AERA Open
October-December 2015, Vol. 1, No. 4 pp. 1–17
DOI: 10.1177/2332858415612751
© The Author(s) 2015. http://ero.sagepub.com

Early Grade Teacher Effectiveness and Pre-K Effect Persistence: Evidence From Tennessee

Walker A. Swain
Matthew G. Springer

Tennessee Consortium on Research, Evaluation, and Development,
Vanderbilt University's Peabody College, Nashville, TN, USA

Kerry G. Hofer

Abt Associates, Inc, Bethesda, MD, USA

In recent years, states have significantly expanded access to prekindergarten (pre-K), and federal policy makers have proposed funding near-universal access across the country. However, researchers know relatively little about the role of subsequent experiences in prolonging or truncating the persistence of benefits for participants. This study examines the interaction between pre-K participation and one of our most important educational interventions—teaching quality. We pair student-level data from a statewide pre-K experiment with records of teacher observation scores from Tennessee's new formal evaluation program to assess whether a student's access to high-quality early grade teachers moderates the persistence of pre-K effects. Our analyses indicate a small positive interaction between teaching quality and state pre-K exposure on some but not all early elementary cognitive measures, such that better teaching quality in years subsequent to pre-K is associated with more persistent positive pre-K effects.

Keywords: *administrative data, prekindergarten, teacher effectiveness*

Advocates of early childhood education generally view preschooling intervention as a vital and underutilized tool to narrow racial and socioeconomic outcome gaps in school and beyond (e.g., Doggett & Wat, 2010). Opponents tend to argue that the benefits of the programs are too short-lived to justify the costs (e.g., Dalmia & Snell, 2008). One important way research can inform this debate is by developing a stronger understanding of the factors that contribute to or inhibit the persistence of preschool benefits. In this article we utilize data from a public pre-K evaluation in Tennessee, matched with school administrative records and data from a new teacher evaluation program, to examine the interaction between pre-K participation and a factor that is as elusive to measure as it is universally accepted as vital to student outcomes—teaching quality.

Pre-K Expansion and Effects

In recent years, state-financed preschool programs have expanded dramatically. Enrollment in the past decade in state programs has more than doubled, with several states going as far as offering universal programs (Hustedt & Barnett, 2011). President Obama has made a concerted effort to push legislation that would make universal pre-K access federal law. However much of the research cited by politicians supporting these types of broad expansions comes from a few resource-intensive targeted experimental programs that have demonstrated remarkable benefits (Campbell et al., 2008; Campbell et al., 2012; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Nores, Belfield, Barnett, & Schweinhart, 2005). Unfortunately, these model programs may have little in common with the type of programs states have implemented at scale and propose to grow in the future.

To answer the question of whether children who participate in a statewide public pre-K program make greater academic and behavioral gains than similar children who do not participate in the program, Vanderbilt University's Peabody Research Institute (PRI) initiated a rigorous evaluation of Tennessee's Voluntary Pre-Kindergarten Program (TN-VPK) in 2009. Funded by the Institute of Education Sciences and with the assistance of the Tennessee Department of Education's Division of Curriculum and Instruction, this project utilized two primary designs, the first of which was a randomized controlled trial (RCT)—to evaluate TN-VPK. The other piece of the project utilizes an age-cutoff regression discontinuity design that exploits a sharp age cutoff requirement to compare 1-year gains for students who just meet the age requirement to those whose birthdays require them to wait a year before enrolling. In this study, we use data from the RCT as it permits evaluation of longitudinal program effects.

It is well documented that while pre-K programs drive early measurable cognitive gains (e.g., Bassok, 2010; Duncan, Bailey, & Yu, 2015; Gormley, 2008; Gormley, Gayer, Phillips, & Dawson, 2005; Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013; Magnuson, Ruhm, & Waldfogel, 2007a; Weiland & Yoshikawa, 2013), the achievement effects tend to fade by third grade or sooner (e.g., Currie & Thomas, 2000; Deming, 2009; Magnuson, Ruhm, & Waldfogel, 2007b). Seemingly contradicting "fade-out" concerns, some report that benefits resurface in the form of improved outcomes like high school graduation, college going, and lower rates of incarceration later in life (Barnett, 1995; Currie & Thomas, 2000; Deming, 2009; Garces, Currie, & Thomas, 2002; Hustedt & Barnett, 2005; Ludwig & Miller, 2007). Sophisticated long-run evaluations of Perry Preschool (Heckman et al., 2010) and TN STAR (Chetty et al., 2011) have suggested that preschool and early elementary school quality (teacher experience, peer test scores, class size) interventions can benefit students' long-run earnings dramatically even where early cognitive measure indicate fade-out, potentially through elevated "noncognitive" skills. However, while nonacademic skill acquisition may be the primary pathway by which early childhood interventions have improved students' life outcomes, in the current education policy context of test score accountability, and a labor market that potentially places increasing weight on educational attainment, there are still reasons to explore levers to facilitate persistence of the apparent early cognitive benefits preschool interventions.

Most analyses of differences in persistence rates have focused on student characteristics, such as race, ethnicity, or some measure of socioeconomic status (Currie & Thomas, 2000; Gormley, 2008; Magnuson et al., 2007a), and have sometimes hypothesized that differential persistence rates, for example between Black and White Head Start participants, could be attributable to subsequent access to higher-quality schooling (Currie & Thomas, 2000; Lee & Loeb, 1995). This study seeks to better examine nonascriptive characteristics that lengthen or cut short the persistence of cognitive gains resulting from students attending a school-based voluntary pre-K program.

The Potential Moderating Role of Instructional Quality

There are several ways one might hypothesize that teaching quality in grades following pre-K could alter the persistence of pre-K effects. If higher quality teachers are better equipped to differentiate instruction and adjust to the higher levels of preparation of pre-K participants, one could expect the benefits of pre-K to persist longer or even grow. Students' who attended pre-K might also be better equipped to benefit from strong teachers who emphasize more academically rigorous content. Alternatively, if better teachers emphasize catching up the least prepared students, the gap between the pre-K participants and control students could close more rapidly as targeting the least prepared student shifts attention away from the more prepared student. Students who had gotten the early preparation of pre-K could essentially flatline in terms of gain, or even have decreasing gains as progress is not maintained. Teachers whose instruction is of lower quality might slow the academic progress of either or both groups.

To our knowledge this is the first study to specifically examine the relationship between the formal evaluation rating of teachers to whom students are assigned after pre-K and the persistence of benefits students may have received from attending pre-K. Prior research has reported on the interaction of early childhood education and other more general school-level indicators of quality, including school test scores (Currie & Thomas, 2000) and measures of safety and academic environment (Lee & Loeb, 1995). Magnuson and colleagues (2007b) used two classroom quality measures—a teacher survey response to questions about the amount of time spent on instruction relative to other activities and data collected on the size of each classroom—to study the relationship between classroom experience and the persistence of preschool effects. They found that better scores on both classroom quality measures (i.e., small class size and high instructional time) were associated with a diminished preschool-related gap in student academic performance, while initial gaps in student performance persisted when children were enrolled in large classes or experienced smaller quantities of reading instruction. However, Bassok, Gibbs, and Latham's (2015) recent working paper, which utilizes the same 1998 ECLS-K sample, as well as the 2010 sample to explore changes in persistence patterns, finds no consistent interaction between preschool experiences and kindergarten year quality measures, including class size, peer preschool experience, full-day kindergarten, and an index for quality.

Contributions of the Current Study

This study contributes to the existing literature on factors that moderate pre-K effect persistence in two of important ways. First, the preschool treatment in our study is relatively clearly defined. In contrast to studies that rely on responses to questions on nationally representative surveys to determine whether a student participated in center or school-based preschool (e.g., Claessens, Engel, & Curran et al., 2014; Deming, 2009; Magnuson et al., 2007b), which can mean very different things depending on where the respondent lives, our study examines a well-defined statewide program with relatively strictly enforced standards set by the Tennessee Board of Education.[1]

Second, while prior studies have often relied on teacher self-reports of classroom conditions, our access to administrative data from Tennessee's newly implemented teacher evaluation programs allows us to incorporate rare substantive information on the perceived quality of early grade teachers, whose students are generally untested and thus lack value-added measures of effectiveness. The primary component of the evaluation for teachers in nontested grades (including kindergarten and first grade) is a series of annual classroom observations conducted by a trained observer and spread across the year (four visits for fully licensed teachers and six visits for those operating on an apprentice license). Observation scores are based on a detailed statewide rubric where teachers are rated on a range of categories, within the domains of instruction, planning, environment, and professionalism (see Online Appendix A for a sample evaluation rubric).[2] Teachers of untested subject areas also have a component of their score determined by a state board–approved achievement test chosen in agreement between the teacher and the evaluator (e.g., SAT 10 or DIBELS), and a school level growth score. Ultimately, the composite of these measures is reduced to a 5-point rating scale, by which teachers are categorized as follows: 1 = *significantly below expectations*, 2 = *below expectations*, 3 = *meets expectations*, 4 = *above expectations*, 5 = *significantly above expectations*.[3] While observations and scoring are of course subjective, the consequences associated with different scores are relatively consistent across teachers. Scores above a Level 3 help teacher's secure teacher tenure, and Level 5 ratings have been tied to salary bonuses and decreased oversight. These admittedly flawed though consequential categories are used as the primary measure for the construct of individual teaching quality in this study.

