ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-28

A Reanalysis of Impacts of the Tennessee Voluntary Prekindergarten Program

Tyler W. Watts
New York University

Greg J. Duncan
University of California
Irvine

Mariela Rivas
University of California
Irvine

We present a reanalysis of the Tennessee Voluntary Prekindergarten Program (TNVPK), a state-funded program designed to promote the school readiness of 4-year-olds from low-income families. Oversubscribed programs used a lottery to randomly assign prospective enrollees a chance to attend TNVPK. We found that assignment to the program had largely null effects on measures of behavior, attendance, and retention collected during elementary school. TNVPK increased enrollment in special education by 4% between kindergarten and grade 3, and generated negative but generally statistically insignificant impacts on third-grade state test scores. We explore reasons for fadeout as well as threats to internal validity.

VERSION: May 2019

Suggested citation: Watts, T.W., Duncan, G.J., & Rivas, M. (2019). A Reanalysis of Impacts of the Tennessee Voluntary Prekindergarten Program (EdWorkingPaper No.19-28). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-28

A Reanalysis of Impacts of the Tennessee Voluntary Prekindergarten Program

BY TYLER W. WATTS*, GREG J. DUNCAN+, AND MARIELA RIVAS+

We present a reanalysis of the Tennessee Voluntary Prekindergarten Program (TNVPK), a state-funded program designed to promote the school readiness of 4-year-olds from low-income families. Oversubscribed programs used a lottery to randomly assign prospective enrollees a chance to attend TNVPK. We found that assignment to the program had largely null effects on measures of behavior, attendance, and retention collected during elementary school. TNVPK increased enrollment in special education by 4% between kindergarten and grade 3, and generated negative but generally statistically insignificant impacts on third-grade state test scores. We explore reasons for fadeout as well as threats to internal validity.

* Steinhardt School of Culture, Education and Human Development, New York University, 627 Broadway, 8th Floor, New York, NY, 10003 (e-mail: tyler.watts@nyu.edu).

+ School of Education, University of California, Irvine, 3200 Education Drive, Irvine, CA 92697-5000

Acknowledgement

This paper developed from a collaboration between the authors and the Principal Investigators of the TNVPK evaluation study – Mark Lipsey and Dale Farran – who were supported by Grant #R305E090009 from the Institute of Education Sciences, U. S. Department of Education. In an attempt to provide an independent replication analysis, Lipsey and Farran have shared their data with us. Our communication with Lipsey and Farran has only involved clarification questions regarding the data and study procedures. We would like to thank Lipsey, Farran, and their coauthor, Kelley Durkin, for openly cooperating with us on this project and supporting our many inquiries regarding the data and study procedures throughout our analytic effort. We would also like to thank Martha Bailey, Steve Barnett, Damon Clark, Kerry Hofer, Bill Gormley, Jade Jenkins, Cybele Raver, Diane Schanzenbach, and Christina Weiland for their helpful contributions.

I. Introduction

Impressed by the striking long-run advantages enjoyed by participants in the Perry Preschool and Abecedarian program and in the Head Start program in its early years (see Elango, Garcia, Heckman, & Hojman, 2015 for review), a growing number of cities and states have begun offering prekindergarten programs to substantial numbers of four-year-olds.[1] In some cases, classrooms are restricted to children from low-income families (e.g., Tennessee); in other cases, access is universal (e.g., New York City). However, few studies have used experimental methods to evaluate the end-of-program and, especially, early-grade impacts of these recently developed public prekindergarten programs.

A recent review by Phillips et al. (2017) of prekindergarten (pre-k) research lists 30 studies of program effects on outcomes at the end of preschool or the beginning of kindergarten and 39 studies of program effects on outcomes measured beyond that point. In the vast majority of cases, the end-of-program impact estimates are based either on regression discontinuities (RD) created by birthday cutoffs in program eligibility (e.g., Gormley, Gayer, Phillips, & Dawson, 2005; Weiland & Yoshikawa, 2013) or on propensity-score matching methods (e.g., Huang, Invernizzi, & Drake, 2012).

Because birthdate-cutoff RD studies rely on comparisons between children who just received preschool and children who will soon enter preschool, these studies cannot estimate program impacts past the end of pre-k. Consequently, virtually all of the estimates of medium- or long-term impacts in the Phillips et al. (2017) study are based on propensity-score matching methods that rarely include matching variables measured prior to the beginning of the pre-k year. Thus, the

[1] State-funded prekindergarten programs enrolled 1.3 million 4-year-old children in 2017, which accounted for approximately one-third of the country's population of 4-year olds (Friedman-Kraus, Barnett, Weisenfeld, Kasmin, DiCrecchio, & Horowitz, 2018).

question of whether public pre-k programs generate long-lasting impacts on children's outcomes remains unanswered.

In contrast to the previous quasi-experimental and correlational work, evaluation evidence on the Tennessee Voluntary Prekindergarten Program (TNVPK) adds to the literature base in several noteworthy ways. First, TNVPK constitutes a large, state-funded pre-k program that annually enrolls approximately 18,000 children across the state (i.e., about 20-25% of the state's 4-year-olds). Second, unlike existing studies, the evaluation of TNVPK relied on random assignment of children to program slots through the implementation of a lottery procedure.[2] Finally, the TNVPK evaluation collected longitudinal follow-up data on children who participated in the lottery through the end of third grade.

A. Previous TNVPK Evaluation Studies

The effects of TNVPK on child outcomes have been reported by several previous evaluation efforts, most of which did not rely on experimental variation. However, these previous studies deserve some mention, as they motivated the evaluation effort reported here.

After the state began expanding pre-k access in 2005, the Tennessee Office of Education Accountability commissioned an external evaluation of the pre-k program's effectiveness. This lead to a number of non-experimental studies that relied on state administrative data to compare pre-k participants to various comparison groups matched on a handful of demographic characteristics. This early work was reviewed in a report by the Strategic Research Group (SRG, 2011), and in general, these initial studies found slight advantages for pre-k attendees on state achievement tests taken in kindergarten and first grade. These

[2] This design has also been used to study high-performing charter schools (e.g., Angrist, Dynarski, Kane, Pathak, & Walters, 2010).

initial advantages were typically gone by second grade, with some tests showing negative effects of pre-k participation by the end of elementary school.

These evaluation efforts were followed by a more rigorous approach led by researchers at Vanderbilt University, who worked in partnership with the TN Department of Education (TNDOE).[3] They designed an experimental study that would randomly assign children to TNVPK through a lottery system implemented at over-subscribed programs. However, their first report (Lipsey, Farran & Hofer, 2015) was limited to a sub-sample of consented children who participated in the randomized study and used propensity scores to match TNVPK participants with non-participants. They reported positive effects of TNVPK enrollment on end-of-preschool cognitive skills, but also found that these effects had all but disappeared by the end of kindergarten and first grade. Surprisingly, they found negative effects of attending TNVPK on second- and third-grade cognitive measures, and on some measures of behavior.

The second report released by the Vanderbilt team (Lipsey, Farran & Durkin, 2018) relied on the lottery design to generate estimates of program impacts, and they reported effects on student outcomes that were similar to the findings reported in the non-experimental work (i.e., positive initial effects, followed by null and negative impacts). However, even with the most recent experimental work, Lipsey and colleagues did not employ program evaluation techniques commonly found in the economics literature. For example, although the lottery was conducted within each participating TNVPK program site, the Lipsey et al. analyses did not control for fixed effects based on the unit of random assignment and instead relied on a hierarchical linear modelling (HLM) approach with random effects that is more typical of research in psychology and education.

[3] Their evaluation study was supported by Grant #R305E090009 from the Institute of Education Sciences, U. S. Department of Education.

Consequently, it remains unclear whether the effects reported in the Lipsey et al. papers might be due to between-site differences that should be controlled in analyses of treatment impacts. Further, Lipsey et al. (2018) defined the "treatment" group in a manner that allowed for the inclusion of students who enrolled in alternative TNVPK sites (i.e., any TNVPK center other than the original over-subscribed site for which they enrolled and were randomly assigned), a choice that could potentially introduce selection factors that might influence program impact estimates. Finally, potential issues regarding attrition and non-compliance deserve further attention from an econometric perspective.

B. Current Study

The current paper constitutes an independent reanalysis of the TNVPK evaluation, and provides new estimates of the impact of the TNVPK program on measures of student academic achievement and behavior. Relying primarily on administrative data measuring student outcomes between kindergarten and grade 3, we find that the offer to attend TNVPK had null effects on indicators of retention, attendance, disciplinary offenses, and placement into gifted and talented programs. However, we also find that TNVPK boosted the probability of placement into special education during elementary school by 4 percentage points, and we find a negative, but only at the margin of statistical significance, effect of the program on third-grade test scores, amounting to -0.08 SDs. Examinations of potential threats to validity suggest that differential rates of attrition and baseline differences between the treatment and control groups had negligible effects on impact estimates. Moreover, analyses of site heterogeneity suggest that site-specific treatment effects were generally normally distributed around the average treatment effect reported in our primary models.

In the final sections of the paper, we leverage detailed measures of cognitive ability and behavior collected on a subsample of students participating

in the study to delve into possible mechanisms to explain the null to negative longer-run effects reported here. With these post-hoc analyses, we investigate whether the temporal pattern of program effects on test scores might be characterized by fadeout, and we examine impacts of the program on non-cognitive skills.

