-0.065*****	**0.028****	**0.034*****
	(0.01)	**(0.01)**	**(0.01)**	**(0.01)**	**(0.01)**	**(0.01)**
Female	-0.020	0.053	0.069	-0.062	0.085	0.075
	(0.07)	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)
White	0.273	-0.016	-0.004	0.224	0.347	0.066
	(0.17)	(0.11)	(0.12)	(0.18)	(0.18)	(0.16)
Low Income	**-0.378*****	**-0.427*****	**-0.361*****	**-0.336*****	**-0.256****	**-0.267****
	(0.07)	**(0.07)**	**(0.07)**	**(0.07)**	**(0.09)**	**(0.10)**
IEP	**-0.311****	**-0.387*****	**-0.280****	**-0.324****	**-0.382****	**-0.510****
	(0.12)	**(0.08)**	**(0.09)**	(0.10)	**(0.15)**	**(0.17)**
N	806	806	805	806	805	803

* p<0.05; ** p<0.01; *** p<0.001. Note: Other controls are schools and indicators for missing income as well as variables on teacher characteristics and class features. Reference groups omitted from the estimation are Males, Non-White, middle to high income, Teacher Education less then Master's degree, Teacher experiences 0-5 years. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

End of 1st grade

The estimates of the associations between the pre-K program, 1st grade classroom features and child characteristics with children's development at the end of the 1st grade year are reported in Table 5. At the end of first grade, the impact of pre-K participation is no longer evident in the reported measures. Classroom quality and classroom features are included as children's 1st grade experiences. Teacher education, certification, and experiences are also included in this estimation as controls.

Table 5. Multivariate analyses of children's 2018 spring standard score in relation to child and classroom characteristics and CLASS domains

	Rec. Vocabularya (PPVT/TVIP)	Literacy (WJ/WM-LW)	Language (WJ/WM-PC)	Math (WJ/WM-AP)	DCCS
Treatment	-0.006	0.056	0.039	-0.007	0.055
	(0.08)	(0.07)	(0.06)	(0.07)	(0.07)
Retained	-0.650***	-0.887***	-0.806***	-0.758***	-0.451*
	(0.16)	(0.15)	(0.15)	(0.14)	(0.20)
Age	-0.024**	-0.081***	-0.077***	-0.078***	0.002
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.043	0.099	0.151**	-0.069	0.099

	(0.06)	(0.05)	(0.06)	(0.07)	(0.07)
White	0.174	0.082	-0.031	0.233	0.214
	(0.12)	(0.13)	(0.12)	(0.15)	(0.15)
Low Income	-0.419***	-0.396***	-0.433***	-0.418***	-0.284***
	(0.08)	(0.07)	(0.06)	(0.07)	(0.08)
IEP	-0.460***	-0.462***	-0.420***	-0.569***	-0.215
	(0.11)	(0.10)	(0.09)	(0.10)	(0.12)
N	827	827	827	827	827

* p<0.05; ** p<0.01; *** p<0.001. Note: Other control included is school, teacher education, teacher experience and certification, class size, inclusion ratios, and CLASS scores. Reference groups omitted from the estimation are Males, Non-White, middle to high income, Teacher Education less then Master's degree, Teacher experiences 0-5 years, . Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

End of 2nd grade

Table 6 and 7 present the estimates of the associations between the pre-K program, 2nd grade classroom features and child characteristics with children's development at the end of the 2nd grade year. In this grade we included the SDQ and MEFS. Like the previous set of estimations, we included individual, teacher and classroom characteristics as controls, as well as number of school days, class size, inclusion ratio and quality as measured by the CLASS. As observed for K and 1st grade, the positive effect of pre-K attendance in no longer observed.

Table 6. Multivariate analyses of children's 2019 spring standard score in relation to child and classroom characteristics and CLASS domains

	Rec. Vocabulary (PPVT/TVIP)	**Literacy (WJ/WM-LW)**	**Language (WJ/WM-PC)**	**Math (WJ/WM-AP)**	**Math (WJ/WM-C)**
Treatment	**-0.141***	-0.009	-0.039	0.002	-0.112
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
Retained	**-0.663*****	**-0.999*****	**-1.076*****	**-1.007*****	**-1.225*****
	(0.10)	(0.10)	(0.12)	(0.11)	(0.11)
Age	**-0.036*****	**-0.072*****	**-0.089*****	**-0.056*****	**-0.083*****
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female	**-0.182****	-0.056	**-0.114***	-0.022	**-0.158***
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
White	**0.261***	0.276*	0.055	0.262*	0.202
	(0.12)	(0.14)	(0.10)	(0.12)	(0.11)
Low Income	**-0.348*****	**-0.307*****	**-0.160****	**-0.292*****	**-0.286*****
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
IEP	**-0.430*****	**-0.582*****	**-0.446*****	**-0.621*****	**-0.492*****
	(0.09)	(0.09)	(0.10)	(0.09)	(0.10)
N	947	954	951	954	955

* p<0.05; ** p<0.01; *** p<0.001. Note: Other control included is school, teacher education, teacher experience and certification, class size, inclusion ratios, and CLASS scores. Reference groups omitted from the estimation are Males, Non-White, middle to high income, Teacher Education less then Master's degree, Teacher experiences 0-5 years. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Table 7 reports estimations on the effects of attending preschool in executive function (as measured by the MEFS) and in socio-emotional (SDQ). For the SDQ we calculate the prosocial behavior scale, as well as the total difficulties scale. No effects were found on either of these. We also estimated the effects of attending preschool on various socio-emotional problem subscales: emotional symptoms, conduct problems, hyperactivity/inattention and-peer relationship problems. While there are differences for females, and low income children among some of subscales and scales, no differences were observed between children that did and did not attend the preschool program.