The main results from the TN-VPK evaluation found evidence of strong program effects on test scores at the end of the pre-K year (effect size of .33 on composite cognitive assessment). However, the cognitive gains experienced by program participants faded rapidly, with treatment and control groups being statistically equivalent on tested measures by the end of first grade (Lipsey et al., 2013). Access to longitudinal outcome measures and extensive administrative records on participants in this RCT presents ideal circumstances to explore the role of school-based factors in determining the persistence of pre-K effects.[4] In this article, we focus specifically on the interaction of TN-VPK exposure and the quality of a student's kindergarten and first grade teacher, as measured by teacher evaluation ratings. Our analyses seek to answer two closely related primary research questions: To what extent does early grade teaching quality moderate the effects of attending TN-VPK? And does the magnitude of that moderated effect vary based on a student's academic preparedness, specifically for students with low baseline cognitive scores and nonnative English speakers?

These questions are relevant to researchers and policy makers alike. For researchers, they offer some insight into the lingering questions about the drivers of fade-out, catch up, or persistence of early interventions. For policy makers, a positive interaction between pre-K participation and teaching quality would indicate that policies promoting the placement and retention of high-quality teachers in early, generally low-accountability grade levels could help prolong the cognitive gains made by students who participate in pre-elementary programs. Furthermore, if we hypothesize that higher rated teachers emphasize higher order skills, consistent with or building on those taught in pre-K, a positive interaction between pre-K participation and our measure of teaching quality should be most pronounced for students with the largest baseline deficits, who would otherwise lack the skills to benefit from instruction focused on more advance material. Evidence of pre-K providing this type of cognitive scaffolding has important implications for efforts to close stubborn academic achievement gaps.

Data and Sample

The data in this study come from two primary sources. First, student information and data were collected by researchers at PRI as part of a large-scale experimental evaluation of the TN-VPK program. Second, teacher evaluation records and supplemental student, teacher, and school information were collected by the Tennessee Department of Education (TNDOE) and processed for research purposes under a partnership between TNDOE and the Tennessee Consortium on Research, Evaluation, and Development at Peabody College, Vanderbilt University.

Construction of TN-VPK Experiment's Intensive Subsample Analytic Sample

Prior to the start of the 2009–2010 and 2010–2011 school years, two cohorts of more than 3,000 total children applied to TN-VPK programs in schools targeted for the RCT portion of the study conducted by PRI.[5] Students were offered admission off randomly ordered applicant lists at oversubscribed sites across the state. Students and their families also had to be willing to participate in the piece of the experimental evaluation that directly assessed students across school years. PRI has followed (and is continuing to follow) both groups of students into later grades, those who were randomly admitted to the TN-VPK program and those who were randomly left on the wait list, collecting state data on the whole sample and directly assessing a smaller subsample (referred to as the Intensive Subsample [ISS]). While the counterfactual is a "business as usual" control that makes no adjustments to account for parents who sought out alternative preschool programs, a parent survey revealed that more than half of the ISS analytic sample stayed at home with a parent or other caregiver during the pre-K year. Just more than 11% enrolled in Head Start, and another 15% had formal private child care.

For a child to be included in the ISS, which is the sample of students we focus on in the current study, the child had to (a) meet all eligibility criteria for participation in the experiment, (b) have parental consent to participate in the study, and (c) be assessed by PRI staff at least once during the pre-K or kindergarten school year.[6] In addition, the consented, eligible children on a school's applicant list were included in the ISS only if there were at least one consented, assessed child who participated in TN-VPK and at least one consented child that did not participate from that randomized applicant list. Furthermore, to construct the analytic sample, PRI restricted the analytic sample to the 1,076 students who were assessed at the end of pre-K. Students are treated as TN-VPK participants where administrative records indicate they attended a TN-VPK program, whether at their assigned school or at another location, for a minimum of 20 days. Conversely, control group nonparticipants were defined as students for whom administrative records do not indicate attendance of at least 20 academic days.

Primary Measures of Interest

The primary outcome variables of interest are generated from individual assessment data collected by the PRI to gauge the effectiveness of the TN-VPK program. Trained research staff assessed students in the fall and spring of the pre-K year and in the spring of each subsequent school year with a selection of subscales from the Woodcock–Johnson III Achievement Battery (WJ; Woodcock, McGrew, & Mather, 2001). The WJ battery is a standardized, longitudinally scaled assessment that is appropriate for a wide age range of test takers. Subtests used across years in the TN-VPK evaluation included two literacy subscales (Letter-Word Identification and Spelling), two language subscales (Oral Comprehension and Picture Vocabulary), and two math subscales (Applied Problems and Quantitative Concepts). A principal components factor analysis revealed that all the WJ scales were rather intercorrelated with high loadings on a single factor. The W-scores, a translation of the raw score which is not age-adjusted, on those scales were therefore averaged together to create a composite measure representing children's overall achievement in literacy, language, and math. Our main analysis uses the composite WJ score as the dependent variable, while secondary analyses present findings using the subscales.

The primary indicator variables of interest come from the statewide educator evaluation system. In Tennessee, annual evaluations differentiate teacher performance based on a composite teacher effectiveness rating score, which is calculated using individual and school-level student growth scores and achievement data as well as teacher observations for teachers in tested and untested subjects and grades.[7] For all teachers included in the current study, 40% of the evaluation is comprised of student achievement data—25% based on school- or districtwide student growth as measured by the Tennessee Value Added Assessment System and 15% based on alternative measures of student achievement approved by the State Board of Education and selected through joint agreement by the educator and evaluator. The remaining 60% of the overall evaluation scores is determined through qualitative measures including teacher observations, student perception surveys, personal conferences, and reviews of prior evaluations and work.

An individual teacher's effectiveness score can range from 0 to 500. TNDOE uses these raw scores to assign a teacher to one of five discrete performance categories. Denoting X as the teacher score, for all models teachers with $X < 200$ are categorized as *significantly below expectation* (Level 1), teachers with $200 \leq X < 275$ as *below expectation* (Level 2), teachers with $275 \leq X < 350$ as *at expectation* (Level 3), teachers with $350 \leq X < 425$ as *above expectation* (Level 4), and teachers with $X \geq 425$ as *significantly above expectation* (Level 5).[8] Ratings reports provided to teachers include the discrete rating but not the underlying score on the 0 to 500 scale. We use these discrete ratings to create a series of teaching quality indicators, used in separate analytic models, including (a) a set of binary indicators for teachers whose overall evaluation rating make them subject to different sanctions or rewards (e.g., above Level 3 rating or below, above Level 4 evaluation score, etc.) and (b) a categorical indicator which represents a teacher's specific final evaluation rating on the 1 to 5 scale.

There are a few concerns with regard to the matching of TN-VPK study student records to teacher evaluation data. First, because the teacher evaluation system was introduced in 2011–2012 school year, we lack kindergarten teacher evaluation records for students in the first experimental cohort who were already in first grade when the policy was introduced. For this reason we limit our primary analysis to the interaction of pre-K participation and first grade teaching quality. We do however present estimates using the second cohort to estimate the relationship between kindergarten teacher ratings and pre-K effect persistence, which did not show a significant interaction effect, a fact we attribute to the lack of variation in K-year teacher ratings (nearly 90% of students' kindergarten teachers were rated 4 or 5).

Second, matching students to teachers relies on the accuracy and completeness of administrative records of course enrollment and teacher evaluation. Of the 1,076 students in the two cohorts of the ISS analytic sample, 972 were matched to at least one first grade core subject teacher (defined as self-contained, reading, language arts, or math).[9] However, a number of students have multiple teachers over the course of a given year due to transfers within or across schools. Following the TNDOE's policy for assigning students to teachers for accountability purposes, in our primary analysis we apply a further restriction that students were taught by the assigned teacher for a minimum of 150 academic days, reducing the number of students with matched teachers to 897. Our final sample restriction is based on completeness of cognitive test scores and baseline student level covariates resulting in an analytic sample of 823 students.