II. Study Design

The Tennessee Voluntary Prekindergarten Program (TNVPK) offers state-funded prekindergarten for children who meet the following eligibility criteria: 1) children must be 4 years old by August 15 prior to the beginning of the new school year and cannot be age-eligible for kindergarten (i.e., five years old) by that date; and 2) children must reside within the area served by the school district. Provided that these criteria are met, priority enrollment is granted to children who qualify for free or reduced-price lunch (FRPL). If space allows, non-FRPL qualifying children who have disabilities, are English Language Learners, are in state custody, or are educationally at risk of failure (as defined by the Individuals with Disabilities Education Act; U.S.C. § 1400 et sec) may enroll as well.[4]

Local districts apply to the TN DOE for funding for preschool classrooms, and may house these classrooms in a number of settings, including elementary schools and child care centers that are limited to preschool classes. By conventional standards, the quality of the program is high: it meets 9 of the 10 National Institute for Early Education Research benchmarks,[5] it is a full-day program with relatively small class sizes (maximum of 20 students), and it

[4] Eligibility criteria are described in detail in the Tennessee Code Annotated (TCA) 49-6-101—104 (found online at https://law.justia.com/codes/tennessee/2010/title-49/chapter-6/part-1/49-6-101/).

[5] These benchmarks were updated in 2016, after the initial findings were released for the evaluation considered here, to include more process-oriented markers of quality. Based on the updated benchmarks, the TNVPK program meets 5 out of 10 NIEER standards (found online at http://nieer.org/state-preschool-yearbooks/yearbook2016).

requires a licensed teacher and teacher aide in each classroom (this keeps the child-to-adult ratio at or below 10 to 1). The curriculum used in the classroom must be chosen from a state-approved list.[6] The program is large; as of 2016, it cost $85 million dollars per year and included 935 classrooms and 18,000 students.[7]

The data for the current study were collected through a partnership between the Peabody Research Institute at Vanderbilt University and the Tennessee DOE. Random assignment to TNVPK was accomplished by randomly ordering application lists as well as using data on actual enrollment to determine the cut point on the application lists that distinguished those assigned and not assigned slots at any given site. Specifically, the TN Office of Early Learning (OEL) recruited TNVPK sites (i.e., centers or elementary schools offering TNVPK) for participation in the study that regularly had more eligible applicants than available enrollment spots and encouraged them to use a lottery procedure to allocate enrollment offers.

Participating sites sent application lists to Vanderbilt, where study personnel then randomly ordered student names and returned the resulting list to the participating TNVPK sites.[8] Staff at the study TNVPK sites then began with names at the top of the list and contacted the families of students to offer admission. They continued down the list until they filled the available number of

[6] The list for the 2018-2019 school year can be found online (https://www.tn.gov/content/dam/tn/education/early-learning/pre-k/prek_approved_curricula.pdf)
[7] See "History of Voluntary Pre-K" section of the TN DOE Voluntary Pre-K website: https://www.tn.gov/education/early-learning/voluntary-pre-k.html
[8] Initially, Vanderbilt randomly ordered all names that appeared on any given list, regardless of eligibility status. Most sites screened children and families for eligibility before allowing parents to place their child's name on an application list. However, 27 children who were later determined to be ineligible (see below for more details) appeared on randomized applicant lists sent to Vanderbilt. Although Vanderbilt did not remove these children from the randomization process (i.e., they were offered a chance to enroll if their name appeared at the top of the list), Vanderbilt researchers disregarded them from study participation. In other words, only children eligible for the program were included in the study sample.

spots at that site. If a contacted family declined the offer of admission, or if TNVPK site staff could not successfully contact the families of eligible students, then the next student on the list was offered a chance at enrollment. Once all enrollment slots were filled, the remaining children on the enrollment lists were placed a waitlist ordered by Vanderbilt's initial randomization. If a child whose family initially accepted an offer failed to enroll when TNVPK classes began, then TNVPK site staff contacted the next-in-line waitlisted child to offer the empty program slot.

For analysis purposes, Vanderbilt researchers determined whether a student had been assigned to TNVPK (hereafter referred to as the "treatment" group) or placed on the waitlist (hereafter referred to as the "control" group) after enrollment data from TNDOE became available. At this point, researchers excluded students from inclusion in the study if their names appeared on an "invalid" randomized list, or if they were determined to be ineligible for TNVPK. Of the 150 randomized applicant lists, 39 (comprised of 483 children) were determined to be invalid because either the list did not produce a control group or the site did not use the list to determine enrollment. Further, researchers disregarded 18 children who applied to enroll in a "blended" preschool program (i.e., a program that included children aged 3 to 5) and 19 children who were either randomized along with a sibling or labeled as exempt from randomization by a site (this was typically done for a relative of someone working in a TNVPK site).

Finally, Vanderbilt researchers also disregarded 27 children from inclusion in the evaluation study who apparently did not meet eligibility requirements for TNVPK (i.e., age, FRPL status or other marker of disadvantage

listed above).[9] After disregarding invalid lists and ineligible students, they counted down from the top of each list until they reached the number corresponding to the actual number of eligible students who enrolled in a given center. For example, if TN DOE records showed that "Center A" enrolled 20 eligible children, then, regardless of actual TNVPK enrollment status, the first 20 eligible children on the randomized applicant list for center A were assigned treatment group status and, also regardless of TNVPK enrollment status, the rest of the children on the randomized applicant list were placed in the control group.

Although this process describes the treatment status indicator used in the current study, it should be noted that Vanderbilt researchers generated an alternative treatment status indicator that was used for impact estimation in the previous evaluation reports (i.e., Lipsey et al., 2018).[10] This alternative indicator took into account students who enrolled in other TNVPK sites by increasing the size of the treatment group to adjust for enrollment slots that were apparently available to children on randomized lists at alternative TNVPK sites. We used our assignment indicator because it adopts the strictest interpretation of the opportunity to enroll in TNVPK based on one's ordered spot on the randomized

[9] In effect, this procedure limited participation in the study to children eligible for TNVPK. Vanderbilt researchers used multiple sources of information to determine eligibility, including the enrollment applications collected by study sites and administrative records made available by the TNDOE. Twenty-five children were considered ineligible owing to age, and two children were disregarded because of family income (i.e., these children had family income over the FRPL limit and did not meet any of the other eligibility criteria).

[10] To generate this indicator, researchers followed the same basic process described above, but moved the cutline between the treatment and waitlist groups down if students on the list had enrolled in other TNVPK sites. For example, if "Center A" enrolled 20 eligible children, but 5 children on the "Center A" randomized list enrolled in other TNVPK sites, then the cutpoint was adjusted and the first 25 students on the list were considered part of the treatment group (rather than arriving at a treatment status group of 20 following the procedure described above).

list, and we examine enrollment in other TNVPK sites as an issue of non-compliance in the analyses that follow.

This sample selection procedure, illustrated in Figure 1, produced a study sample of children attending TNVPK in 79 sites operating in 29 school districts. These sites were not recruited to be representative of the entire population of TNVPK programs, but they were drawn from both rural and urban areas. Lipsey et al. (2018) present descriptive information comparing study sites and the larger TNVPK population and find more similarities than differences.[11]

As shown in Figure 1, sites participating in the study produced 111 valid randomized applicant lists that included 3,131 TNVPK-eligible children. Study evaluators were able to locate TN DOE records for 2,990 of these children, who comprise the analytic sample for the current paper. The randomization procedure was conducted for two cohorts participating in successive years: Cohort 1 in the 2009-2010 school year (58% of the analysis sample; n = 1,744) and Cohort 2 in the 2010-2011 school year (42% of the analysis sample; n = 1,246).[12]

Table 1 provides preschool attendance information for the two groups; 1,614 children were listed above the enrollment cutoff on their respective randomized applicant list and were considered part of the treatment group, and 1,376 fell below the cutoff and were considered part of the control group.

[11] Lipsey et al. (2018) demonstrate that the sample of oversubscribed sites included in the current study were distributed geographically across TN, with sites sampled in urban, suburban and rural locations. However, sites operating the Nashville area were slightly overrepresented in the current study. Children included in the current study were broadly similar to the population on an array of demographic characteristics observed, with the exception being that the current sample was less likely to be White (49% in the current sample vs. 60% in the population) and more likely to be Hispanic (22% in the current sample vs. 9% in the population).

[12] Of the 79 participating sites, 29 participated in both cohorts of the study. Consequently, these 29 sites generated 2 unique randomized applicant lists that correspond to each year of participation in the study. Three sites generated 2 lists during one year of participation; this occurred because enrollment was determined on a rolling basis. Thus, for these three sites, Vanderbilt randomly ordered the first enrollment list produced by the site, then a second enrollment list as more spots became available.

Compliance in the treatment group was high, with 89% of treatment children attending at least 20 days. However, non-compliance in the control group was substantial; 40% (n = 550) of control-group children attended at least 20 days of TNVPK, with the majority of them (n = 363) enrolling in a TNVPK study site. The relatively low rate of compliance in the control group presents a potential challenge to the interpretation of results, as many parents of waitlisted children found alternative ways to enroll their children in preschool. In the following analyses, we present both intent-to-treat (ITT) and local average treatment effect (LATE) estimates, the latter of which were generated using random assignment as an instrument for actual attendance.