Table 7. Multivariate analyses of children's 2019 spring MEFS and SDQ score in relation to child and classroom characteristics and CLASS domains

	MEFS	Pro social	Total Difficulty	Emotion	Conduct	Hyperactivity	Peer Problem
Treatment	-0.120	-0.127	-0.067	-0.108	-0.088	0.008	0.095
	(0.07)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
Age	**-0.044*****	0.002	0.013	0.017*	-0.005	-0.003	**0.014***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.028	**0.334*****	0.001	**0.200****	-0.073	**-0.337*****	0.109
	(0.07)	(0.08)	(0.07)	(0.06)	(0.08)	(0.08)	(0.07)
White	0.146	0.043	-0.061	-0.009	-0.138	0.032	-0.070
	(0.11)	(0.14)	(0.13)	(0.11)	(0.18)	(0.16)	(0.13)
Low Income	**-0.298*****	**-0.263*****	**0.169***	**0.206****	0.129	0.060	-0.113
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
IEP	**-0.225***	**-0.330*****	0.138	**0.265****	0.078	-0.109	-0.055
	(0.09)	(0.10)	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)
Retained	**-0.726*****	-0.181	0.183	0.185	0.159	-0.005	0.029
	(0.10)	(0.15)	(0.12)	(0.11)	(0.15)	(0.14)	(0.13)
N	932	905	905	905	905	905	905

* p<0.05; ** p<0.01; *** p<0.001. Note: Other control included is school, teacher education, teacher experience and certification, class size, inclusion ratios, and CLASS scores. Reference groups omitted from the estimation are Males, Non-White, middle to high income, Teacher Education less then Master's degree, Teacher experiences 0-5 years. Standardized scores are used for PPVT, and WJ or WM. Errors are clustered by classroom.

Longitudinal trends

Figures 1 and 2 below summarize effects over time for the longitudinal cohort. These are reported in effect sizes, which is calculated as the the estimated association standardized by dividing it by the standard deviation of the control group. The figure 1 reports the effects at kindergarten entry, at the end of kindergarten, at the end of 1st grade scores, and at the end of 2nd grade. Beneficial effects of language and literacy at the beginning of the kindergarten diminished through kindergarten and through the second grade year. Positive impacts of pre-K on math were consistent during the kindergarten year but the differences between children that attended and did not attend the preschool program started to converge by the end of first grade and continuously through the end of second grade.

Figure 1. Effect Size for Receptive Vocabulary, Literacy, Language and Math

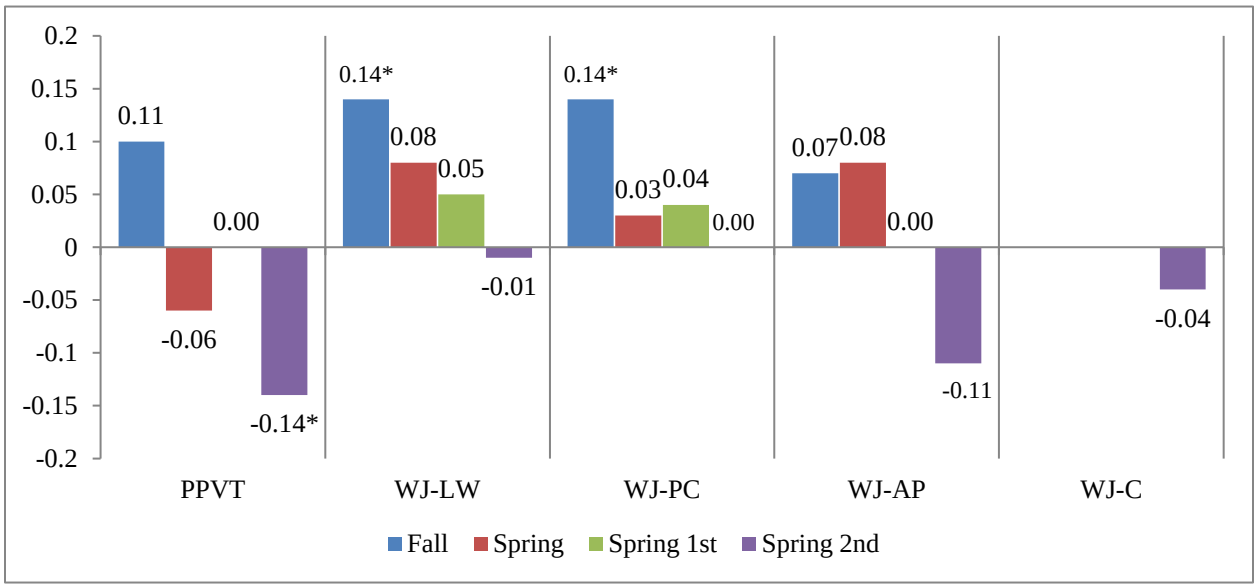
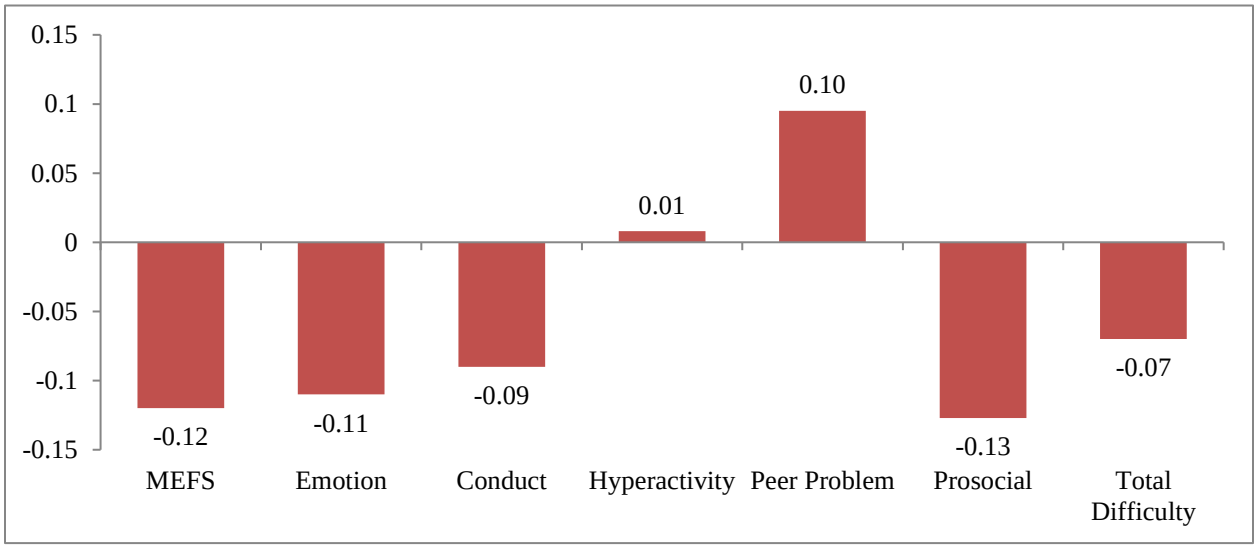


Figure 2 shows effects of executive functioning skills (MEFS) and social-emotional behaviors (SDQ) at the end of 2nd grade. No significant differences were found in either of these dimensions. Children who participated in pre-K program were less likely to show problems in emotional symptoms, conduct problems, and total difficulties, but these differences were not significant. Similarly, they showed lower prosocial behaviors as reported by teachers, and higher levels of peer problems.