Student-Level Characteristics and Baseline Group Equivalence

Baseline group equivalence is an important analytic and conceptual consideration in any group comparison, particularly given the consent rate issues in similar randomized studies. To that end, we first examined baseline differences on pretest scores and demographics (primarily from a parent questionnaire collected toward the beginning of the pre-K year) between the treatment and control groups. As evidenced by the unadjusted means for our analytic sample displayed in Table 1, students in the treated sample scored significantly higher on 3 out of the 6 baseline tests, were younger at the time of the pretest, were less likely to be Hispanic and nonnative English speakers, and had higher levels of parental education than the students in the control group. Nonparticipants also experienced a longer lag between the beginning of the pre-K year and their preassessments due to logistical challenges of administering tests to students who were not enrolled in school-based pre-K.

To adjust for these baseline differences, we use a multilevel logit propensity score as a covariate in all models.[10] Although this method results in two statistically similar comparison groups across all but 2 of the baseline variables (Table 1, column 5), the propensity score adjustment may still be inadequate to account for bias contributed by unobserved characteristics that may have resulted in differential participation rates and are the primary threat to internal validity (Rosenbaum & Rubin, 1983).

Teacher Characteristics

Because the primary focus of this study is the relationship between pre-K participation and subsequent teaching quality, it is also worth examining briefly the characteristics of the students' teachers, and any evidence of differential sorting associated with treatment or control status. The average first grade teacher of an experiment participant was a White female teacher, around 40 years of age, with roughly 12 years teaching experience, and an overall rating of 3.7, equating to a high *meets expectation* rating. There were no significant demographic differences between treatment and control group teachers. However, treatment students were slightly more likely to be taught by teachers rated Level 5, and less likely to be taught by teachers rated Level 4 by roughly 10 percentage points.[11] Notably, the propensity score adjustment for baseline imbalances in student characteristics, makes these differences statistically insignificant for the primary analytic sample, reducing concerns about sorting of students to different first grade teachers (Table 2, column 5).

Differential Attrition

Another potential source of bias is differential attrition. Since the ISS analytical sample attrition is addressed extensively by Lipsey and colleagues' (2013), and it appears to be relatively minor and remain balanced across the treatment and control groups (attrition is less than 5% and 4% for the treatment and control group, respectively), we simply present Figure 1, which is a visual consort diagram that shows where participants are excluded at various stages prior to the generation of the sample matched with TEAM teacher evaluation data. Overall, attrition in the ISS sample is of minimal concern as more than 95% of the sample was located and included at the end of each year, and attrition was not differential by experimental condition. Of the analytic sample, 897 students (823 with complete set of controls and cognitive scores) at the end of first grade were matched with TEAM teacher evaluation data (those students who were not enrolled in their first grade year could not have first grade TEAM scores).[12] However, it is worth noting, as discussed above, that low consent rates functionally present the same threats as conventional attrition, for which we ultimately attempt to mitigate through quasi-experimental strategies described in the section that follows.

Analytical Strategy

Our primary research question is: to what extent does early grade teaching quality moderate the persistence of pre-K effects? To inform this question, we focus on the interaction of pre-K participation and teaching quality as it correlates with student achievement, which we interpret as approximating the effect of early grade teaching quality on the persistence of pre-K effects. The moderating effect is identified using two primary configurations of overall teacher ratings: (a) a categorical indicator which represents a teacher's specific final evaluation rating on the 1 to 5 ordinal scale and (b) a set of binary indicators for comparing teachers who surpass a performance threshold (i.e., rated above 2, above 3, above 4) to those teachers who score below it.

Ideally we would derive this effect estimate from two overlapping experiments, where students were randomly assigned to pre-K and then randomly assigned to teachers with different quality ratings. However, baseline imbalances between "treatment" and "control" students combined with an inability to rule out the potential of students differentially sorting to teachers of varying effectiveness levels necessitate the use of an extensive list of control variables and a

TABLE 1

Student Differences at Baseline for Analytic Sample

	TN-VPK	Nonparticipants	Mean Diff.	*SE*	Adj. Mean Diff.	*SE*
Age at pretest[a]	1655.53	1682.38	−26.85*	11.63	0.71	13.58
Black	0.22	0.22	0.00	0.04	0.02	0.04
Hispanic	0.15	0.27	−0.12**	0.04	0.04	0.03
Male	0.48	0.47	0.01	0.04	0.01	0.05
Native English speaker	0.84	0.71	0.13*	0.05	−0.05	0.04
Library card[b]	0.89	0.87	0.02	0.06	−0.09	0.08
Newspapers[c]	0.37	0.37	−0.01	0.07	−0.04	0.09
Magazines[c]	0.28	0.26	0.02	0.04	−0.02	0.04
Mother’s ed[d]	2.11	1.96	0.15*	0.06	0.00	0.06
Working parents	1.26	1.25	0.01	0.07	0.02	0.06
Test interval	918.69	902.83	15.85**	3.32	−4.63*	2.19
Test lag	67.75	82.66	−14.91**	4.83	4.50*	1.91
Baseline scores on Woodcock–Johnson Achievement Battery						
WJ composite	395.01	391.95	0.17	0.12	−0.07	0.11
Letter Word	319.53	314.26	0.20*	0.10	0.06	0.10
Spelling	349.80	351.66	−0.07	0.09	−0.03	0.12
Picture Vocabulary	457.18	449.39	0.34*	0.13	−0.09	0.11
Oral Comprehension	444.41	441.04	0.21†	0.12	−0.05	0.13
Quantitative Concepts	407.77	407.52	0.02	0.10	−0.10	0.11
Applied Problems	391.37	387.85	0.13	0.13	−0.14	0.11
n	606	217				
N	823					

Note. Means of baseline characteristics for the analytic sample treatment and control groups are presented in the first two columns. Mean difference column significance based on regression *t* test with *SE* clustered at the randomization list level. Mean differences on baseline tests are reported as standardized effect sizes for continuity with later tables. Adjusted mean difference column includes control for propensity score for TN-VPK participation to account for baseline imbalances in student characteristics. Analytic sample defined by students with complete records on all covariates included in the primary model.

[a] Age at pre test is presented in years for the mean and days for the mean difference for clarity of interpretation.

[b] Scale from 1 to 3 for regularity of usage.

[c] Number of subscriptions form (0–4 or more) .

[d] Mother’s education is on a 4-point scale from *less than high school* to *more than associate’s*.

†*p* < .1. **p* < .05. ***p* < .01.

propensity score as covariate strategy to attempt to minimize the role of selection in what is ultimately a nonexperimental framework. Our primary ordinary least squares (OLS) regression model takes the following form,

$$Y_{ijt} = \beta_1 preK_{ijt} + \beta_2 rating_{ijt} + \beta_3 preK_{ijt} * rating_{ijt} + \beta_4 \omega_{it} + \lambda_i \beta_5 + \phi_i + \alpha_t + u_s \qquad (1)$$

where, Y_{ijt} is the composite WJ test score measure in first grade for student *i*, in classroom *j*, at time *t*. $preK_{ijt}$ is a binary indicator variable that takes a value of 1 if the student participated in TN-VPK (i.e., the treatment group) and 0 for all other arrangements (i.e., the control group). $rating_{ijt}$ is the moderating variable of interest, and takes the form of one of the two indicators of teaching quality described in the previous paragraph. The ω_{it} represents a control for students’ baseline cognitive composite test scores on the WJ, which was administered in the fall of the year students applied to participate in pre-K, and λ_i is a vector of the full set of available baseline student characteristics, including age at pretest, gender, race/ethnicity, parental education, number of parents that work, indicators for home literacy activities, the time lag between start of school and pretest, the interval between pretest and posttest, and whether English is a student’s primary language. We also include the estimated propensity score for the likelihood of a particular student participating in TN-VPK as expressed by ϕ_i. The α_t represents a cohort year fixed effect, and the u_s represents the student error term. In all reported results robust standard errors are clustered at the school randomization list level.[13]

The estimate on β_1 represents the main effect of TN-VPK participation on first grade test scores. The estimate on β_2 represents the average difference in achievement for control group students who have teachers with higher ratings in the year of the first grade test, while β_3 represents the average difference in the effect of TN-VPK participation for students