A. Baseline Measures

We first determined whether the lottery procedure produced observationally equivalent groups, based on baseline measures taken from TN DOE administrative records at the time of preschool enrollment (Table 2). The p-values shown in the third and fourth columns were generated from regression models that adjusted standard errors for TNVPK site clustering. The first set of p-values was generated from bivariate regression models, whereas the second set was generated from regressions that included randomized applicant-list (r-list) fixed effects (g = 111).[13] The models that included r-list fixed effects adjust for the actual design of the study; they control for differences between sites and between their respective enrollment lists, that should not affect results because random assignment occurred *within* sites. Unadjusted differences in the demographic characteristics of students between the two groups often appear to

[13] In our analyses, we control for randomized applicant-list fixed effects (g = 111), because random assignment was conditional on the specific list in which a given child appeared. However, as we detail in the results section, we also tested models that instead used TNVPK-site fixed effects (g = 79), and results were nearly identical to models that included randomized applicant-list fixed effects.

be quite substantial (e.g., 54% of students in the "assigned to TNVPK" group were white, compared with 43% in the waitlisted group), yet these discrepancies disappeared when fixed effects were included. Thus, statistically significant differences from unadjusted p-values reflect differences between r-lists and sites in the characteristics of student applicants coupled with variation in the probability of enrollment across applicant lists.

Table 2 also presents results from joint F-tests assessing whether the entire set of baseline characteristics differed between the treatment and control groups. In these models, we regressed the treatment status indicator on the entire set of baseline characteristics shown in Table 1, and we found little indication that the set of characteristics differed across the study conditions. For the model that included r-list fixed effects, the F-test produced a p-value of 0.907. In contrast, the model without fixed effects produced a statistically significant result (p = 0.012), again indicating that differences at baseline appear more substantial when the clustered design of the study is not taken into account. The tests that included adjustments for r-list fixed effects provide confidence that the treatment and control groups were balanced on the few demographic measures available at baseline, but of course not cannot assess balance across unobserved characteristics (e.g., academic ability). In the analyses that follow, we evaluate baseline balance by presenting results that did and did not include baseline covariates, and we also detail results for a subsample of students that included a larger set of baseline covariates, including tests of cognitive ability.

B. Outcome Measures and Sample Inclusion

Our key outcome measures for TNVPK were taken from TN DOE administrative records when students were enrolled in kindergarten through grade 3. Outcome measures available for each grade included special-education placement, gifted and talented placement, referrals for serious disciplinary

offenses, absences, and retention. Our measure of absences was a continuous variable that represented the total number of days a child was recorded absent during a given school year, and the other administrative outcome measures were binary indicators of whether an event had occurred (e.g., special-education placement in kindergarten is coded "1" if a student was recommended for special education during the kindergarten year). We considered each outcome separately for each respective wave of data between kindergarten and grade 3, and we also summed outcomes over the entire elementary school period. For the binary indicators, the outcome summation variable was coded to "1" if a student was ever designated for each outcome between kindergarten and grade 3 and "0" otherwise. For absences, the outcome summation was the total number of absences over this same period.

Children also completed a series of state achievement tests in mathematics, reading, and science in grade 3.[14] Because these tests were all highly correlated with one another (r ranged from 0.71 to 0.75), and to avoid multiple testing bias, we averaged scores to create an achievement test score composite. In the supplementary material, we present treatment impacts on the disaggregated test scores and found that these impacts did not differ substantially across subject areas.

In our analyses of each respective wave, we included in the sample only the students who had non-missing data on all outcome measures for the given wave. For the sake of simplicity, we labeled each wave according to the grade level in which students would have enrolled if they had been on track (i.e., not retained). However, the actual measurement points correspond to academic years,

[14] These tests are part of the Tennessee Comprehensive Assessment Program (TCAP), which was introduced in 1988. See the "History and Current Landscape of Assessment in Tennessee" section of the 2015 TN DOE report *Tennessee Task Force on Student Testing and Assessment* for an overview of the TN state assessment program at the time during which our data were collected (https://www.tn.gov/content/dam/tn/education/testing/tst_assessment_task_force_report.pdf).

and we included students in each wave who were retained in a previous grade. For example, the first-grade wave represents the 2010-2011 academic year for cohort 1, and children who were retained in kindergarten but had non-missing data for that academic year were still included in our grade-1 wave. Because of limitations in our state administrative dataset, we had access only to third-grade test-score measures and retention data for students who were not retained in a previous grade. However, we found few indications that treatment status affected retention rates across our various models, and we also describe analyses below that were generated using the grade-4 wave of data for children in the first cohort, and these estimates included children who had been retained and were missing grade-3 wave data as a result. Further, we also tested models for our set of administrative outcomes that excluded children who had ever been retained, and results were nearly identical to our main specifications (Appendix Table S1). Finally, the "outcome summation" wave included students who had at least one non-missing measure for each outcome (n = 2,925); this provided us with a set of outcome measures that included 98% of children who had participated in the preschool lottery.

Table 3 displays sample means on the set of outcome measures at each wave, as well as p-values reflecting treatment and control group differences generated by regressing each respective outcome measure on treatment status. These regressions include neither baseline covariates nor r-list fixed effects. A few descriptive values from Table 3 are worth noting. First, special-education placement rates were high, especially among children in the treatment group, with 21% of the treatment group and 14% of the control group placed in special education at least once between kindergarten and grade 3 (p = 0.003). In contrast, only 1% of treatment and control children were ever designated for gifted and talented programming (p = 0.499). Approximately 12% of treatment children and 10% of control children had been retained by the end of the grade-3 wave (p =

0.106), and most of this retention occurred in either kindergarten or first grade. Overall, Table 3 shows little indication of differences between the treatment and control groups on measures of disciplinary offenses, absences, or our third-grade test score composite.

Table 4 shows treatment- and control-group attrition across our various follow-up waves, where "attrition" for a given wave is defined as lacking complete administrative data for that wave (see Table 3 for a list of each measure included in each wave). We observed relatively low rates of attrition for kindergarten (approximately 4%), grade 1 (approximately 5%), and grade 2 (approximately 7%), and found no indication that these rates differed between children who were and were not assigned to TNVPK. The grade-3 attrition rate was higher (19%), primarily owing to the inclusion of test scores, but we again saw little indication that the third-grade attrition rate differed between the two groups ($p = 0.165$). We return to the issue of attrition in Section IV and assess whether our treatment impact estimates might have been affected by attrition related to the baseline characteristics of students who were likely to leave the sample.

III. Treatment Impacts

Table 5 presents our key TNVPK-impact estimates – intent-to-treat (ITT) estimates generated from OLS regressions with cluster-adjusted standard errors (adjusted at the site level, $g = 79$) using the Huber-White estimator in Stata 15.0. For each outcome, we relied on two sets of model specifications. The first included only the treatment status indicator and r-list-level fixed effects ($g = 111$; effectively controlling for site and cohort), and the second included both fixed effects and the baseline covariates shown in Table 2. We standardized the test-score composite measure and our various measures of absences using control-group means and standard deviations. We used linear probability models to generate treatment impacts for the remaining binary measures. In Table 5, we

present results only for the treatment status variable; coefficients and standard errors for the baseline covariates are provided in supplementary material Table S2.

Estimated impacts were highly consistent across models that did and did not include baseline covariates, which suggests that baseline imbalance does not pose a serious threat to validity. Focusing on the models that included the full set of controls, we found that students assigned to TNVPK slots were consistently more likely to be placed in special education between kindergarten and third grade. By the spring of kindergarten, children assigned to TNVPK had a 3.2 percentage-point higher placement rate than control-group children (the control-group placement rate for kindergarten was 7.4%), and this effect remained relatively consistent through grade 3. Across all the grades considered, we found that treatment children were 4.4 percentage points more likely to be placed in special education between kindergarten and third grade.

For gifted and talented placement, we found negative, small, and statistically non-significant point estimates, and estimated program impacts on measures of serious disciplinary offenses, retention, and student absences were also largely null.

We found a negative and marginally statistically significant effect of assignment to TNVPK on our grade-3 state test-score composite measure of -0.082 SD (*SE*= 0.045). In the supplementary material (Table S3), we present test-score impacts for the disaggregated math, reading, and science tests, and show that the negative and marginally significant impacts were consistent across these three domains (math: β = -0.084, *SE*= 0.044; reading: β = -0.066, *SE*= 0.043; science: β = -0.070, *SE*= 0.046).

These largely null, and possibly adverse, effects of assignment to the TNVPK program are surprising. In the following sections, we investigate possible threats to the validity of the results (i.e., non-compliance, attrition, site

heterogeneity) and we also examine whether poor program quality, fadeout, or negative impacts on children's non-cognitive skills are responsible for the null effects.

IV. Threats to Validity

A. Non-Compliance

As shown in Table 1, the TNVPK study encountered substantial non-compliance in the control group, as 41% of children assigned to the waitlist found their way into a TNVPK classroom (supplementary information Figure S1 presents the attendance distribution patterns for children in the study). To estimate the impact of actually attending TNVPK, we turned to 2SLS models, using the offer of TNVPK enrollment as an instrument for TNVPK attendance.