Figure 2. Effect Size for MEFS and SDQ

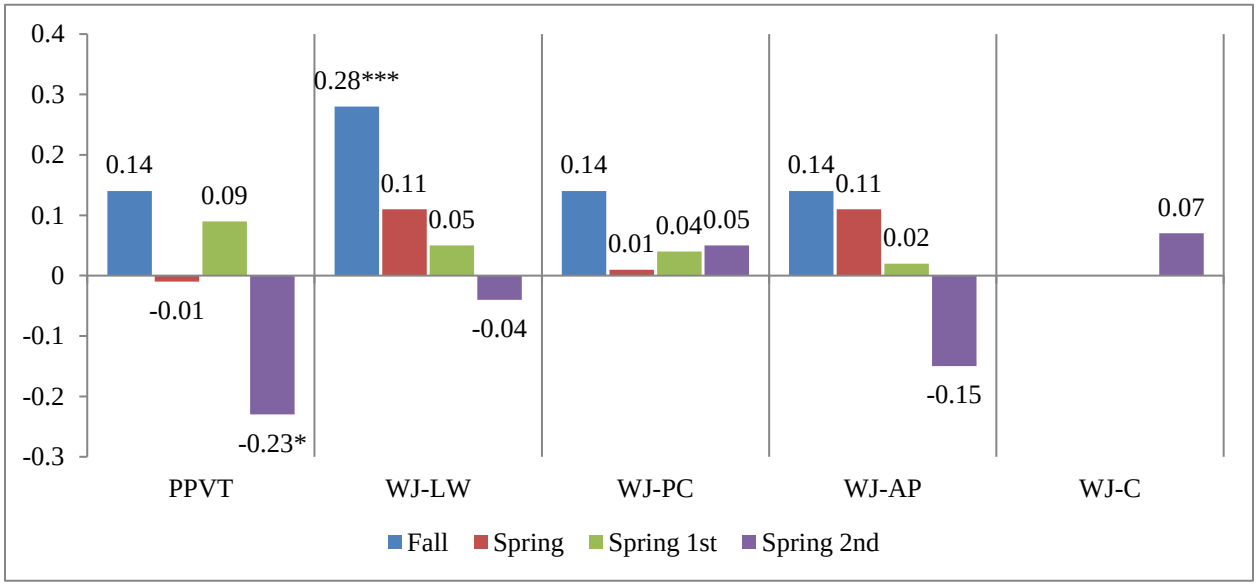


2. Are there child subgroups (as defined by low income or child gender) benefit more from the prekindergarten program than others?

We further analyzed the association between having attended pre-K and children's development levels for low income children and females. Figures 3 and 4 depicts the results from low income children at kindergarten entry, at the spring of kindergarten, at the spring of first grade and at the spring of the second grade. Estimations for females are presented in Figure 5 and 6 below.

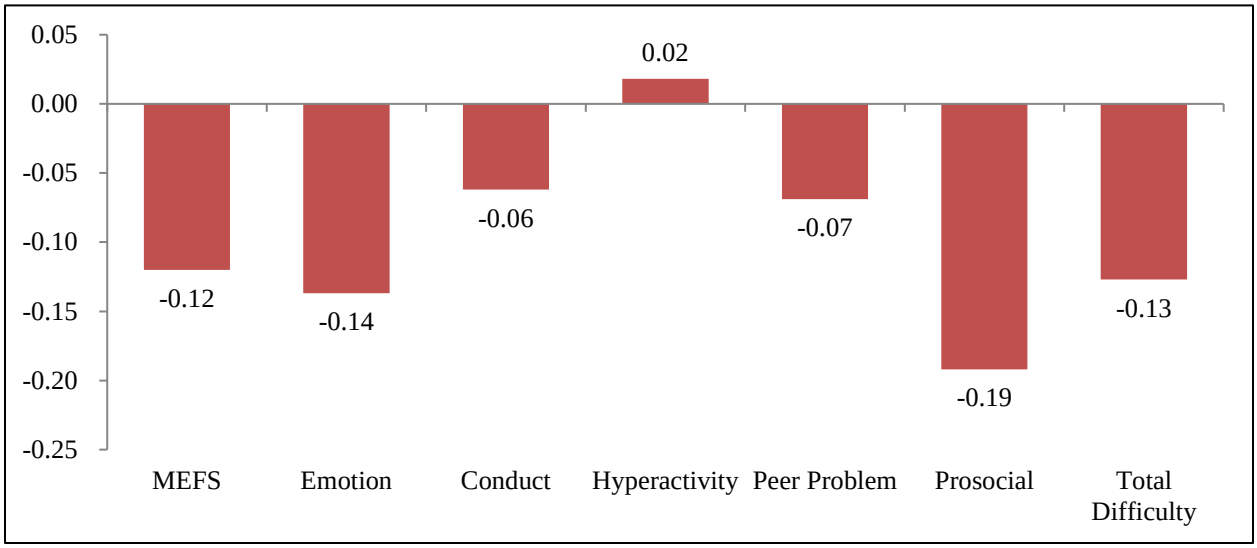
At kindergarten entry, children identified as low-income families show significant benefits in literacy from having attended pre-K. These effects are not observed for vocabulary or math. At the spring of kindergarten, low income children show small effects from pre-K participation sustained through the kindergarten year in literacy and math (see figure 3), albeit these are not significant. By the end of first grade, there is convergence in outcomes between children that did and did not attend pre-K. Language measured by WJ passage comprehension subscale and math measured by the WJ calculation subscale showed very small positive eof pre-K participation at the end of second grade (see figure 3).

Figure 3. Effect Size for Receptive Vocabulary, Literacy, Language and Math: Low Income only



There was no significant impact of pre-K participation in 2nd grade chdilren's executive functioning and behavioral adjustment. Children who participated in the pre-K program reported less problems in emotional symptoms, conduct problems, peer relationship problems and total difficulty, as well as lower prosocial behaviors (see figure 4).However, none of these differences were statistically significant.