TABLE 2
Teacher Characteristics of Students in the Sample

	TN-VPK	Nonparticipants	Mean Diff.	*SE*	Adj. Mean Diff.	*SE*
First grade teacher characteristics (*n* = 823 students)						
Female	0.99	1	−0.01	0.01	−0.03**	0.01
Black	0.06	0.07	0.00	0.02	0.01	0.02
White	0.93	0.91	0.01	0.02	−0.01	0.03
Other race/ethnicity	0.01	0.02	−0.01	0.01	0.01	0.01
Years experience	12	11.78	0.22	0.78	−0.56	0.97
Salary	44703.43	45208.55	−505.13	635.10	−67.26	786.91
Age	40.86	41	−0.14	0.91	−0.78	1.11
Overall rating (1–5)	3.76	3.65	0.11	0.08	0.09	0.10
Level 5	0.33	0.23	0.10**	0.04	0.05	0.04
Level 4	0.27	0.36	−0.09*	0.04	−0.02	0.04
Level 3	0.24	0.24	−0.01	0.03	−0.03	0.04
Level 2	0.16	0.17	0.00	0.03	−0.01	0.04
n	606	217				
Kindergarten teacher characteristics (*n* = 600 students)						
Female	0.99	0.99	0.00	0.01	−0.01	0.02
Black	0.04	0.06	−0.03	0.02	−0.03	0.02
White	0.95	0.93	0.02	0.02	0.01	0.02
Other race/ethnicity	0.01	0.01	0.00	0.01	0.02	0.01
Years experience	13.16	12.62	0.54	0.93	0.18	1.26
Salary	43604.79	44098.45	−493.66	701.35	522.6	875.44
Age	40.75	41.52	−0.78	1.04	−0.43	1.50
Overall rating (1–5)	4.25	4.06	0.19*	0.08	0.03	0.13
Level 5	0.48	0.42	0.06	0.05	0.11†	0.06
Level 4	0.35	0.32	0.03	0.04	−0.14*	0.06
Level 3	0.13	0.18	−0.05	0.03	−0.02	0.05
Level 2	0.04	0.08	−0.04†	0.02	0.05	0.04
n	430	170				

Note. Mean teacher characteristics for the treatment and control groups of the analytic sample are presented in the first two columns. Mean difference column significance based on regression *t* test. Adjusted mean difference column includes control for propensity score for TN-VPK participation to account for baseline imbalances in student characteristics. Analytic sample defined by students with complete records on all covariates included in the primary model.
†$p < .1$. *$p < .05$. **$p < .01$.

with a higher quality teacher compared with those whose first grade teachers are rated lower. Here, we are most interested in the estimate on β_3 as it allows us to understand the extent to which first grade teaching quality moderates the persistence of pre-K effects.

Take, for example, when $rating_{ijt}$ equals 1 if a teacher's overall evaluation rating is a Level 5 (highly effective) and 0 if a teacher received an overall evaluation rating below 5 (exceeds expectation or lower). The impact of TN-VPK on the first grade test scores of students with a Level-5-rated teacher can be found by adding the estimate on the coefficient from the interaction term to those of the main effect of TN-VPK and the Level 5 rating indicator ($\beta_1 + \beta_2 + \beta_3$). TN-VPK effects on students with teachers who are not rated Level 5 are captured by the estimate on the β_1 coefficient. To estimate the first grade gap in achievement between TN-VPK participants and control students with Level-5-rated teachers, we can combine the estimates from the β_1 and β_3 coefficients.

For our subgroup analysis, we estimate a variant of Equation 1 that includes a three-way interaction. A difference-in-difference-in-difference estimand allows us to address our secondary research question: does the magnitude of the moderated effect of TN-VPK on first grade skills based on teaching quality vary based on a student's academic preparedness, specifically for students with low baseline cognitive scores and nonnative English speakers? The model mirrors that described above, but adds a three-way interaction term, interacting exposure to TN-VPK, first grade teaching quality, and the student characteristic of interest, to assess heterogeneity in response to pre-K and subsequent teaching quality.

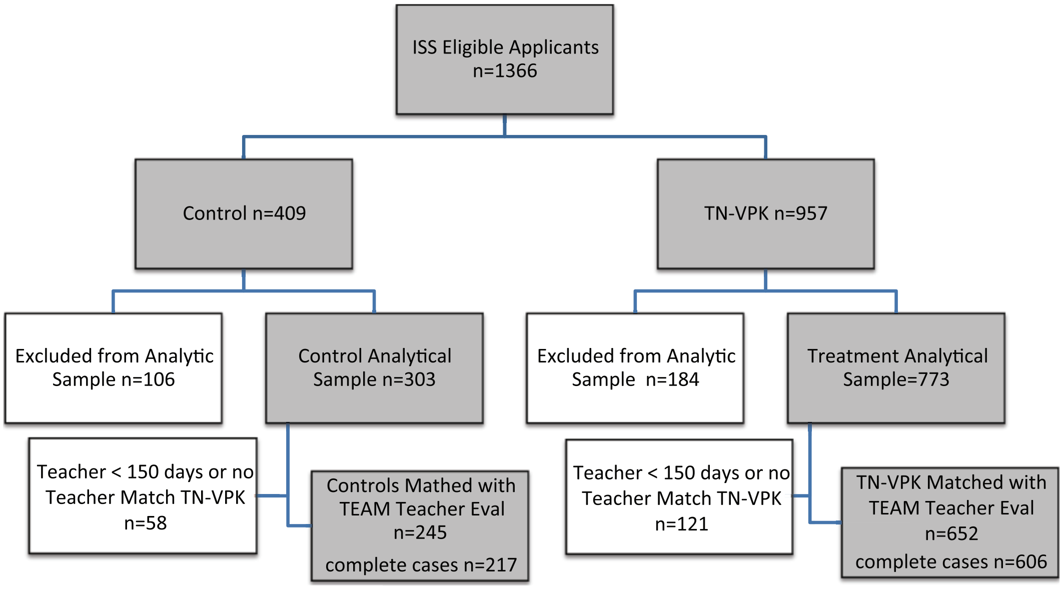
ISS Eligible Applicants n=1366
Control n=409
TN-VPK n=957
Excluded from Analytic Sample n=106
Control Analytical Sample n=303
Excluded from Analytic Sample n=184
Treatment Analytical Sample=773
Teacher < 150 days or no Teacher Match TN-VPK n=58
Controls Mathed with TEAM Teacher Eval n=245
complete cases n=217
Teacher < 150 days or no Teacher Match TN-VPK n=121
TN-VPK Matched with TEAM Teacher Eval n=652
complete cases n=606

FIGURE 1. *Consort diagram for both cohorts of pre-K study participants.*

Results

Our primary finding is that students who participated in TN-VPK consistently perform better in first grade with higher rated first grade teachers than students with similarly rated teachers who did not participate in TN-VPK. While differences are generally small in magnitude (effect sizes from .10 to .24) and somewhat sensitive to model specifications, the general result of a positive interaction between pre-K participation and first grade teacher ratings is markedly consistent. Figure 2 illustrates this relationship graphically, plotting the covariate adjusted mean WJ composite scores for “treatment” and “control” students over time based on whether or not they were ultimately taught by teachers who earned the highest overall rating. The gap between the lines for TN-VPK participants and similarly situated (by teacher rating) nonparticipants approximates the “pre-K effect.” Students in all groups exhibit convergence after the end of pre-K. However, in Level-5-rated classrooms, the TN-VPK students continue to slightly outperform control students in similarly rated classrooms by the end of first grade, while the pre-K participants are overtaken by the comparison group in all other classrooms.

Teaching Quality, Pre-K Participation, and Composite Cognitive Scores

We next quantify the positive interaction between first grade teacher ratings and TN-VPK using a series of related regression models, and again find that students who participated in TN-VPK consistently perform better in first grade with higher rated first grade teachers. As discussed above, we analyze separately the linear effect of steps up the teacher effectiveness rating scale, and binary indicators for each of the performance thresholds. Table 3 presents five distinct regression models and report estimates as standardized mean difference effect sizes. Model 1 presents the main effect of first grade teacher ratings using the 1 to 5 ordinal scale on students’ achievement scores, without accounting for pre-K participation but controlling for the full set of covariates, and shows that in general students with higher rated teachers perform roughly 0.06 standard deviations better for each 1-unit scale improvement in teacher rating.

Model 2 examines the moderating effect of the same overall teacher rating (1–5) on effect estimates for TN-VPK participation on first grade scores using the same propensity score and full baseline controls as covariates strategy discussed in Lipsey and colleagues (2013). The coefficient on overall teacher rating can now be interpreted as the achievement difference for control group students with higher rated teachers, which is now statistically indistinguishable from zero. The large negative coefficient on the pre-K indicator should be interpreted as the theoretical average performance of a pre-K student with a zero-rated teacher, and the positive coefficient on the interaction term indicates that for each improvement in teacher rating, students who participated in pre-K perform on average roughly 0.10 standard deviations better. This continuous model illuminates an important trend: teaching quality as measured by overall evaluation ratings appears to matter more for students who participated in pre-K, than those who did not. However, it is difficult to interpret practically, as it imposes a linear relationship on a teacher rating that is not normally distributed.