The TN DOE sets 20 days as the threshold for full enrollment, and we found that the lottery-generated offer of TNVPK attendance strongly predicted whether a child attended 20 days or more (β= 0.392, *SE*= 0.034, t = 12.88). In the third column of Table 6, we present 2SLS estimates of the impact on the third-grade test-score composite, as well as the "outcome summation" measures, of attending TNVPK for at least 20 days.[15] For purposes of comparison, we also present the corresponding ITT estimates from Table 5 as well as OLS estimates that were generated by regressing each respective outcome measures on a dummy for "whether attended 20 days or more" and baseline controls. All models included r-list fixed effects and site-cluster-adjusted standard errors.

As expected, the 2SLS models effectively scaled up the ITT effects (and corresponding standard errors) shown in Table 5. Attending at least 20 days of TNVPK had a marginally statistically significant negative impact of -0.207 SD

[15] In results available upon request, we tested different thresholds for attendance, including 50 or more days, 100 or more days, or 120 or more days. We found substantial overlap between these groups (i.e., students who attended at least 20 days tended to attend the full year) and 2SLS results were largely similar across the various cutoff levels.

(*SE* = 0.119) on the third-grade test-score composite and increased the probability of placement in special education between K and grade 3 by about 11 percentage points (*SE* = .046). These estimates suggest that the local average treatment effect (Angrist & Pischke, 2009) of attending TNVPK on student test scores was negative and non-trivial, and that TNVPK attendance led to higher rates of placement in special education. The LATE interpretation suggests that the TNVPK program lowered third-grade test scores for children who were induced to attend preschool based on winning an offer of attendance in the lottery. It should be noted, however, that all children in the sample had some desire to attend the program, since everyone enrolled in the study had initially applied for admission to a TNVPK center.

Across our other outcome measures, we again found largely null results for 2SLS models. Somewhat surprisingly, the naïve OLS model results shown in the second column found no effect of attendance on the third-grade test-score composite, an interesting result in light of the large number of correlational studies showing that non-random attendance in preschool positively predicts later test scores, especially for disadvantaged students (e.g., Magnuson, Ruhm & Waldfogel, 2007; Vandell, Belsky, Burchinal, Steinberg, & Vandergift, 2010). Nevertheless, the collection of models in Table 6 suggests that TNVPK attendance among this sample had null or possibly negative effects on achievement and other school outcomes.

B. Attrition

Although we found little evidence of statistically significant differential attrition between the treatment and control group at each wave, we pursued further checks to ensure that non-selective attrition from the sample did not substantially bias outcomes. Even if the share of students who left the sample did not differ between the treatment and control groups, estimates might be biased if

the characteristics of students who left the treatment group differed from the characteristics of those who left the control group.

To test this question, we ran a series of logistic regression models in which we modeled the probability of remaining in the sample at each outcome wave as a function of treatment status, r-list fixed effects, and baseline covariates. To see a graphical distribution of the predicted likelihoods of attrition for both treatment and control students at each wave, see supplementary information file Figure S2. After generating a predicted probability of attrition for each student at each wave, re-ran our treatment-impact models with weights for the inverse of the probability of remaining in the sample. These "attrition-adjusted" models effectively weight up students in the model who remained in the sample but whose baseline characteristics appeared to be similar to those of students who left. Table 7 presents both the ITT estimates taken from Table 5 and the attrition-adjusted estimates that included the inverse probability weights. Across Table 7, we saw little indication that attrition substantially influenced our ITT models. For the third-grade test-score effect, the ITT estimate with no adjustment for attrition was -0.082 SD (*SE* = 0.045) and the estimate produced by the weighted model was -0.073 SD (*SE*= 0.047).

C. Site Heterogeneity

Because students were placed on a randomized applicant list (r-list) within each TNVPK site, we rely on models that include r-list fixed effects for our primary ITT estimates (recall that 50 sites generated a single r-list, while 29 sites participated in both cohorts of the study, thus generating multiple r-lists). In essence, these models estimate the unique treatment effect for each site r-list; an average effect across the r-lists is then calculated by weighting up the r-lists by the number of students. If a few large sites had treatment effects that were different from those of the rest of the sites in the study, then this weighted average

might be a misleading indicator of the treatment effect that was common across most sites. Similarly, a handful of small sites with either very positive or very negative impacts might skew the composite ITT estimate.

To investigate heterogeneity based on sites and r-lists, we first estimated r-list-level treatment impacts for our key outcomes: the composite third-grade state-achievement test score and the "outcome summation" measures shown in the right-hand column of Table 5. These treatment effects (displayed in supplementary materials Table S4), which did not take into account the number of students on each r-list, largely resembled the effects shown in Table 5. The r-list-level TNVPK effect on third-grade test scores was smaller (-0.036 SDs) and the effect on special-education placement rate between K and third grade was 4%. The exception was the estimated effect on total absences from kindergarten through third grade, as the r-list level effect doubled the individual level estimate shown in Table 5 (β = 0.119).

These r-list-level estimates effectively treat as equal r-lists that included smaller and larger numbers of students, and these models suggest that our main ITT estimates were not skewed by outlier impacts found for r-lists that included the largest numbers of students. To examine whether small r-lists might have skewed our ITT estimates, we plotted each r-list's unique treatment effect against the sample size for each list. In Figure 2, we display this plot for the third-grade achievement-test score (plots for the other administrative outcomes are available upon request). For the third grade test score composite, we observed an approximately normal distribution of r-list effects centered around our ITT point estimate of -0.082, providing little indication that our effects were driven by small r-list outliers.

Finally, because some sites generated multiple r-lists, we also tested whether estimates might differ if we adjusted for site fixed effects, rather than r-

list fixed effects. These results, shown in supplementary Table S5, were nearly identical to the main estimates shown in Table 5.

V. Extensions

A. Effects Beyond Grade 3

The null and negative effects estimated through grade 3 are concerning, and raise questions as to how the impacts of the TNVPK program will continue to unfold as children progress in school. Because TN does not start statewide testing until grade 3, it is unclear whether the negative test score impacts are indicative of a longer pattern of poor academic performance following exposure to the preschool program, or if the third-grade effect was merely a chance estimate that may not be representative of impacts in other periods. Our data do not contain measures of adolescent or adult functioning. However, we did obtain administrative records for the fourth-grade year for children in cohort 1. Because these records were available for only one cohort, we place less emphasis on these results, but they do offer insight into how impacts might unfold in later periods.

We first tested whether our earlier kindergarten-through-grade 3 ITT treatment impacts were consistent across cohorts, and we found no statistically significant interactions between cohort status and the treatment group indicator (cohort interactions are shown in supplementary materials Table S6). However, we found that the third grade test score impact for children in cohort 1 was slightly more negative than the estimate generated for the larger sample (β = 0.115, SE = 0.054; full results for only cohort 1 are shown in Table S7). Thus, although this difference was not statistically significant, it remains possible that the effects of TNVPK could have been slightly more adverse for children in cohort 1 when compared with children in the second cohort.

In Table 8, we present results for cohort 1 on our set of administrative outcomes, including the test-score composite, for the grade 4 wave. Recall that our fourth-grade-wave of data includes children (n = 279) who had been

previously retained and, as a result, did not have test score data at grade 3. For these students, we separately standardized their test scores, and we also tested fourth-grade-wave models that did not include students who had ever been previously retained. Results from these models, shown in supplementary materials Table S8, were nearly identical to those shown in Table 8.

The results for all cohort 1 students with fourth-grade-wave data show patterns of effects similar to those observed for both cohorts through grade 3. We again found positive, though not statistically significant, effects on placement into special education, as well as negative statistically significant effects on placement in gifted and talented programs (ITT estimate: β = -0.022, *SE*= 0.010). The test score composite effect was again negative, but larger; our ITT estimate for TNVPK was -0.165 SD (*SE*= 0.053), and the 2SLS effect ballooned to -0.449 SD (*SE*= 0.168). As with the results for the full sample, we again found little indication that attrition substantially affected treatment impact estimates (see Table 8, Column 4). These results provide further indication that the elementary-school effects of TNVPK may have been adverse.

B. Fadeout

The long-term null to negative effects observed between kindergarten and third grade are consistent with two different explanations as to how TNVPK may have affected students over the long term. First, it may be that children gained less while in the TNVPK program than they would have in other care settings. This would lead to null or negative impacts at the end of preschool that persisted at least through third grade, and would imply that the quality of TNVPK was lower than the quality of other early-care alternatives. On the other hand, these TNVPK sites may have produced positive effects on student cognitive skills initially, but these effects may have faded in the years following the end of preschool.

Fortunately, the TNVPK evaluation study contains some data that can shed light on this question. The initial study design included a collection of student behavioral and test-score measurements from the beginning of preschool through the end of third grade. However, recruitment into this more intensive version of the study, which initially required the active consent of participating families, was plagued by low response rates on data-collection consent forms sent to study families.[16] Worse, the recruitment problem was especially problematic for children assigned to the control condition. For the first cohort, only 20% of the treatment group and 14% of the control group were successfully recruited into the sub-study. In the second cohort, 72% of the treatment group and 52% of the control group agreed to participate. Thus, of the 2,990 students included in the TNVPK study, 1,065 participated in the intensive sub-study rounds of data collection (subsequently called the intensive sub-study sample; ISS). Supplementary file Figure S3 presents a diagram detailing the process of sample inclusion for students in the ISS.