Figure 4. Effect Size for MEFS and SDQ: Low Income only



At kindergarten entry, pre-K program effects on receptive vocabulary is highest for girls, though not significant (See figure 5). At the spring of first grade, females showed small positive but not significant By the end of second grade, children that attended the program show convergence in all the measured outcomes relative to children that did not attend the pre-K program.

At the spring of second grade, females showed small positive program effects in social-emotional behavior measures except for prosocial behavior. However, all of the differences were not significant. For executive functioning measured by MEFS, significant negative impact of pre-K participation are present in 2nd grade.

Figure 5. Effect Size for Receptive Vocabulary, Literacy, Language, and Math: Females only

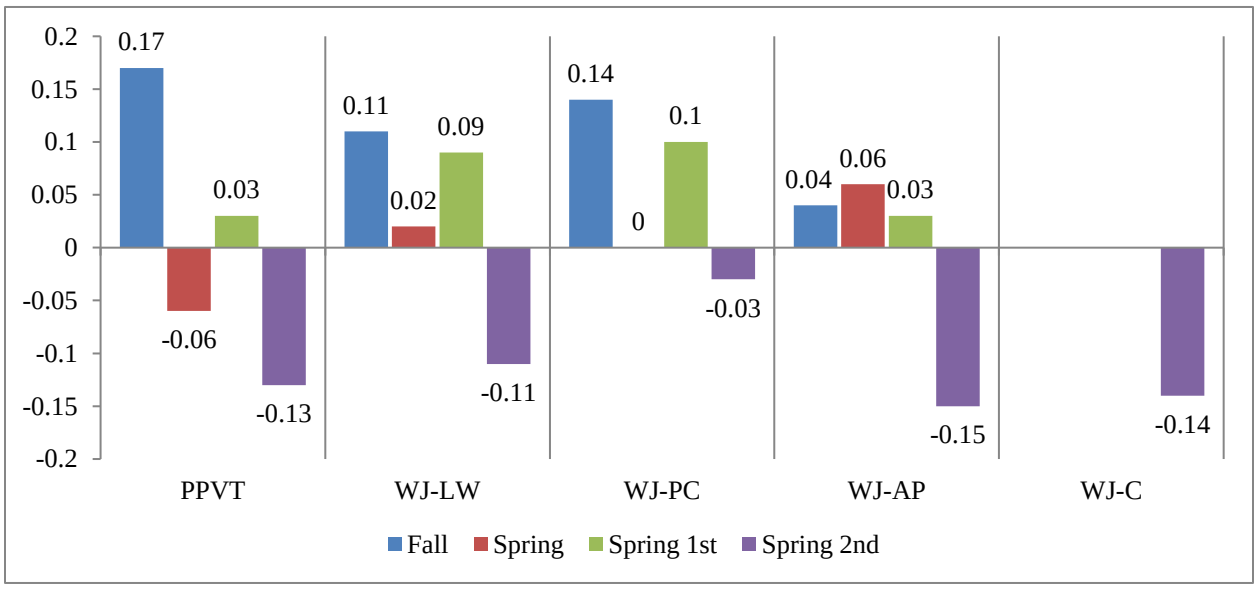
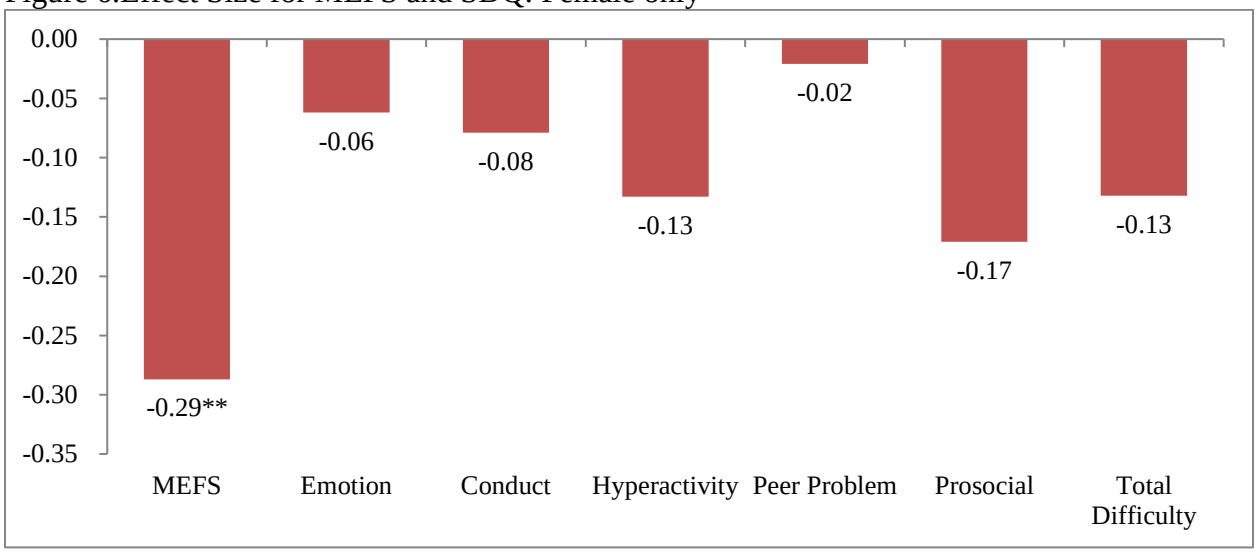


Figure 6.Effect Size for MEFS and SDQ: Female only



3. What is the overall observed quality experienced by the longitudinal children from pre-K through to second grade?

To understand possible drivers for the convergence trends observed above, we report classroom quality for the longitudinal cohort over the years. Table 6 and Figures 4-6 report classroom process quality for children in the sample using the CLASS. CLASS scores from pre-K classrooms in the spring of 2016, kindergarten classrooms in the spring of 2017, first grade classrooms in the spring of 2018, and second grade classrooms in the spring of 2019 are reported and illustrated below. Children in that attended pre-K experienced lower classroom quality in kindergarten across all CLASS domains, and even lower quality in 1st grade. In 2nd grade, however, quality was higher than in 1st grade across all three domains. Moreover, it was higher than the quality the longitudinal cohort experienced in their preschool year for the Emotional Support and Classroom Organization domains, while still lower for the Instructional Support domain. The low CLASS IS over time could be a strong contributor to the converging trends observed through the years in the pre-K group. Statistically significant differences in scores across the years are marked with an asterisk.