For ease of interpretation, Models 3–5 present a series of binary interactions comparing the effect of pre-K

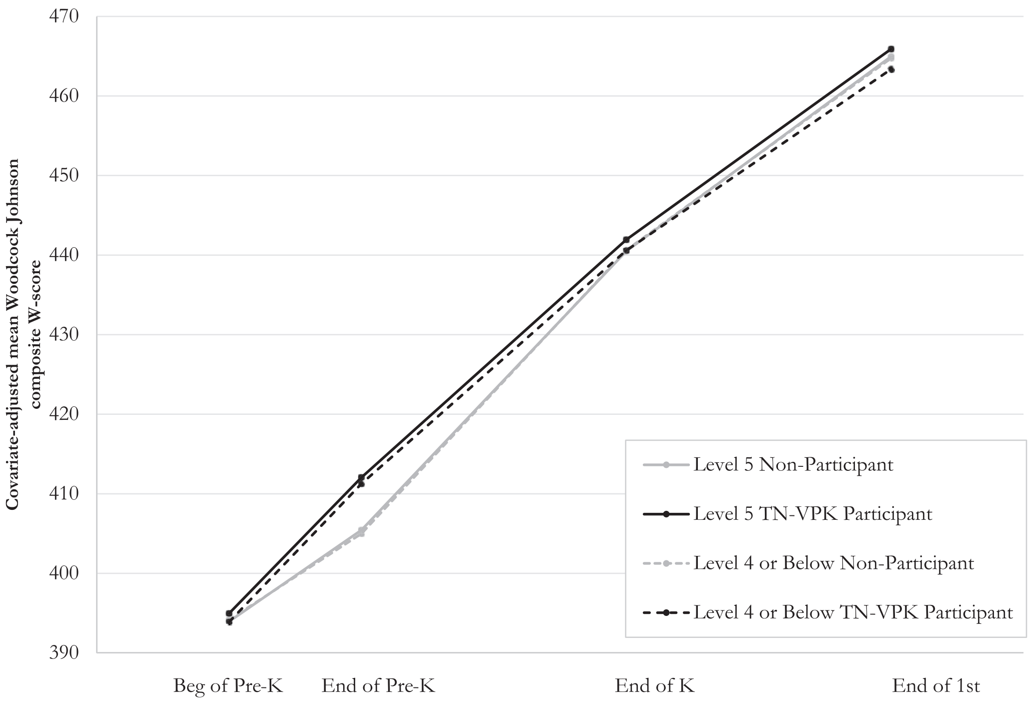

FIGURE 2. *Adjusted mean WJ composite W scores for TN-VPK vs. nonparticipants, by first grade teacher effectiveness ratings (Level 5 vs. non–Level 5 teacher effectiveness rating).*

participation on students taught by teachers on either side of a performance threshold. In each of the 3 comparisons, TN-VPK participants perform worse than nonparticipants with the lower rated teachers. In Model 3, which focuses on the comparison of low-rated teachers (1 and 2) compared to more highly rated teachers, none of the estimates are statistically significant, though patterns are consistent with those for the other cutoffs. In classrooms where teachers are rated below 4 (column 4 of Table 3), TN-VPK students on average score a statistically significant 0.17 standard deviations below their control group peers.

However, the positive coefficient on the interaction term for pre-K and above-3 indicates that pre-K participants perform 0.20 standard deviations (significant at the .10 level) better with teachers rated 4 or better. Notably, the difference in performance for the group across this teaching quality threshold is not large enough in magnitude to equate to persistence of a pre-K effect. Across all models, control group students consistently score no better with higher rated teachers than control students with lower rated teachers (coefficient is negative though statistically insignificant).

Column 5 shows a similar pattern of positive interaction between teacher rating and pre-K participation, with a slightly larger magnitude. Also, the negative effect or pre-K participation in the lower rated teacher group is smaller here, which should not be surprising as it now includes teachers rated as high as Level 4. Combining the relevant coefficients, pre-K students with teachers rated Level 5 on average outperform similarly situated control group student by roughly 0.12 standard deviations.[14]

Differences by Grade Level

While our primary results focus on the role of first grade teaching quality (where we have teacher ratings for both cohorts of the TN-VPK experiment) in relation to the persistence of pre-K effects, we also assess whether the same patterns hold for teachers of students in their kindergarten year. As stated above, because the TN teacher evaluation system was instituted after the first cohort of the study already completed kindergarten, any multiyear analysis is restricted to the second cohort. For simplicity, we focus here on the model estimating the interaction of TN-VPK exposure and the continuous (1–5) teacher rating, with the sample restricted to the analytic sample of cohort 2 (*n* = 430 treatment, 170 control). Notably, the interaction between first grade teacher ratings and TN-VPK exposure is qualitatively similar to that of the combined cohort estimate. The second and third columns of Table 4 shows that with each step up the teacher rating scale, students who went to TN-VPK perform 0.10 standard deviations better than their control group peers, when we account for K-year teacher ratings (column 3) and their interaction with pre-K participation (column 2). However, the interaction with kindergarten teacher ratings is statistically zero when predicting end of K cognitive scores (column 1).

TABLE 3

First Grade Teacher Quality × Pre-K Interactions on Woodcock–Johnson (Both Cohorts)

	(1)	(2)	(3)	(4)	(5)
Overall rating (1–5)	0.06*	−0.02			
	(0.03)	(0.04)			
Rating × pre-K		0.11*			
		(0.05)			
Pre-K		−0.45*	−0.17	−0.17*	−0.12†
		(0.17)	(0.12)	(0.08)	(0.07)
Teacher rated > 2			−0.03		
			(0.09)		
Pre-K × rated > 2			0.15		
			(0.13)		
Teacher rated > 3				−0.01	
				(0.09)	
Pre-K × rated > 3				0.20†	
				(0.11)	
Teacher rated > 4					−0.06
					(0.11)
Pre-K × rated > 4					0.24†
					(0.14)
N	823	823	823	823	823

Note. Coefficients represent differences in standardized WJ composite W-scores at the end of first grade. All models include the full set of baseline covariates including the composite pretest and propensity score. The first column presents the main effect of first grade teacher ratings, column 2 examines the interaction of the continuous rating (1–5) with pre-K participation, columns 3–5 present the interaction for pre-K participation and a series of binary indicators for teacher rating thresholds. Robust standard errors clustered at the r-list level are presented in parentheses.

$^{\dagger}p < .1$. $*p < .05$.

One plausible explanation for this apparent discrepancy is that the principal ratings of kindergarten teachers could have less to do with instructional techniques that might promote larger cognitive gains than their ratings of first grade teachers who are closer to a tested grade year. To test this hypothesis we constructed a crude value-added-like measure, predicting students' test score gains based on their teacher rating. In their first grade year, students taught by higher rated teachers made significantly larger gains on the composite cognitive measure (more than 0.5 standard deviations per step up the rating scale). However, a student's kindergarten teacher rating had no discernable relationship to the their cognitive test score gains in the kindergarten year.

Differences by Student Academic Preparedness

One potential mechanism by which teaching quality might moderate the effects of pre-K participation is by serving to counteract home-based barriers to academic success. Students who had particularly low cognitive scores prior to enrollment in pre-K and students with limited English language backgrounds could be particularly receptive to the combination of preschool and subsequent higher quality instruction. To test these hypotheses, we examine whether the relationship between pre-K participation and teacher ratings differs based on students baseline cognitive scores or English language background through separate three-way interaction models. Tables 5 and 6 present the results of separate regressions mirroring the form presented above but including a three-way interaction term for binary indicators for a student being a nonnative English speaker or scoring in the bottom quartile of the sample on the baseline composite WJ battery.[15] The three-way interaction terms for each model show a consistent and large, significant trend. For both groups, students with lower relative baseline cognitive scores and students who do not speak English as a native language, the interaction between teaching quality and pre-K participation is particularly important (effect size roughly 0.5 for Level 4 teachers and 0.9 for Level 5).

While the number of students in the subsets that drive these three-way interactions is admittedly small (210 and 158 for low-baseline test score and nonnative English speakers, respectively), the large magnitudes of the coefficients are indicative of something more than a chance relationship. These estimates indicate that one possible explanation for the positive interaction between teacher ratings and pre-K participation is that TN-VPK participation helps elevate students' capacity to benefit from high-quality instruction by

TABLE 4

Teacher Rating × Pre-K Interaction on WJ Composite at K and First Grade (Cohort 2 Only)

Outcome	End of K WJ Composite	First Grade WJ	First Grade WJ
Pre-K	−0.004	−0.504*	−0.414[†]
	(0.229)	(0.238)	(0.208)
Teacher rating K year	0.002	−0.017	−0.007
	(0.039)	(0.031)	(0.040)
Teacher rating K × pre-K	0.016	0.022	
	(0.049)	(0.045)	
Teacher rating first grade		0.001	0.003
		(0.053)	(0.050)
Teacher rating first grade × pre-K		0.098	0.098[†]
		(0.060)	(0.057)
N	600	600	600

Note. Coefficients represent differences in standardized WJ composite W-scores at the end of kindergarten (column 1) and first grade (columns 2–3). All models include the full set of baseline covariates including the composite pre-test and propensity score. The first column presents the main effect of kindergarten teacher ratings on end-of-kindergarten test scores. Column 2 examines the interaction of kindergarten and first grade teachers' continuous rating (1–5) with pre-K participation; column 3 replicates the primary model but adds a control for kindergarten teacher rating. Robust standard errors clustered at the r-list level are presented in parentheses.