Table 9 shows baseline characteristics from administrative data comparing the students included in the intensive sub-study sample (ISS) with the students who were excluded from participation owing to recruitment failure. Students in the ISS were more likely to be white (55% for ISS vs. 45% for excluded students), but this difference was not statistically significant when r-list fixed

[16] The recruitment problems for both cohorts are described in detail in Lipsey, Farran, and Durkin (2018). For the first cohort, TN DOE staff sent consent forms to all students who appeared on a randomized applicant list. Only 24.4% of the consent forms sent out were returned (this includes students initially on r-lists who were eventually disregarded because of ineligibility or a lack of state administrative data), though nearly all of the returned forms indicated active consent (i.e., few parents refused to participate). For the second cohort, the TN DOE allowed Vanderbilt researchers to approach parents to request consent as a part of the TNVPK application process. As a result, 67.8% of consent forms were returned, the vast majority of them again actively consenting to participate. In addition, Vanderbilt researchers were able to obtain consent for 11 children who were not included in the sample considered here, because no TN DOE administrative records were available for them.

effects were included (p = 0.110). The F-statistic assessing the overall equivalence between the ISS and excluded students on the set of baseline characteristics was not statistically significant ($F(7,110)= 1.82$, p = 0.106), which indicates that the two samples were similar on these few measures.

Supplementary material Table S9 presents information regarding attrition from the treatment and control groups in the ISS, and shows relatively low rates of attrition at each wave (11% of children left the sample by kindergarten, and only 13% had left the sample by third grade). The rate of attrition did not differ between the treatment and control groups at any wave. Although the substantial problems with study recruitment raise concerns about the pre-treatment equivalence of the treatment and control groups included in the ISS, the relative stability of the sample across each follow-up wave suggests that the ISS may provide useful time-series information on treatment/control differences.

In the supplementary materials, we provide extensive information on the high-quality developmental measures collected for the ISS, which included the Woodcock Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew and Mather, 2001), a widely used assessment of cognitive functioning and academic achievement that includes individual tests of math, language, reading, and general cognitive ability. Children's non-cognitive skills were assessed beginning in kindergarten by teachers using the Cooper-Farran Behavioral Ratings Scales (Cooper & Farran, 1988) and the Academic Classroom and Behavior Record (Farran, Bilbrey, & Lipsey, 2003). These two teacher reports measure student learning behaviors, attitudes about school, and social skills. Parents of children in the sub-study were given a short survey on home parenting practices at the beginning of the preschool year, and students completed the WJ-III during the fall of preschool.

Table S9 shows baseline descriptive information for the children included in the treatment and control groups on the extensive set of measures collected for

the sub-study sample. In general, we found evidence of balance across most individual measures. Although point estimates on WJ-III subtest scores often favored the treatment group, these differences were not statistically significant. However, the overall F-test was statistically significant ($F(21,75)= 2.12$, $p = 0.014$), suggesting some degree of difference across the entire set of baseline characteristics between the two groups.

To understand the comparability between the larger sample of children who participated in the lottery study and the ISS, we estimated treatment impacts for the set of state administrative outcomes shown in Table 5 for only children included in the ISS (supplementary Table S11), and we tested for interactions between ISS status and the treatment indicator (Table S12). For special education placement, gifted program placement, and disciplinary offenses, we observed statistically insignificant interactions between treatment status and ISS inclusion. However, for the third grade state test score, we found indications that assignment to TNVPK had a larger negative effect on students in the ISS (ISS-only effect of assignment to TNVPK: $\beta= - 0.210$, $SE= 0.073$, Table 11; interaction term coefficient: $\beta= -0.169$, $SE= 0.096$, Table S12). Further, we found evidence that assignment to TNVPK had a positive effect on absences for students in the ISS (see Table S11) and some evidence that assignment may have lowered retention rates for students in this sample (see Table S12). Despite these differences, observing the longitudinal pattern of effects within the ISS (provided that random assignment produced balanced groups *within* the ISS) still provides a useful picture of how treatment effects might have unfolded over time.

In Table 10, we present estimates of TNVPK treatment impacts on a composite WJ-III measure of cognitive ability, beginning with the end of preschool (see supplementary Table S13 for descriptive information for the outcome measures shown in Table 10). The composite was a standardized (to the control group mean and standard deviation) average of available WJ-III subtests

(impacts on disaggregated measures of math and reading are shown in supplementary Table S14). Because we found some evidence of baseline imbalance between the ISS treatment and control groups, we included two sets of control variables. We first included only the set of administrative controls available for the full sample (i.e., estimates shown in Column 1), then added the entire set of controls available for the ISS (i.e., estimates shown in Column 2; a full list of controls is presented in Table S10). All models shown in Table 10 included r-list fixed effects, and all standard errors were adjusted for site clustering.

Most notably, at the end of preschool we observed a positive effect on WJ-III scores of 0.158 SDs (*SE*= 0.048), and this coefficient was reduced somewhat when the full set of ISS baseline controls was added (β = 0.117, *SE* = 0.029). In kindergarten and first grade, we found non-statistically significant effects close to 0. For second-grade WJ-III scores, our estimated effects were remarkably similar to the composite-score effects estimated on third-grade state-achievement scores for the full sample, as the fully controlled model produced a negative TNVPK coefficient of -0.083 SD (*SE* = 0.039). In third grade, effects were still negative, though slightly smaller and not statistically significant (fully controlled effect: β = -0.061, *SE* = 0.047). All in all, this pattern of declining impact suggests that the TNVPK longitudinal results can be characterized as fadeout; the program positively impacted student cognitive skills at the end of prekindergarten, and this positive effect declined over course of elementary school.

In Table 10, we also present estimates from OLS and 2SLS models using the ISS sample to examine the impact of actual TNVPK attendance on WJ-III scores. The effect of random assignment to TNVPK on attending for at least 20 days was again strong for students in the ISS (β = 0.473, *SE*= 0.037, *t* = 12.64), indicating that assignment to TNVPK was a sound instrument for attendance. Column 4 of Table 10 shows 2SLS impacts on the WJ-III for each of the follow-

up waves, and effects were again scaled up to adjust for non-compliance in the control group. At the end of preschool, attending TNVPK had a positive impact on cognitive skills (0.266 SDs), but this effect was close to 0 by the spring periods of kindergarten and first grade and negative and moderate in magnitude by the end of second grade (β= -0.188, *SE*= 0.089). The third-grade effect was also negative and similar in magnitude, but not statistically significant (β= -0.145, *SE*= 0.110).

Figure 3 provides information on the relative performance of TN students. Specifically, it shows unadjusted (i.e., with no controls for site or baseline measures) differences between treatment and control groups on a standardized measure of the WJ-III at each time point for students in the ISS. Because the WJ-III has been nationally normed, students in both groups can be compared with the national average at each wave. At each measurement point, a student scoring at the national average would produce a score of 100; as Figure 3 shows, students in our study scored well below the national average at the beginning of preschool.[17] At the end of the program, both groups moved closer to the national average, although the treatment group gained at an accelerated pace ($p < 0.001$). By the end of kindergarten, both groups were scoring slightly above the national average, and the difference was no longer statistically significant ($p = 0.124$). By the end of third grade, both groups had again dropped below the national average, and the unadjusted difference was again statistically insignificant ($p = 0.716$).

C. Non-Cognitive Skills and Fadeout

Based on the evidence from Perry Preschool and other early childhood interventions, some have hypothesized that the benefits of early childhood education manifest themselves in such non-cognitive skills as self-regulation, grit,

[17] Figure 3 also shows a slight advantage for students in the TNVPK group over control students on the baseline tests, but this difference was not statistically significant ($p = 0.125$).

and conscientiousness (e.g., Heckman & Kautz, 2014; Raver, 2004). On the other hand, recent discussions around the TNVPK program have raised the possibility that exposing children to rigorous academic instruction during preschool may cause burnout and a loss of interest in school at later stages (e.g., Christakis, 2016). To test whether TNVPK influenced non-cognitive skills, we relied on teacher assessments of the behavior of children in the ISS beginning in kindergarten.

Beginning in kindergarten, teachers rated students on a measure designed to gauge students' interest in school (called the *Feelings About School Scale*) – a potential measure of burnout or loss of interest. Other teacher ratings included assessments of social skills, behavior problems, and learning-related behaviors (e.g., paying attention in class); all were scaled positively and averaged together to create a composite of behavioral adjustment. Table 11 shows generally null treatment impacts for students in the ISS on these two measures of non-cognitive skills in kindergarten, grade 1 and grade 3. In second grade, however, we found similar negative program impacts on both the *Feelings About School Scale* (fully controlled ITT effect: β= -0.172, *SE*= 0.068) and the standardized behavioral composite (fully controlled ITT effect: β= -0.140, *SE*= 0.057). Thus, the evidence suggesting that the TNVPK program caused burnout is ambiguous: we only found a negative effect on the *Feelings About School Scale* at second grade. Further, this effect was not distinguishable from the effect found on composite score that averaged together other measures of behavioral adjustment.

D. Special-Education Placement

Given that TNVPK increased the probability of placement into special education beginning in kindergarten,[18] we explored possible links between such placements and negative impacts on test scores in later elementary school. The vast majority (90%) of referrals to special education in kindergarten were made because of "language or speech impairment," which usually indicates referrals for short doses of supplemental speech therapy outside of class. But if students placed in special education received more remedial instruction over the course of elementary school, then this may have led to poor test scores by second or third grade. To determine whether special-education placement may account for the negative test-score effects, we returned to the full TNVPK sample and tested the third-grade test-score treatment impact model, but also included as control variables indicators for whether the student was placed in special education during elementary school. This "mediational model" (shown in supplementary materials Table S15) reduced the negative treatment impact on the third-grade composite achievement score from -0.082 SDs to -0.061 SDs. That it accounted for about one-quarter of the third-grade test-score effect indicates that special-education placement may have played some but not a major role in producing lower test scores for children in the treatment group.