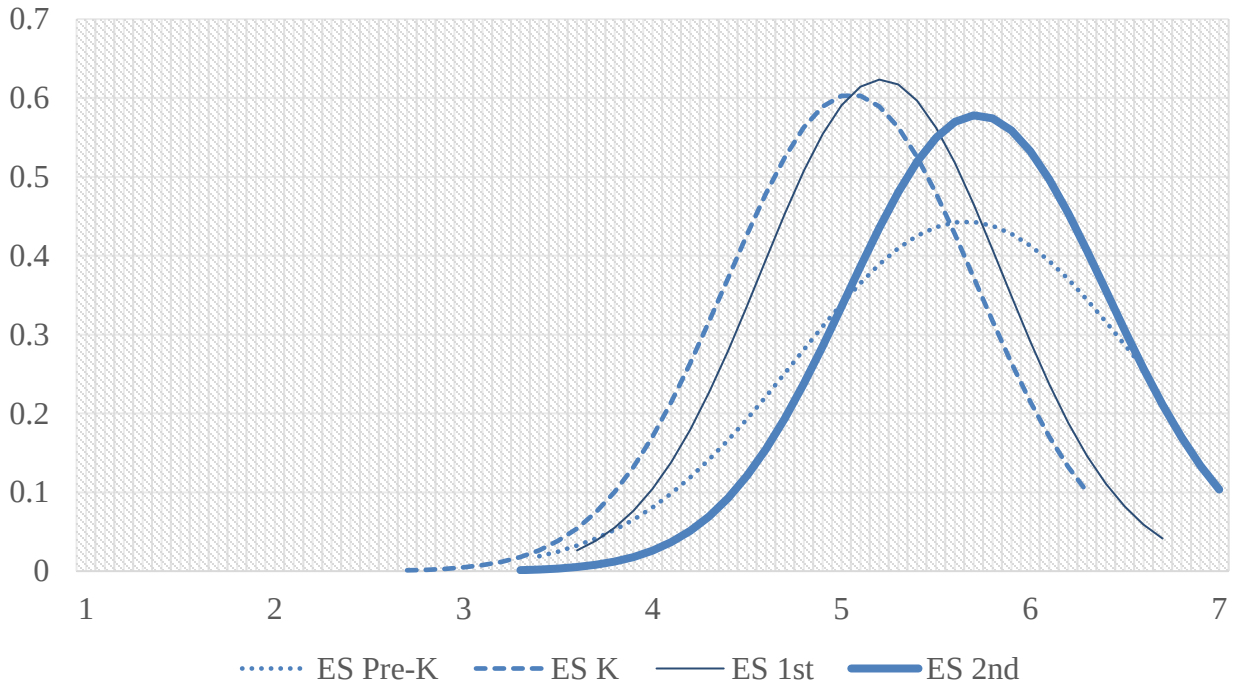
Table 6. CLASS Domains across the years for pre-K attenders in the Longitudinal Cohort

	Pre-K 2016 (N=105)		K 2017 (N=140)		1st grade (N=142)		2nd grade (N= 135)	
	Mean (range)	(SD)	Mean (range)	(SD)	Mean (range)	(SD)	Mean (range)	(SD)
Emotional Support	5.66***a (2.35-6.95)	(0.90)	5.05 (2.75-6.25)	(0.66)	5.21*b (3.60-6.70)	(0.64)	5.72***c (3.35-7.00)	(0.69)
Classroom Organization	5.09*a (1.33-6.87)	(1.16)	4.81 (2.27-6.40)	(0.81)	5.20***b (2.93-6.67)	(0.66)	5.58***c (2.53-6.87)	(0.72)
Instructional Support	2.65***a (1.13-5.33)	(0.83)	2.06 (1.00-4.93)	(0.72)	1.66***b (1.07-3.53)	(0.31)	2.24***c (1.00-4.33)	(0.76)

* p<0.05; ** p<0.01; *** p<0.001. Note. [a]PreK and K means are significantly different. [b]K and 1st grade means are significantly different. [c]1st grade and 2nd grade means are significantly different.

CLASS ES domain scores experienced by the longitudinal cohort in Figure 4 show lower CLASS ES levels in K and 1st grade, but 2nd grade CLASS ES levels similar to what the children experienced in their preschool year. There are no statistical differences between the CLASS ES distributions of preschool and 2nd grade.

Figure 4. CLASS ES Domain across the years for pre-K attenders in the Longitudinal Cohort



In CLASS CO, while K and 1st grade scores were quite low, 2nd grade shows higher CO levels experienced by the longitudinal cohort, than they experienced in any grade before, even in preschool. And the latter difference is statistically significant.

Figure 5. CLASS CO Domain across the years for pre-K attenders in the Longitudinal Cohort