[†]$p < .1$. *$p < .05$.

mitigating problematic baseline deficits. One might also argue that early grade teaching quality moderates the persistence of pre-K gains most for students whose home environments might otherwise counteract the benefits of their early childhood investment.

Sensitivity Analysis—Alternative Modeling Strategies, Sample Compositions, and Tests for Positive Sorting

We conducted a series of analyses to explore whether our findings are sensitive to alternative modeling strategies and sample composition, or subject to bias in the form of positive sorting. First, to account for the role of potential interclass correlations, we replicated the above analyses using a three-level hierarchical linear model with random intercepts at the school randomized list and district levels. Like the primary models, we include the full set of covariates and propensity scores. Results do not differ substantially from those presented in our primary OLS models with clustered standard errors (see Online Appendix E). We also estimate school fixed effect models that include dummy variables for each school that students attend in first grade, making coefficients interpretable as driven by within-school variation (Online Appendix F), and randomization list fixed effects that could be interpreted as within neighborhood estimates, as students generally applied to programs proximate to their homes (Online Appendix F). Although significance levels vary based on the amount of variance left in the comparison group (e.g., within school, within neighborhood), the magnitudes of effect size estimates are qualitatively similar across all modeling strategies.

Next, we examined if our estimates were sensitive to alternative sample composition. Our primary analytic sample includes all students, for whom the full set of cognitive measures and student level covariates are available and that were accurately matched with an evaluated teacher three years after pre-K application (first grade for 95% of students). The question of how to address the role of students' kindergarten grade-level retention is complicated by the fact that the retentions can be understood as intermediate pre-K effects. Thus, adding an indicator for "on grade level" to the model would introduce bias to the estimates of pre-K effects. However, because the pre-K participants were less likely to be retained in their kindergarten year than students in the control group as detailed by Lipsey, Weiland, Yoshikawa, Wilson, and Hofer (2015), we might be concerned that teacher evaluations for the control group in the third year of the study were more often reflecting ratings of kindergarten teachers, who might be evaluated differently, for example receiving higher ratings for an emphasis of lower-level concepts in their instruction. To simplify the construct of "first grade teaching quality," we re-create our primary estimates with an alternative sample limited to students who are on grade level, and thus taught by first grade teachers. Estimates from this reduced sample (Online Appendix C) are slightly smaller in magnitude and lose some statistical significance, but are qualitatively similar to those from the more inclusive preferred model.

On the other end of the inclusiveness spectrum, we estimate the same models presented throughout the article with missing data imputed for the full 1,076 students included in the Lipsey et al.'s ISS analytic sample. Online Appendices

TABLE 5

Three-Way Interaction With Indicator for Low (Bottom Quartile) Student Baseline Scores

	> Level 3	> Level 4
Pre-K	−0.08	−0.05
	(0.10)	(0.07)
Low baseline WJ	0.32*	0.21†
	(0.15)	(0.12)
Pre-K × low baseline	−0.34†	−0.21
	(0.17)	(0.14)
Teacher rated > 3	0.12	
	(0.10)	
Pre-K × rated > 3	0.08	
	(0.12)	
Rated > 3 × low baseline	−0.46*	
	(0.20)	
Pre-K × rated > 3 × low baseline	0.47*	
	(0.22)	
Teacher rated > 4		0.18†
		(0.10)
Pre-K × rated > 4		0.03
		(0.12)
Rated > 4 × low baseline		−0.97**
		(0.27)
Pre-K × rated > 4 × low baseline		0.85**
		(0.29)
N	823	823

Note. Coefficients represent differences in standardized WJ composite W-scores at the end of first grade. All models include the full set of baseline covariates including the composite pretest and propensity score. The columns present the three-way interaction among pre-K participation, low baseline cognitive scores (bottom quartile), and a series of binary indicators for teacher rating thresholds. Robust standard errors clustered at the r-list level are presented in parentheses.

$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$.

TABLE 6

Three-Way Interaction With Indicator for Nonnative English Speakers

	(Nonnative English)	(Nonnative English)
	(> Level 3)	(> Level 4)
Pre-K	−0.08	−0.07
	(0.08)	(0.07)
Teacher rated > 3	0.08	
	(0.09)	
Pre-K × rated > 3	0.06	
	(0.11)	
ELL	0.22	0.23†
	(0.15)	(0.12)
Pre-K × ELL	−0.26	−0.09
	(0.18)	(0.14)
ELL × rated > 3	−0.31†	
	(0.17)	
Pre-K × rated > 3 × ELL	0.54*	
	(0.21)	
Teacher rated > 4		0.16†
		(0.09)
Pre-K × rated > 4		0.01
		(0.11)
ELL × rated > 4		−0.93**
		(0.26)
Pre-K × rated > 4 × ELL		0.92**
		(0.31)
N	823	823

Note. Coefficients represent differences in standardized WJ composite W-scores at the end of first grade. All models include the full set of baseline covariates including the composite pretest and propensity score. The columns present the three-way interaction among pre-K participation, nonnative English speaker (ELL), and a series of binary indicators for teacher rating thresholds. Robust standard errors clustered at the r-list level are presented in parentheses.

$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$.

G–I present results from a multiple imputation strategy using multivariate normal regression techniques. We replace missing values with multiple sets of simulated values to construct an analysis file, apply standard analyses to each completed dataset, and adjust the obtained parameter estimates for missing-data uncertainty (Marchenko, 2010; Rubin, 1987). Here, the objective is not to predict missing values as close as possible to the true ones but to handle missing data in a way resulting in valid statistical inference. Given that fewer than 10% of cases had missing data for any given analytic variable, it is not surprising that results were qualitatively similar to analyses that were restricted to complete cases.

Next, we checked for the possibility of students differentially sorting into classrooms which could introduce bias into the estimated interaction terms that are the focus of this study. Using the continuous form of the students' teacher rating as the outcome variable, we run a series of models with the same set of covariates describe above to check for an interaction between student scores the previous year and pre-K participation status when predicting the rating of the teacher to whom the student will be assigned. If the coefficient on the interaction were significant, it would elevate concerns that higher-scoring pre-K students were assigned to higher-rated teachers more frequently than their similarly situated control group peers. However, the fact that these coefficients are statistically insignificant is consistent with the argument that, after controlling for the observed characteristics, preschool exposure does not alter the nature of the sorting into classrooms. The small, though statistically significant, coefficient on test scores from the end of pre-K indicates that there may be some positive sorting of students into classrooms, where higher scoring students end up with higher-rated teachers. However, the near zero interaction between the test score and pre-K participation when predicting first grade or average 2-year teaching quality indicates that the sorting mechanisms are similar for both treatment and control students (see Online Appendix J).

To further allay concerns that positive interactions with first grade teaching quality simply highlight differences in pre-K effects among students that were derived prior to their placement with higher or lower rated teachers, we also model a simple falsification test, where we replicate our primary models, but substitute the outcome of kindergarten cognitive scores. Because these exams were administered prior to any interaction with the first grade teachers, significant interaction terms would be indicative of a false attribution of causality (though we acknowledge the inability to draw causal conclusions). None of the coefficients in the model are statistically significant. However, the direction of the relationship is consistent with our primary findings, allowing for the possibility of an underlying trend that preceded the subsequent significant relationship (Online Appendix K).

Discussion

The results reveal some important patterns. The relationship between TN-VPK participation and first grade teaching quality appears to matter. All else equal, students who participate in TN-VPK appear to consistently perform better with higher rated teachers than nonparticipants. This positive interaction between first grade teaching quality and pre-K participation also seems to be most important for students who entered pre-K scoring in the bottom portion of the sample or with limited English skills. One possible explanation for these findings is that the high-rated teachers in our study have adjusted their teaching emphasis to account for higher levels of preparedness due to the expansion of TN-VPK in recent years, or have received training that emphasized more challenging early academic curriculum in response to accountability policies. To the extent that the teacher teaches to the full group, this elevated content emphasis could have negative effects on the control students but benefit students who participated in pre-K.