VI. Conclusion

Our analysis of data from the random-assignment evaluation of the Tennessee Voluntary Prekindergarten program found null effects on measures of behavior, attendance, and retention collected during elementary school. TNVPK

[18] Because children in the treatment group were placed in the TNDOE system a year earlier than control-group children, many of these placements into special education occurred during the prekindergarten year. Specifically, 62% of children from the treatment group who had a special education designation during the kindergarten year were also assigned to special education during prekindergarten. These effects are curious, given that the results shown in Table 11 for the ISS do not indicate that teacher ratings of children's behavior were lower for children in the treatment group during kindergarten, and these teacher behavior ratings included scales measuring children's readiness to learn.

increased enrollment in special education by 3% between kindergarten and grade 3. Although end-of-pre-K impacts on achievement may well have been positive, we found negative but generally insignificant impacts on third-grade state-test scores.

These results largely confirm the findings reported by previous evaluations of TNVPK (see Lipsey et al., 2015; 2018; SRG, 2011). In comparison with the Lipsey et al. (2018) analysis, we used a different definition of the "treatment" group based on a more stringent cutoff between the students likely to have been offered a chance of attendance and students placed on the waitlist. We also employed different modelling techniques, including using r-list and site fixed effects to adjust for the design of the study, and we used alternative approaches to handling missing data, attrition, and non-compliance. However, we found a broadly similar pattern of effects to those reported by Lipsey et al. Interestingly, this pattern of effects has also been reported by non-experimental evaluations of TNVPK, as the earliest state-commissioned work that used only a handful of demographic characteristics to match TNVPK enrollees to non-participants also found small initial benefits that quickly turned to null, or even negative, impacts in later grades (e.g., SRG, 2011).

Indeed, many evaluations of early childhood interventions have found that positive test score impacts disappear over time (Bailey et al., 2017) and there is every indication that TNVPK impacts disappeared as well. Unfortunately, our data were unable to offer satisfying insights into why this fadeout pattern might have occurred. Swain, Springer, and Hofer (2015) found some indication that teacher quality moderated fadeout of test scores after TNVPK, but further analyses have shown that other measures of school quality had no discernible effect (Springer et al., under review).

Even though the program produced null or negative effects on the skills measured by achievement tests, one might still hope that other beneficial impacts,

such as changes in grade repetition and special-education placement, might be generated and persist. We found no evidence that this was the case. Indeed, if anything, a TNVPK offer appeared to increase a child's chances of enrollment in special education in kindergarten and beyond. Of course, it remains possible that higher rates of placement in special education should be counted as a positive outcome for the TNVPK program, if children who needed additional services were provided easier access to them. Empirical studies have struggled to estimate a clear effect of placement into special education services on achievement outcomes (e.g., Hanushek, Kain, & Rivkin, 2002; Morgan et al., 2010), and our data suggested that placement into special education explained little of the negative TNVPK effect on third grade test scores.

All in all, it is disappointing that benefits of enrollment in the TNVPK program did not appear to persist beyond the pre-K year. One possible reason why such benefits disappeared is that although the cognitive and non-cognitive skills fostered in TNVPK classrooms meet many of the quality criteria considered to be important by early education researchers, these skills may not play a central role in longer-run school success. Another factor may be a misalignment of the curricula in pre-K and in the early elementary grades (see Engel, Claessens, Watts, & Farkas, 2016). If kindergarten teachers are focusing on content that TNVPK students have already mastered, control-group children will quickly catch up. Whether and why impacts may have turned negative by third grade is yet another puzzle for the TNVPK research agenda.

References

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP?. *Journal of policy Analysis and Management*, *31*(4), 837-860.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7-39.

Christakis, E. (2016, January/February). The new preschool is crushing kids. *The Atlantic.* Retrieved from: https://www.theatlantic.com/magazine/archive/2016/01/the-new-preschool-is-crushing-kids/419139/

Cooper, D.H., & Farran, D.C. (1988). Behavioral risk factors in kindergarten. Early Childhood Research Quarterly, 3, 1–19.

Elango, S., Garcıa, J. L., Heckman, J. J., & Hojman, A. (2015). Early Childhood Education. *National Bureau of Economic Research Working Paper Series*, No. 21766. doi:10.3386/w21766

Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher*, *45*(5), 293-300.

Farran, D.C., Bilbrey, C. & Lipsey, M.(2003).Academic and Classroom Behavior Record. Unpublished scale available from D.C. Farran, Peabody Research Institute, Vanderbilt University, Nashville, TN.

Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G. G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). *The State of Preschool 2017: State Preschool Yearbook.* New Brunswick, NJ: National Institute for Early Education Research.

Gormley Jr, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental psychology*, *41*(6), 872.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities?. *Review of Economics and Statistics*, *84*(4), 584-599.

Heckman, J. & Kautz, T. (2014). Achievement tests and the role of character in American life. In J. Heckman, J. E. Humphries, & T. Kautz (Eds.), *The myth of achievement tests: The GED and the role of character in American life* (pp. 3-56). Chicago, IL: University of Chicago.

Huang, F. L., Invernizzi, M. A., & Drake, E. A. (2012). The differential effects of preschool: Evidence from Virginia. *Early Childhood Research Quarterly, 27*(1), 33-45.

Lipsey, M. W., Farran, D. C., & Hofer, K. G. (2015). *A randomized control trial of the effects of a statewide voluntary prekindergarten program on children's skills and behaviors through third grade* (Research Report). Nashville, TN: Vanderbilt University, Peabody Research Institute.

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly, 45*(4), 155-176.

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

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*(4), 236-254.

Phillips, D. A., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, P., Duncan, G., Dynarski, M., Magnuson, K. and Weiland, C. (2017). *The Current State of Scientific Knowledge on Pre-Kindergarten Effects* https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

Raver, C. C. (2004). Placing emotional self- regulation in sociocultural and socioeconomic contexts. *Child Development*, *75*(2), 346-353.

Strategic Research Group. (2011). *Assessing the impact of Tennessee's Pre-kindergarten program: Final report*. Columbus Ohio: Strategic Research Group. Retrieved from: http://www.comptroller.tn.gov/Repository/RE/SRG%20PreK%20Final%20Report%202011.pdf

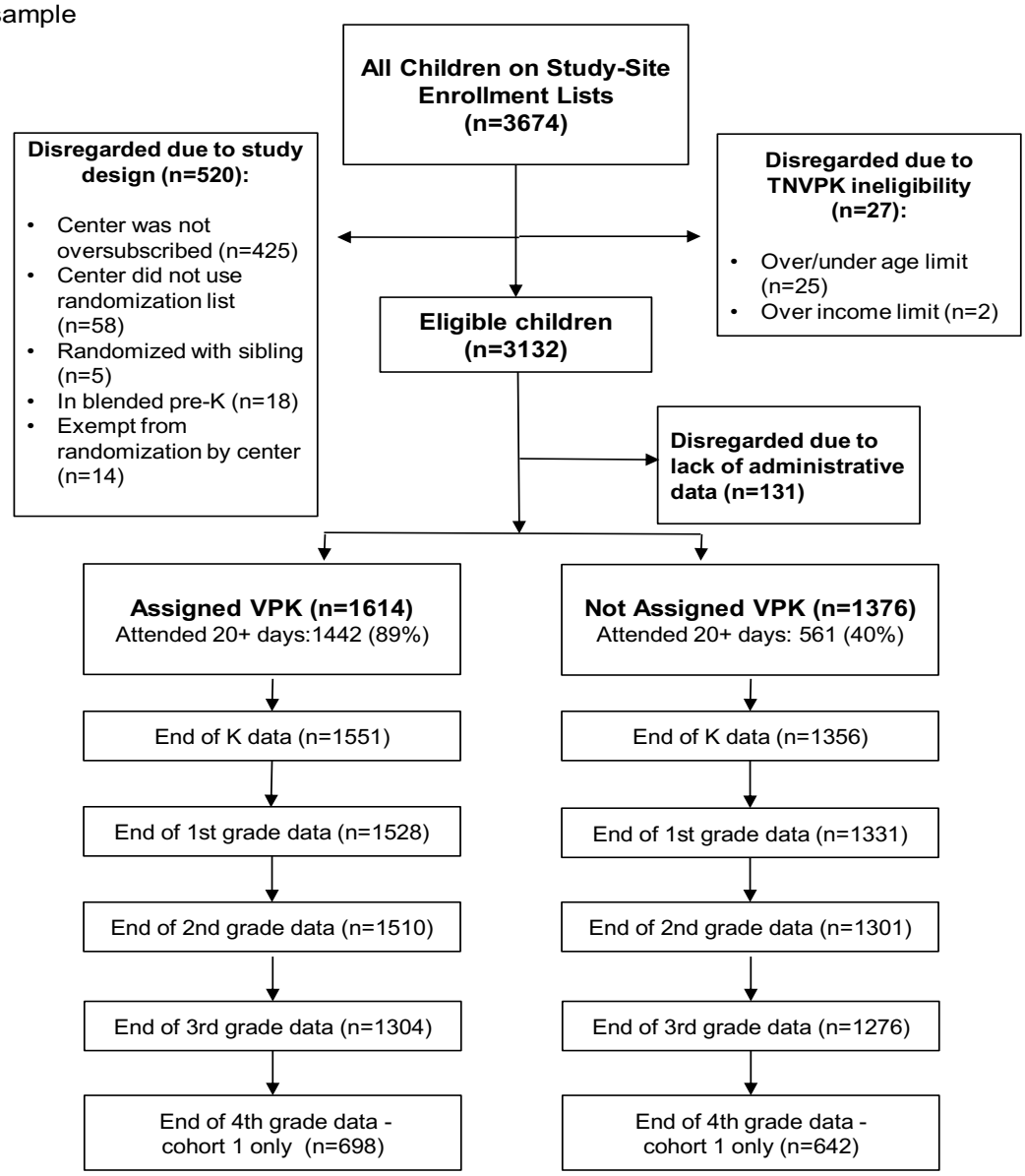
Swain, W. A., Springer, M. G., & Hofer, K. G. (2015). Early grade teacher effectiveness and Pre-K effect persistence: Evidence from Tennessee. *AERA Open*, *1*(4), 1-17. doi: 10.1177/2332858415612751

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., & Vandergrift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Development*, *81*(3), 737-756.