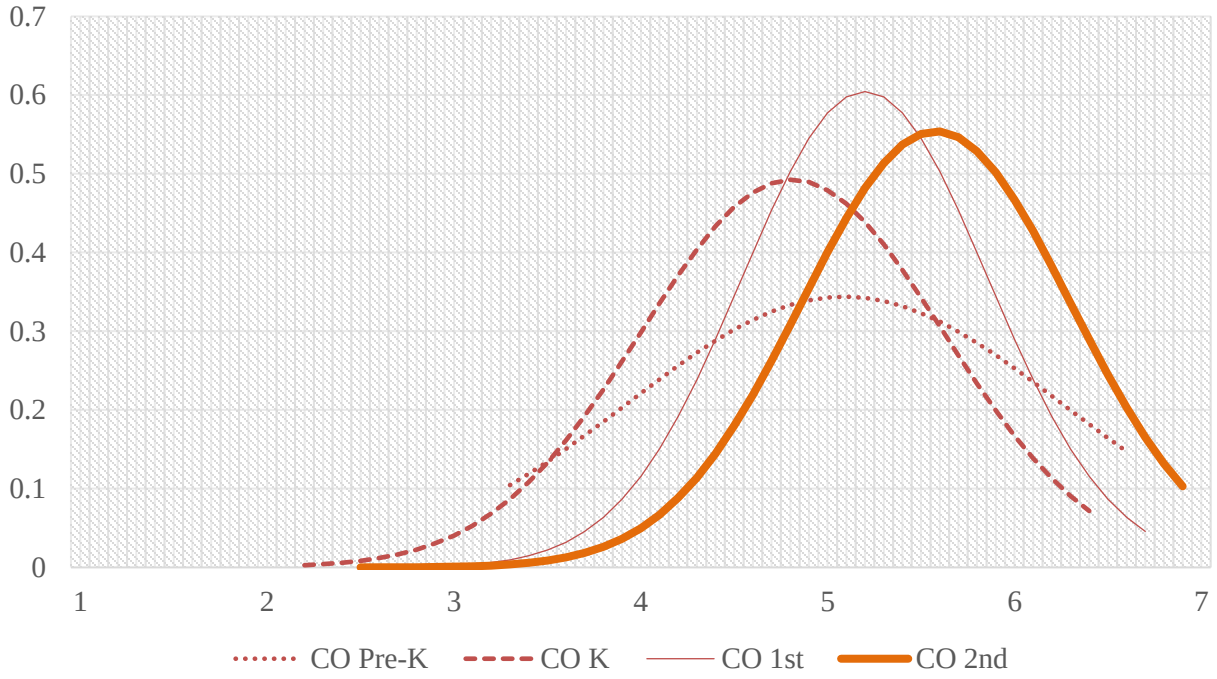
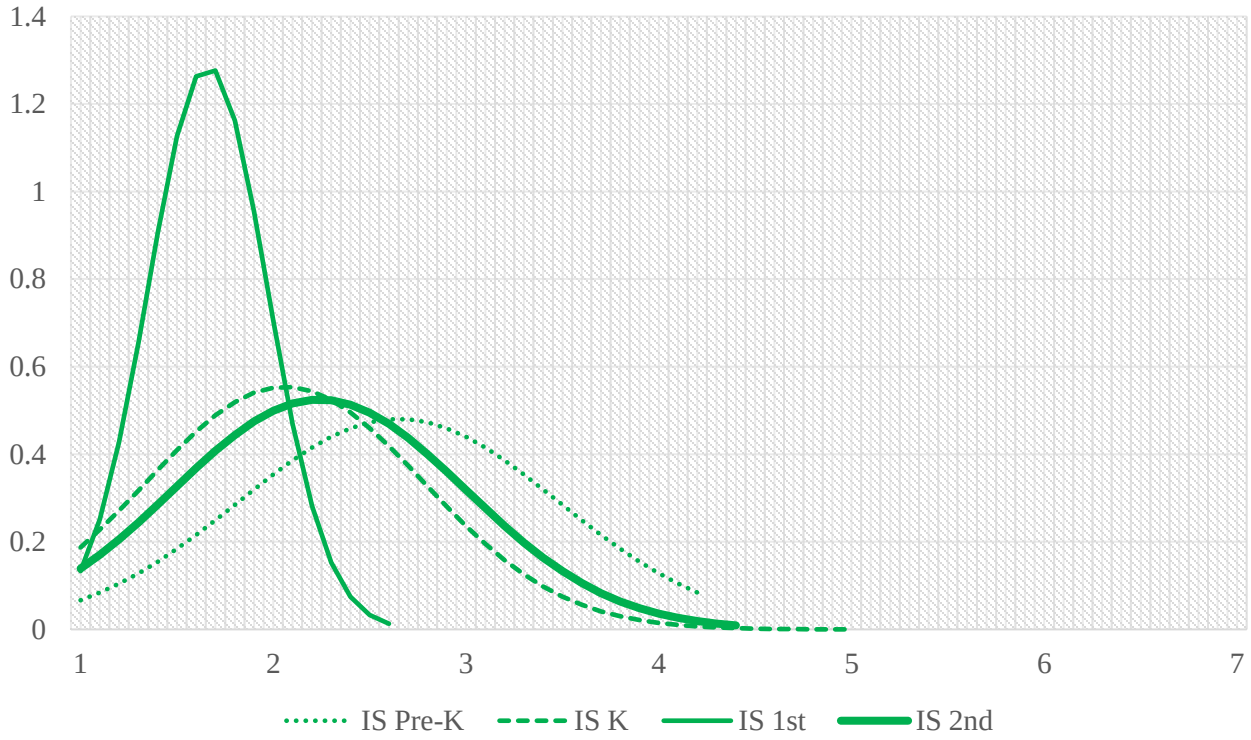


Figure 6 reports trends in CLASS IS. Unlike 1st grade scores, 2nd grade scores are closer to those observed in K and pre-K. Yet, these are lower than scores in the pre-K year. Preschool CLASS IS scores are statistically significantly higher than 2nd grade CLASS IS scores.

Figure 6. CLASS IS Domain across the years for pre-K attenders in the Longitudinal Cohort



Summary

This report presents findings from the fourth year (2018-19) of the WV Universal Pre-K Program evaluation. We investigated the impact of West Virginia's universal preschool program on children's language, math, literacy, and executive functioning skills through second grade. We also investigate results for low income children and females. The classroom quality the longitudinal cohort of children experienced in pre-K, kindergarten, 1st grade, and 2nd grade classrooms is reported.

We find positive impacts of attending the pre-k program on children's learning and development at kindergarten entry in language and literacy at kindergarten entry. The positive effects found for pre-K participation were greater for low income children on literacy and for girls on receptive vocabulary. Estimates for the spring of kindergarten and the spring of first grade show that some positive associations remain but are not significant anymore. That is, we observe convergence in the developmental trajectories of children that did and did not attend the preschool program. This is still evident in second grade scores. In the second grade testing wave we added a measure of socio-emotional difficulties. We observe, positive (albeit not significant) impacts on the behavior adjustment scale including in emotional symptoms, conduct problems, and total difficulty.

These results show strong consistency with a large number of evaluations of preschool programs. Rigorous evidence has shown positive short-terms impacts of such programs on children's development (Yoshikawa, et. al, 2013, Maxwell, et. al, 2001). The results suggest convergence between children that attended the preschool experience and children that did not,

similar to Yoshikawa, et. al's (2013) summary which states that convergence appears quite present over the elementary grades (p.9). Despite such convergence, the authors state that in many cases, there is evidence of effects on adult outcomes. In West Virginia, the convergence in the measured outcomes may likely be due to the low quality observed in kindergarten and first grade for the longitudinal cohort. Such low quality experiences would then fail to sustain the gains generated in children through their preschool experience. Another explanation for convergence may derive from issues with differentiation in classroom instruction in early elementary. That is, having children ready and with positive experiences entering K, in parallel to less prepared children may make teachers teachers focus on those starting behind, without having effective differentiation that would more successfully support growth across all children. Given that the counties in our study did have lower pre-K enrollment rates than other counties in the state to start with, it is feasible that a larger fraction of children enter K needing stronger supports due to their lack of preschool experience. Results may differ once a critical mass of students experience the preschool program as these may result in lesser differences between children in their early elementary experiences.