Recent work by Engel, Claessens, and Finch (2013) demonstrated a negative association between overemphasis of low-level math concepts and student achievement. Similarly, if poorly rated teachers struggle with classroom management, the disruptive environment could result in a convergence of cognitive measures, as pre-K students fail to build on earlier gains. If the higher-rated first grade teachers in our study had stronger classroom management skills or spent more time on challenging concepts and tasks, both of which are factors in the Tennessee observation rubric, it is plausible that such an environment would facilitate greater learning among the pre-K participants than their control group peers, who do not appear to benefit measurably from placement in higher rated classrooms.

The positive interaction between TN-VPK participation and teaching quality could also be interpreted as TN-VPK preparing students (particularly those with language barriers or more pronounced early cognitive deficits) to benefit more from high-quality teachers. If cognitive scaffolding (e.g., Berk & Winsler, 1995) from TN-VPK helped prepare students who might have otherwise struggled with literacy, language, and math, it is not surprising that we see large positive interactions between teaching quality and pre-K participation in these areas.

It is also worth noting that while our primary measure of teaching quality (an overall rating from the consequential statewide evaluation system) has the merits of being more objective than self reports and being composed of multiple measures, it is still subject to the threat of a considerable amount of measurement error in capturing the elusive construct of "high teaching quality." Any noise in the measure, or competing elements (e.g., high-level instruction vs. differentiation) in the construct of "good teaching" itself, have the potential to attenuate estimated effects toward zero. Moving forward, research should further examine alternative and more specified metrics of teaching practices and their interaction with students' earliest educational experiences.

The question at the core of this study is the intersection between two of the most prominent aims of contemporary education policy—expanding access to preschool learning opportunities and access to high-quality teachers for traditionally disadvantaged students. Our findings are consistent with the understanding that either intervention without the other is inefficient, if not all together inadequate. When we invest public dollars in providing preschool to students from impoverished homes, the longitudinal test-score effects fade rapidly in elementary classrooms with poor quality teachers. Alternatively, students who have experienced the TN preschool program, tend to benefit more from high-quality teachers than they otherwise would. These findings suggest that policy makers interested in maximizing the cognitive impacts of preschool initiatives should work to ensure access to high-quality early grade teachers. This could be facilitated through policies that discourage school administrators from shifting experienced or effective teachers to later tested grades or through the institution of recruitment and retention bonuses for talented early grade teachers in hard-to-staff schools.[16]

As the students from the Tennessee Pre-K experiment progress into tested grade areas, our analysis will expand to estimate cognitive effects for the full sample (approximately 3 times the size and not subject to bias from differential consent rates) and to include teacher test score value-added measures in our construct of teaching quality. Also, the inclusion of additional years of teacher evaluation data will enable more rigorous techniques to account for the potential bias contributed by nonrandom sorting of students into classrooms, which (along with differential consent rates) poses a substantial barrier to causal interpretations of the present results. Furthermore, as students progress to grades where documented disciplinary actions or absenteeism are more common, we can explore the interactions of teaching quality measures and some proxies for the noncognitive effects, which Chetty and colleagues (2011) have shown to better predict long-term impacts. If the findings presented here hold, we can begin to establish a working theory around the ingredients necessary for measures of preschool cognitive gains to persist.

Acknowledgments

We appreciate helpful comments and suggestions from Dale Ballou, Amy Claessens, Dale Farran, Mimi Engel, Gary Henry, and Mark Lipsey, as well as seminar participants at the annual meetings of the Association for Education Finance and Policy, Association for Public Policy Analysis and Management, and Society for Research on Educational Effectiveness. We would also like to acknowledge the many individuals at the PRI, the Tennessee Consortium, and the Tennessee Department of Education for providing data and expert insight to conduct our analyses, in particular Carol Bilbrey, Janie Hughart, Chris Ham, Rick Feldser, Ilknur Sekmen, Susan Burns, Trish Kelly, Erin O'Hara, Art Peng, Matthew Pepper, and Nate Schwartz, as well as assessors across the state. Any errors remain the sole responsibility of the authors. The views expressed in this article do not necessarily reflect those of sponsoring agencies or individuals acknowledged. Kerry G. Hofer was senior research associate at Peabody Research Institute during the completion of this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Peabody Research Institute (PRI) and the Tennessee Consortium on Research, Evaluation, and Development (the Consortium), both of which are located at Vanderbilt University's Peabody College. PRI conducted the Tennessee-Voluntary Pre-K experiment with funding from the Institute of Education Sciences (Grant R305E090009). The consortium is funded by the state of Tennessee's Race to the Top grant from the U.S. Department of Education (Grant S395A100032).

Notes

1. The Tennessee Board of Education requires that each classroom have a teacher licensed with an early childhood endorsement, a student-to-adult ratio of no more than 10:1, and a maximum class size of 20, among other requirements.

2. McClellen and Donoghue (2014) found reliability (Cronbach's alpha) of roughly .92, indicating a high degree of internal consistency.

3. A sample distribution of the continuous teacher composite score and performance category thresholds for the state in 2012 is included as Online Appendix B.

4. A number of studies have documented the important long-term predictive power of early math and reading cognitive scores from school entry (Duncan et al., 2007), kindergarten (Claessens, Duncan, & Engel, 2009), and first grade (Watts et al., in press). Researchers have highlighted a series of mechanisms by which these early cognitive measures might influence long-term outcomes, including in school academic ability tracking (e.g., Bui, Craig, & Imberman, 2011), development of executive functioning (Clark, Pritchard, & Woodward, 2010), and self-concept of ability (e.g., Marsh, Byrne, & Yeung, 1999).

5. For a more detailed description than what is provided below of the randomization process, site selections, survey items, achievement measures, and construction of the Intensive Subsample that forms the foundation for this analysis, see Lipsey and colleagues' two extensive 2013 reports on the underlying experiment and main effects analyses.

6. To be eligible for inclusion in the randomized sample, children had to meet certain criteria: (a) assigned to either Tennessee's Voluntary Pre-Kindergarten Program (TN-VPK) or the control condition on the basis of their position on the randomized applicant list (i.e., not automatically let in because of sibling preference, etc.), (b) age eligible (age and subsequent grade progression indicated that the child was old enough to attend kindergarten the next school year), (c) income eligible (based on the exclusion of those children who did not meet standard for the free or reduced-price lunch program, and (d) placed in a regular TN-VPK classroom (not a blended or special education classroom.

7. As of July 2011, the Tennessee State Board of Education approved four teacher evaluation models—the Tennessee Educator Acceleration Model (TEAM), Project Coach, Teacher Effectiveness Measure (TEM), and Teacher Instructional Growth for Effectiveness and Results (TIGER). Although implementation is quite different from one model to the next, the evaluation models all follow the requirements set forth by Tennessee's Teacher Effectiveness Advisory Committee and adopted by the State Board of Education and have the same goals—to monitor teacher performance and encourage teacher development. During the 2012–2013 school year, more than 80% of teachers across Tennessee used TEAM as their evaluation model, while TEM is the second most frequently used (11%), followed by Project COACH (5%) and TIGER (2%) (Ehlert et al., 2013). In our analytic sample, only a small number of first grade teachers were evaluated under models other than TEAM, which we account for by including dummy variables for alternate assessment tools in our analytic models.

8. Since the inception of Tennessee's teacher evaluation system, a teacher's performance has been legislatively tied to various consequences and rewards. As part of the state's effort to win favor under the federal Race to the Top competition, the teacher evaluation system was designed to inform human capital decisions, including but not limited to individual and group professional development plans, hiring, assignment and promotion, tenure and dismissal, and compensation. In fact, Section 11 of the Tennessee First to the Top Act of 2010 directly stipulates that provisions within a teacher's contract shall provide for consequences when performance standards are not met. In addition, if performance standards are met or exceeded, the performance contract may also

provide for bonuses beyond base salary (Tennessee First to the Top Act of 2010).In a later bill passed in 2011, the Tennessee General Assembly voted to explicitly tie evaluation scores to tenure decisions (Tennessee Senate, 2011). Under this law, tenure eligibility was limited to new teachers who receive an overall performance rating of *above expectations* or *significantly above expectations* (the highest two categories on the state's 5-point scale) during the final 2 years of the 5-year probationary period. Teachers who do not receive tenured status at the end of their 5-year probationary period may either be rehired under a year-to-year contract or dismissed, while teachers who have tenure may also be reverted to probationary status if they receive one of the lowest performance ratings (*below expectations* or *significantly below expectations*) for two consecutive years. Furthermore, new regulations made it possible that low evaluation scores may be used as an example of "inefficiency" in a tenured teacher's dismissal proceedings.