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.

Figure 1

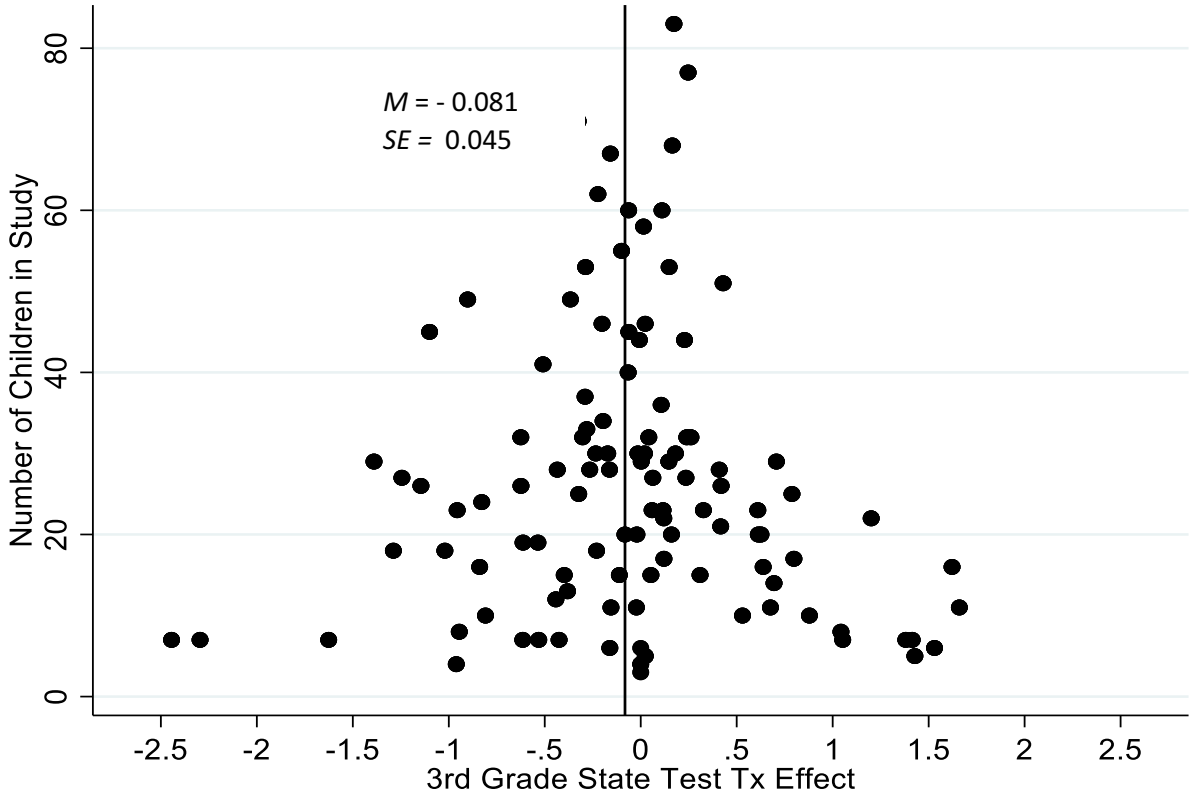
CONSORT diagram describing process that determined inclusion in study sample



Note. At each follow-up wave, our analyses includes students with complete data on all measures for each respective wave. Children not included in a given wave might be included in a later wave if their data for that wave are complete. End of 4th grade data only includes cohort 1.

Figure 2

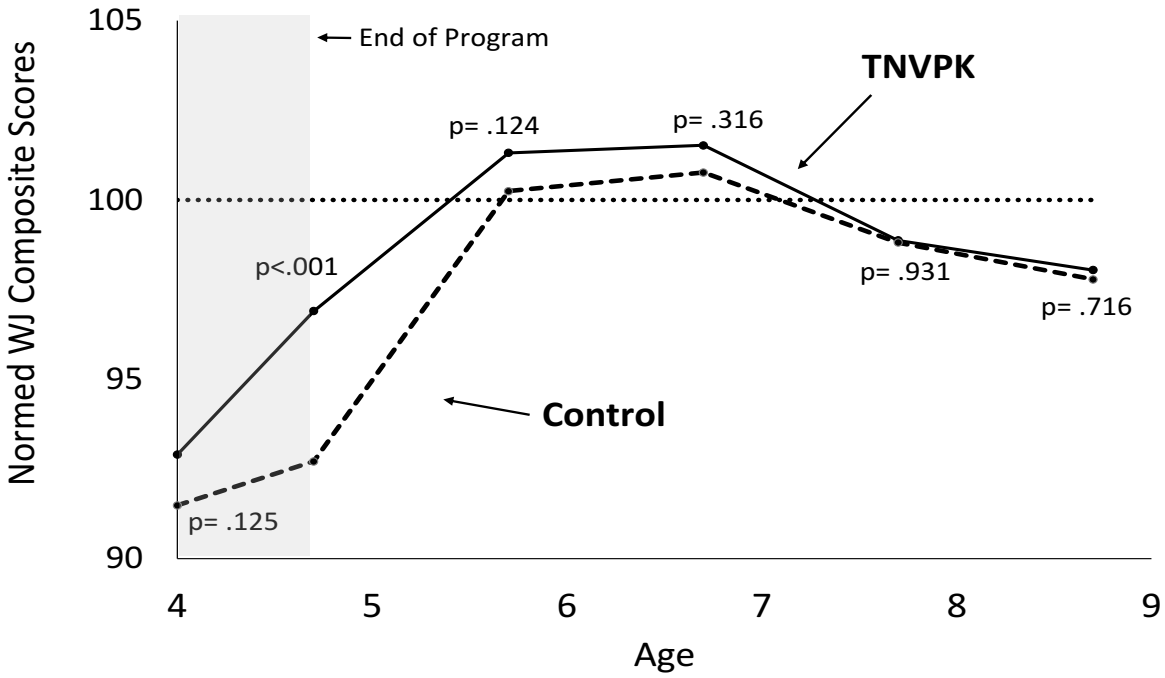
Third Grade Test Score TNVPK Impact by Random Assignment List



Note. Each dot on the graph represents a unique random assignment list included in the study. The vertical line represents the average effect calculated across the r-lists, with each r-list weighted by the number of student observations.

Figure 3

Treatment and Control Group Woodcock Johnson Normed Composite Scores by Age for Children in the ISS



Note. Estimates on the graph were generated from a sample of students in the ISS who had complete data for the entire panel (n = 910). The p-values were generated from models with full control variables, and the shaded area represents the TNVPK preschool period. At each time-point, Woodcock Johnson scores were normed to the national standard (M = 100; SD= 15).

Table 1

Attendance Based on TNVPK Random Assignment

	Full Sample	
	Assigned TNVPK	Not Assigned
Total number of children in sample	1614	1376
Attended any preschool	1438 (89%)	559 (41%)
Average number of days attended	148.18	132.42
Attended 20 days or more	1424 (88%)	550 (40%)
Attended a study school	1366 (84%)	363 (26%)
Attended a non-study school	72 (5%)	196 (14%)

Note: Total number of children in the sample refers to children with at least one wave of state data from waves 2 through 5 for the full sample (see Figure 1). Average number of days attended does not include zeros, and therefore reflects behavior for those who attended. The percentages provided in parentheses represent the share of children in either the treatment or control group who fit into the given category listed.

Table 2

TNVPK Baseline Descriptives

	Assigned TNVPK M (SD)	Not Assigned M (SD)	P-Value of Difference	P-Value (w/ FE)
Age at Preschool Entry (months)	53.22 (3.43)	53.27 (3.50)	0.561	0.753
Female	0.51	0.51	0.890	0.704
White	0.54	0.43	0.001	0.213
Black	0.25	0.29	0.103	0.256
Hispanic	0.19	0.27	0.004	0.584
English Primary Lang	0.81	0.70	0.001	0.979
Study Design Elements				
Cohort 1 (2009-2010)	0.65	0.51	0.001	NA
	F- Test Results			
F (6, 78); no FE's =	2.95		$p = 0.012$	
F (6, 78); w/ FE's =	0.35		$p = 0.907$	
Observations	1614	1376		

Note. P-values were derived from a series of regressions in which each respective baseline characteristic was regressed on the treatment status indicator. The p-values in the first column were generated from bivariate regressions with standard errors adjusted for site-level clustering; the p-values in the second column were generated from regressions with random assignment list (r-list) fixed effects, and standard errors were adjusted for site-level clustering. P-values below 0.001 have been rounded to 0.001. The F-statistic tested whether the entire set of baseline characteristics (excluding cohort) jointly differed between the treatment and control groups. The first F-test p-value was taken from a model without r-list fixed effects, whereas the second p-value included fixed effects.