Recommendations

Some recommendations emerge from this study. Strengthening of instructional supports seems warranted across all early elementary grades. Intentional and strong supports would also support teachers through professional learning opportunities focusing on student' differentiation so that gains in Pre-K are sustained in K and into 2nd grade. It also would be relevant to further investigate why quality is lower in early elementary, and to target efforts at increasing process quality across all these grades.

Understanding differences between 1st and 2nd grade (Are teachers different? Who is assigned where?) would allow understanding why 1st grade quality observed what particularly low, and more so in relation to pre-K and 2nd grade quality. All of the findings point towards an opportunity to increase children's trajectory through a strong and purposeful focus on instructional supports in the elementary grades.

Lastly, the possibility of problems in differentiation in the early elementary grades due to only having a fraction of the children with preschool experience can be further investigated by assessing how and if districts with high pre-K coverage perform differently over time.

References

Barnett, S., Jung, K., Friedman-Krauss, A., Frede, E., Nores, M, Hustedt, J., Howes, C., & Daniel-Echols, M. (2018). State Prekindergarten Effects on Early Learning at Kindergarten Entry: An Analysis of Eight State Programs. *AERA Open 4*, 1-16.

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.

Carlson, S. M. (2017). *Minnesota Executive Function Scale: Technical Report*. St. Paul, MN: Reflection Sciences, Inc.

Carlson, S.M. & Zelazo P. D. (2014). Minnesota Executive Function Scale - Test Manual. Reflection Sciences; Saint Paul, MN.

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., ... & 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.

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G. & Gardiner, B. A. (2020). The State of Preschool 2019: State Preschool Yearbook. New Brunswick, NJ: National Institute for Early Education Research.

Goodman, R. (1999). T he extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent. Journal of Child Psychology and Psychiatry 40,791–801.

Goodman, R. (2001). Psychometric properties of the strength and difficulties questionnaire. Journal of the American academy of Child and Adolescent Psychiatry, 40, 1337-1345.

Maxwell, K. L., McWilliam, R. A., Hemmeter, M. L., Ault, M. J., & Schuster, J. W. (2001). Predictors of developmentally appropriate classroom practices in kindergarten through third grade. Early Childhood Research Quarterly, 16(4), 431-452.

Nores, M., Jung, K, Riley-Ayers, S., Kent, A., Contreras, C., & M. Allenger (2019). Evaluation of West Virginia Universal Pre-K. Second Year Longitudinal Outcomes. New Brunswick, NJ: National Institute for Early Education Research.

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

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

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

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

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

Appendix

Estimation: Full set of results and sensitivity analysis.

In the following appendix tables, we present effect sizes for estimations from standard scores and raw scores including various covariates. In the estimations, child characteristics such as age, gender, race/ethnicity, low income, and IEP status are included in all of the estimation model. First, effect size from full sample presented, followed by group of low income and female. Effect sizes from standard score at the end of second grade, raw score at the end of second grade, and standard score gain and raw score gain between spring first grade and spring second grade will be presented in order. Preferred models are those summarized in the main document. These tables provide sensitivity analyses.

Table A1. Effect Size for Standard Score at the End of 2nd Grade

Standard score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.16	-0.14	-0.16	-0.15	-0.14	-0.15
Letter-Word Identification(01)	-0.04	-0.01	-0.07	-0.04	-0.01	-0.06
Calculation (05)	-0.11	-0.04	-0.10	-0.11	-0.04	-0.10
Passage Comprehension (09)	-0.09	-0.00	-0.11	-0.08	-0.00	-0.10
Applied Problems (10)	-0.20	-0.11	-0.19	-0.18	-0.11	-0.19
MEFS	-0.12	-0.12	-0.13	-0.11	-0.12	-0.13
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A2. Effect Size for Raw Score at the End of 2nd Grade

Raw score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.17	-0.14	-0.16	-0.16	-0.14	-0.16
Letter-Word Identification(01)	-0.05	-0.02	-0.07	-0.04	-0.01	-0.06
Calculation (05)	-0.11	-0.03	-0.09	-0.11	-0.03	-0.09
Passage Comprehension (09)	-0.09	0.00	-0.10	-0.08	0.01	-0.09
Applied Problems (10)	-0.20	-0.11	-0.19	-0.19	-0.11	-0.19
MEFS	-0.10	-0.11	-0.12	-0.09	-0.11	-0.11
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A3. Effect Size for Standard Score Gain during 1st and 2nd Grade

Standard score	M5	M6	M7	M8	M9	M10
Receptive Vocabulary	-0.07	-0.07	-0.07	-0.06	-0.07	-0.07
Letter-Word Identification(01)	-0.01	0.02	-0.01	-0.00	0.03	-0.01
Passage Comprehension (09)	-0.05	0.03	-0.06	-0.05	0.04	-0.06
Applied Problems (10)	-0.13	-0.02	-0.13	-0.13	-0.02	-0.12
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A4. Effect Size for Raw Score Gain during 1st and 2nd Grade

Raw score	M5	M6	M7	M8	M9	M10
Receptive Vocabulary	**-0.08**	**-0.06**	**-0.08**	**-0.07**	**-0.06**	**-0.07**
Letter-Word Identification(01)	-0.02	0.02	-0.02	-0.02	0.02	-0.02
Passage Comprehension (09)	-0.06	0.04	-0.06	-0.05	0.05	-0.05
Applied Problems (10)	-0.14	-0.01	-0.13	-0.13	-0.01	-0.12
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A5. Effect Size for SDQ for 2nd Grade

Score	M5	M6	M7	M8	M9	M10
Emotion	-0.10	-0.11	-0.10	-0.09	-0.11	-0.10
Conduct	-0.01	-0.09	-0.01	-0.01	-0.09	-0.01
Hyperactivity	-0.06	0.01	-0.06	-0.05	0.01	-0.05
Peer Problem	0.09	0.10	0.10	0.10	0.10	0.10
Prosocial	-0.11	-0.13	-0.14	-0.12	-0.13	-0.15
Total difficulty	-0.07	-0.07	-0.07	-0.05	-0.07	-0.06
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Estimations from Low Income only