9. All but two students who were not matched with teacher evaluations were no longer present in the Tennessee public school enrollment files. For simplicity, we refer to the teachers, with whom students were matched 3 years after pre-K application, as first grade teachers. However, just fewer than 5% of the students in the analytic sample were retained between pre-K and first grade. Thus, the measure for teaching quality for that small subset of students is actually associated with kindergarten teachers. Estimates for an alternate sample, excluding these students, are presented in Online Appendix C. Findings are consistent regardless of sample decision rules.

10. We found that the propensity score as covariate approach was more effective than alternative weighting and matching methods at balancing baseline covariates. We followed the approach adopted by Lipsey and colleagues (2013) where they also offer an extensive comparison of alternative specifications. Results are qualitatively similar when we simply include the full set of controls in regression analysis.

11. For the purpose of our analyses, teachers who are rated Level 1 are grouped with Level 2 teachers, due to their extremely low numbers (only 15 students in first grade taught by Level 1 and 3 in kindergarten) and the fact that consequences for being rated 1 and 2 were identical at this point in the development of the Tennessee teacher evaluation system.

12. For consistency, we also present estimates where missing data are imputed to match the 1,076 analytic sample, which was established by the research team in the primary TN-VPK early outcomes reports. This approach is discussed in greater detail later in the article.

13. In an alternative specification with school fixed effects, results are consistent in magnitude with those presented in this article, though sometimes less statistically significant for specific binary interactions. This is not surprising due to the relatively small samples within schools, which range from 2 to 50 students per randomized list.

14. To further explore the interaction between early grade teaching quality and pre-K effects, we estimate the same series of models as above but use the subscale scores from the WJ cognitive assessment as the dependent variables of interest. This allows us to determine whether teaching quality and pre-K interactions are more pronounced in different curricular areas. Across the six measures for which we have pre- and postdata (i.e., Letter Word Recognition, Spelling, Oral Comprehension, Picture Vocabulary, Applied Problems, and Quantitative Concepts), interaction terms for teaching quality tend to follow a fairly consistent pattern and are statistically significant for at least one teaching quality indicator for four of the six WJ cognitive assessment subscales. Results of these models are shown in Online Appendices D and H.

15. While the two indicators are positively correlated (.45) and represent similar constructs, they are worth treating separately, including controls for the other to demonstrate the consistent pattern across language backgrounds and to promote further research on the potential for specific needs of nonnative English speaker students.

16. For a study documenting sorting of highly effective teachers in response to No Child Left Behind, see Chingos and West (2011). For studies on recruitment and retention bonus programs, see Springer, Swain, and Rodriguez (2014) and Clotfelter, Glennie, Ladd, and Vigdor (2008).

References

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *Future of Children*, *5*(3), 25–50. doi:10.2307/1602366

Bassok, D. (2010). Do Black and Hispanic children benefit more from preschool? Understanding differences in preschool effects across racial groups. *Child Development*, *81*(6), 1828–1845.

Bassok, D., Gibbs, C. R., & Latham, S. (2015). *Do the effects of early childhood interventions systematically fade? Exploring variation in the persistence of preschool effects*. Retrieved from http://curry.virginia.edu/uploads/resourceLibrary/36_Preschool_Fade_Out.pdf

Bassok, D., & Loeb, S. (2015). Early childhood and the achievement gap. In H. F. Ladd & M. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 510–528). New York, NY: Routledge.

Berk, L. E., & Winsler, A. (1995). *Scaffolding children's learning: Vygotsky and early childhood education* (NAEYC Research into Practice Series Vol. 7, NAEYC 146). Washington, DC: National Association for the Education of Young Children.

Bui, S. A., Craig, S. G., & Imberman, S. A. (2011). *Is gifted education a bright idea? Assessing the impact of gifted and talented programs on achievement* (w17089). Washington, DC: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17089

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., . . . Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology*, *48*, 1033–1043. doi:10.1037/a0026644

Campbell, F. A., Wasik, B. H., Pungello, E., Burchinal, M., Barbarin, O., Kainz, K., . . . Ramey, C. T. (2008). Young adult outcomes of the Abecedarian and CARE early childhood educational interventions. *Early Childhood Research Quarterly*, *23*, 452–466. doi:10.1016/j.ecresq.2008.03.003

Chingos, M. M., & West, M. R. (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review*, *30*(3), 419–433.

Claessens, A., Duncan, G. J., & Engel, M. (2009). Kindergarten skills and fifth grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*(4), 415–427.

Claessens, A., Engel, M., & Curran, F. C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal*, *51*, 403–434.

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology*, *46*(5), 1176–1191.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5), 1352–1370.

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *Journal of Human Resources*, *35*(4), 755–774. doi:10.2307/146372

Dalmia, S., & Snell, L. (2008, August 22). Protect our kids from preschool. *Wall Street Journal*, A15.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*, 111–134. doi:10.1257/app.1.3.111

Doggett, L., & Wat, A. (2010). Why preK for all? *Phi Delta Kappan*, *92*(3), 8–11.

Duncan, G. J., Bailey, D., & Yu, W. (2015, March). *Fadeout in human capital interventions: Death, miracles, and resurrection*. Paper presented at the annual conference of the Society for Research on Educational Effectiveness, Washington, DC.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1466.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis)alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*(2), 157–178.

Garces, E., Currie, J., & Thomas, D. (2002). Longer-term effects of Head Start. *American Economic Review*, *92*, 999–1012. doi:10.1257/00028280260344560

Gormley, W. T. (2008). The effects of Oklahoma's pre-K program on Hispanic children. *Social Science Quarterly*, *89*(4), 916–936. doi:10.1111/j.1540-6237.2008.00591.x

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, *41*(6), 872–884. doi:10.1037/0012-1649.41.6.872

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*, 114–128. doi:10.1016/j.jpubeco.2009.11.001

Hustedt, J. T., & Barnett, W. S. (2005). Head Start's lasting benefits. *Infants and Young Children*, *18*, 16–24. doi:10.1097/00001163-200501000-00003

Hustedt, J. T., & Barnett, W. S. (2011). Financing early childhood education programs: State, federal, and local issues. *Educational Policy*, *25*, 167–191. doi:10.1177/0895904810386605

Lee, V. E., & Loeb, S. (1995). Where do Head Start attendees end up? One reason why preschool effects fade out. *Educational Evaluation and Policy Analysis*, *17*, 62–82. doi:10.3102/01623737017001062

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and first grade follow-up results from the randomized control design* (Research report). Nashville, TN: Vanderbilt University, Peabody Research Institute.

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression discontinuity design: Methodological issues and implications for application. *Education Evaluation and Policy Analysis*, *37*, 296–313.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*(1), 159–208.

Marchenko, Y. (2010, July). *Multiple-imputation analysis using Stata's mi command*. Paper presented at the Stata Conference, Boston, MA.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007a). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*(1), 33–51.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007b). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly*, *22*(1), 18–38.

Marsh, H. W., Byrne, B. M., & Yeung, A. S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist*, *34*(3), 155–167.

McClellen, C. A., & Donoghue, J. R. (2014). *Evaluation of TEAM rubric*. Nashville: Tennessee Consortium on Research, Evaluation, and Development. Working paper available at http://www.tnconsortium.org/data/files/gallery/ContentGallery/Evaluation_of_TEAM_Rubric_Report.pdf

Nores, M., Belfield, C. R., Barnett, W. S., & Schweinhart, L. (2005). Updating the economic impacts of the High/Scope Perry Preschool Program. *Educational Evaluation and Policy Analysis*, *27*, 245–261. doi:10.3102/01623737027003245

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2014). *Effective teacher retention bonuses: Evidence from Tennessee*. Nashville: Tennessee Consortium on Research, Evaluation, and Development. Working paper available at http://www.tnconsortium.org/data/files/gallery/ContentGallery/Effective_Teacher_Retention_Bonuses_Evidence_from_TN.pdf

Tennessee First to the Top Act of 2010. Tenn. Code Ann. § 49-1 (2010).

Tennessee Senate, Bill No.1528, Public Chapter 70 (2011).

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., . . . Susperreguy, M. I. (in press). The role of mediators in the development of longitudinal achievement associations in mathematics and reading. *Child Development*.

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.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson Tests of Cognitive Abilities–III*. Itasca, IL: Riverside.

Authors

WALKER A. SWAIN is a PhD student in Leadership and Policy Studies at Vanderbilt University's Peabody College of Education and Human Development. His research focuses on the impacts of education and social policy on traditionally disadvantaged populations.

MATTHEW G. SPRINGER, PhD, is an assistant professor of public policy and education at Peabody College of Vanderbilt University and the director of the National Center on Performance Incentives. His research focuses on incentives, accountability, and compensation.

KERRY G. HOFER is an associate/scientist at Abt Associates, Inc. She holds a PhD in teaching and learning with a focus on research methodology from Vanderbilt University. The majority of her research involves rigorous evaluations of early educational programs and practices.