Table A6. Effect Size for Standard Score at the End of 2nd grade： Low Income Only

Standard score	M5	M6	M7	M8	M9	M10
Receptive Vocabulary	-0.25	-0.23	-0.26	-0.24	-0.23	-0.26
Letter-Word Identification(01)	-0.06	-0.05	-0.10	-0.06	-0.04	-0.09
Calculation (05)	-0.03	0.07	-0.04	-0.03	0.07	-0.05
Passage Comprehension (09)	-0.08	0.04	-0.11	-0.08	0.05	-0.11
Applied Problems (10)	-0.19	-0.15	-0.22	-0.18	-0.15	-0.22
MEFS	-0.17	-0.12	-0.19	-0.16	-0.12	-0.18
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A7. Effect Size for Raw Score at the End of 2nd grade: Low Income Only

Raw score	M5	M6	M7	M8	M9	M10
Receptive Vocabulary	-0.25	-0.21	-0.27	-0.25	-0.22	-0.27
Letter-Word Identification(01)	-0.07	-0.07	-0.10	-0.06	-0.06	-0.10
Calculation (05)	-0.03	0.08	-0.05	-0.04	0.08	-0.05
Passage Comprehension (09)	-0.06	0.05	-0.10	-0.07	0.06	-0.10
Applied Problems (10)	-0.19	-0.15	-0.23	-0.18	-0.15	-0.23
MEFS	-0.14	-0.09	-0.16	-0.14	-0.08	-0.16
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A8. Effect Size for Standard Score Gain during 1st and 2nd Grade: Low Income Only

Standard score	M5	M6	M7	M8	M9	M10
Receptive Vocabulary	-0.21	-0.24	-0.21	-0.20	-0.24	-0.20
Letter-Word Identification(01)	-0.03	0.02	-0.03	-0.03	0.01	-0.03
Passage Comprehension (09)	-0.06	0.05	-0.08	-0.06	0.05	-0.08
Applied Problems (10)	-0.17	-0.12	-0.21	-0.17	-0.13	-0.21
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x

with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A9. Effect Size for Raw Score Gain during 1st and 2nd Grade: Low Income Only

Raw score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.22	-0.23	-0.21	-0.22	-0.22	-0.21
Letter-Word Identification(01)	-0.04	0.00	-0.04	-0.04	0.00	-0.04
Passage Comprehension (09)	-0.04	0.09	-0.06	-0.04	0.09	-0.06
Applied Problems (10)	-0.18	-0.13	-0.22	-0.18	-0.14	-0.22
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A10. Effect Size for SDQ for 2nd Grade: Low Income Only

Score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Emotion	-0.16	-0.13	-0.18	-0.15	-0.14	-0.18
Conduct	0.02	-0.06	0.00	0.02	-0.06	0.01
Hyperactivity	-0.01	0.00	-0.13	-0.00	0.02	-0.02
Peer Problem	-0.03	-0.06	-0.06	-0.04	-0.07	-0.06
Prosocial	-0.21	-0.20	-0.24	-0.21	-0.19	-0.24
Total difficulty	-0.12	-0.13	-0.15	-0.11	-0.13	-0.14
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Estimations from Female only

Table A11. Effect Size for Standard Score at the End of 2nd Grade: Female Only

Standard score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.15	-0.12	-0.14	-0.15	-0.13	-0.14
Letter-Word Identification(01)	-0.12	-0.12	-0.11	-0.11	-0.11	-0.10
Calculation (05)	-0.20	-0.13	-0.17	-0.20	-0.14	-0.18
Passage Comprehension (09)	-0.12	-0.04	-0.11	-0.11	-0.03	-0.10
Applied Problems (10)	-0.20	-0.15	-0.18	-0.19	-0.15	-0.18
MEFS	-0.22	-0.29	-0.21	-0.22	-0.29	-0.21
no f.e.	x			x		

with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A12. Effect Size for Raw Score at the End of 2nd Grade: Female Only

Raw score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.17	-0.15	-0.16	-0.16	-0.15	-0.15
Letter-Word Identification(01)	-0.12	-0.12	-0.12	-0.11	-0.11	-0.11
Calculation (05)	-0.21	-0.12	-0.17	-0.20	-0.13	-0.18
Passage Comprehension (09)	-0.11	-0.02	-0.09	-0.10	-0.01	-0.08
Applied Problems (10)	-0.20	-0.16	-0.19	-0.19	-0.16	-0.19
MEFS	-0.21	-0.28	-0.20	-0.22	-0.28	-0.20
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A13. Effect Size for Standard Score Gain during 1st and 2nd Grade: Female Only

Standard score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04
Letter-Word Identification(01)	-0.04	-0.00	-0.04	-0.03	0.00	-0.04
Passage Comprehension (09)	-0.01	0.14	-0.00	-0.02	0.15	-0.00
Applied Problems (10)	-0.11	0.01	-0.12	-0.1	0.01	-0.11
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A14. Effect Size for Raw Score Gain during 1st and 2nd Grade: Female only

Raw score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Receptive Vocabulary	-0.03	-0.04	-0.03	-0.03	-0.04	-0.04
Letter-Word Identification(01)	-0.05	0.00	-0.05	-0.05	0.00	-0.05
Passage Comprehension (09)	-0.01	0.16	0.00	-0.01	0.16	0.00
Applied Problems (10)	-0.12	0.00	-0.13	-0.11	0.00	-0.12
no f.e.	x			x		
with school f.e.		x			x	

with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x

Table A15. Effect Size for SDQ for 2nd Grade: Female Only

Score	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
Emotion	-0.05	-0/06	-0.08	-0.05	-0.06	-0.08
Conduct	-0.01	-0.07	-0.03	-0.01	-0.08	-0.03
Hyperactivity	-0.16	-0.13	-0.14	-0.16	-0.13	-0.15
Peer Problem	0.00	-0.02	-0.00	0.01	-0.02	0.00
Prosocial	-0.21	-0.16	-0.23	-0.23	-0.17	-0.23
Total difficulty	-0.11	-0.13	-0.13	-0.10	-0.13	-0.12
no f.e.	x			x		
with school f.e.		x			x	
with county f.e.			x			x
with % inclusion	x	x	x	x	x	x
with teacher characteristics	x	x	x	x	x	x
with CLASS dimensions				x	x	x