Table 3

TNVPK Outcome Measures Taken from Administrative Data

	Special Ed Placement		p-val	Gifted Placement		p-val	Disciplinary Offense		p-val	Absences		p-val	Retention		p-val	Test Composite		p-val
	PK	CTL		PK	CTL		PK	CTL		PK	CTL		PK	CTL		PK	CTL	
Kindergarten	0.13	0.07	0.001	0.00	0.00	0.526	0.02	0.01	0.121	9.09	8.34	0.005	0.05	0.05	0.960	-	-	
n = 2,876										(7.48)	(6.77)							
Grade 1	0.15	0.10	0.001	0.00	0.01	0.552	0.03	0.02	0.171	7.76	7.59	0.473	0.05	0.03	0.008	-	-	
n = 2,828										(6.36)	(6.41)							
Grade 2	0.15	0.11	0.004	0.01	0.01	0.284	0.03	0.03	0.699	7.29	6.94	0.106	0.01	0.01	0.447	-	-	
n = 2,783										(5.90)	(5.92)							
Grade 3	0.12	0.10	0.217	0.01	0.01	0.924	0.05	0.03	0.032	6.39	6.30	0.762	0.01	0.01	0.334	752.73	752.64	0.952
n = 2,417										(5.87)	(6.57)					(32.52)	(30.84)	
Outcome Summation	0.21	0.14	0.001	0.01	0.01	0.481	0.08	0.06	0.164	30.15	28.53	0.037	0.12	0.10	0.115			
n = 2,925										(21.05)	(20.73)							

Note. Mean values are presented in each cell, with standard deviations in parentheses. "PK" stands for "assigned to TNVPK" and "CTL" stands for "not assigned to TNVPK" (i.e., "control"). The presented sample sizes represent the number of students who had all administrative measures at each respective wave. The "outcome summation" row is coded "1" if a given outcome occurred within the respective measurement window for each outcome. For example, for special education placement, this variable would be coded to "1" if a student was ever placed in special education between kindergarten and grade 3. For absences, it represents the total number of days recorded absent between kindergarten and grade 3. P-values were derived from a series of bivariate regressions in which each outcome was regressed on treatment status, with no fixed effects for random assignment list included.

Table 4

TNVPK Attrition At Each Wave

	Assigned TNVPK	Not Assigned	P-Value of Difference	P-Value (w/ FE)
Kindergarten	0.039	0.037	0.725	0.950
Grade 1	0.053	0.055	0.788	0.894
Grade 2	0.064	0.075	0.262	0.573
Grade 3	0.192	0.191	0.951	0.129
Missing from "summation" measures	0.022	0.022	0.981	0.969
Observations	1614	1376		

Note. For each outcome wave, we considered a student missing if they did not have all measures for that wave (see Table 3 for list of measures at each wave). The first p-value column was generated without considering school fixed effects, and the second p-value column included school fixed effects. The final row represents the proportion of students who did not have data at all outcome waves (i.e., kindergarten through grade 3).

Table 5

Impact Estimates for the TNVPK Program

	Kindergarten		Grade 1		Grade 2		Grade 3		Outcome Summation
	FE Only	Full Controls	FE Only	Full Controls	FE Only	Full Controls	FE Only	Full Controls	Full Controls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Special Ed	0.030	0.032	0.027	0.030	0.027	0.030	0.019	0.021	0.044
	(0.011)	(0.011)	(0.015)	(0.015)	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)
CTL Mean	0.074		0.095		0.108		0.099		0.135
Gifted	-0.000	-0.000	-0.004	-0.004	-0.008	-0.008	-0.010	-0.009	-0.009
	(0.002)	(0.002)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.006)
CTL Mean	0.003		0.006		0.011		0.013		0.014
Discipline Off.	0.004	0.004	0.002	0.002	-0.010	-0.010	0.016	0.016	0.004
	(0.005)	(0.004)	(0.007)	(0.006)	(0.008)	(0.008)	(0.009)	(0.009)	(0.012)
CTL Mean	0.010		0.021		0.027		0.027		0.065
Absences (std)	0.074+	0.083+	0.022	0.029	0.040	0.049	0.046	0.047	0.056
	(0.044)	(0.044)	(0.043)	(0.043)	(0.039)	(0.039)	(0.042)	(0.042)	(0.041)
CTL Mean	0.002		0.000		0.002		-0.019		0.000
Retention	-0.015	-0.015	0.013	0.014	-0.006	-0.006	0.006	0.006	-0.001
	(0.010)	(0.010)	(0.009)	(0.009)	(0.005)	(0.005)	(0.004)	(0.004)	(0.013)
CTL Mean	0.052		0.029		0.014		0.005		0.099
Test Composite (std)	-	-	-	-	-	-	-0.089	-0.081	-
							(0.047)	(0.045)	
CTL Mean							0.000		
Controls									
R-List F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls		Inc.		Inc.		Inc.		Inc.	Inc.
Observations	2876		2828		2783		2417		2925

Note. Robust standard errors were adjusted for site-level clustering and are presented in parentheses. Continuous variables (i.e., test scores and absences) were standardized using the control group mean and standard deviation for each wave, so coefficients can be likened to effect sizes. All other outcomes were binary, and estimates were generated from linear probability models. All models included fixed effects for the 111 TNVPK random assignment lists (r-list). Each set of estimates was derived from a separate model. For each respective outcome at each respective wave, we present two estimates. The first (i.e., the estimates shown in odd numbered columns) was derived from a model that only included the treatment status indicator and r-list fixed effects. The second (i.e., the estimates shown in even numbered columns) included the treatment status indicator, r-list fixed effects, and the set of baseline controls shown in Table 1. The "outcome summation" column presents results indicating whether the outcome ever occured over the measurement period for special education placements, gifted and talented selction, serious disciplinary offenses, and retention. For absences, the "outcome summation" column models used the total number of absences recorded between kindergarten and grade 3. For each outcome, we also present the mean for the control group.

Table 6

TNVPK: Impacts Adjusted for Non-Compliance

	ITT	OLS	2SLS
	(1)	(2)	(3)
3rd Grade Test Composite	-0.081	0.014	-0.207
	(0.045)	(0.049)	(0.114)
Observations		2417	
Outcome Summations			
Special Education	0.044	0.048	0.114
	(0.018)	(0.017)	(0.046)
Gifted and Talented	-0.009	-0.001	-0.023
	(0.006)	(0.005)	(0.015)
Behavioral Off.	0.004	-0.008	0.011
	(0.012)	(0.011)	(0.029)
Absences	0.056	-0.016	0.144
	(0.041)	(0.047)	(0.106)
Retention	-0.001	-0.026	-0.003
	(0.013)	(0.013)	(0.033)
Observations		2925	

Note. Robust standard errors were adjusted for site-level clustering and are presented in parentheses. For the administrative data outcomes, we used the "outcome summation" variables (see Table 5). All models included random assignment list fixed effects and baseline controls. The ITT estimates are identical to the estimates shown in Table 5. The OLS estimates came from a series of regressions in which each outcome was regressed on a dummy variable for "whether attended 20 days" and covariates. The IV estimates came from a series of 2SLS runs, and variation in "whether attended 20 days" was produced solely by the random assignment indicator (i.e., the insrument).

Table 7

TNVPK Attrition-Adjusted Impacts

	Kindergarten		Grade 1		Grade 2		Grade 3	
	Full Controls	Attrition Adjusted	Full Controls	Attrition Adjusted	Full Controls	Attrition Adjusted	Full Controls	Attrition Adjusted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Special Ed	0.032	0.033	0.030	0.028	0.030	0.024	0.021	0.027
	(0.011)	(0.014)	(0.015)	(0.015)	(0.017)	(0.015)	(0.018)	(0.015)
CTL Mean	0.074		0.095		0.108		0.099	
Gifted	-0.000	-0.000	-0.004	-0.003	-0.008	-0.006	-0.009	-0.009
	(0.002)	(0.004)	(0.007)	(0.005)	(0.007)	(0.006)	(0.008)	(0.006)
CTL Mean	0.003		0.006		0.011		0.013	
Discipline Off.	0.004	0.001	0.002	0.000	-0.010	-0.005	0.016	0.009
	(0.004)	(0.005)	(0.006)	(0.006)	(0.008)	(0.008)	(0.009)	(0.008)
CTL Mean	0.010		0.021		0.027		0.027	
Absences (std)	0.083+	0.078	0.029	0.011	0.049	0.041	0.047	0.034
	(0.044)	(0.051)	(0.043)	(0.047)	(0.039)	(0.044)	(0.042)	(0.039)
CTL Mean	0.002		0.000		0.002		-0.019	
Retention	-0.015	-0.014	0.014	0.012	-0.006	-0.012	0.006	0.007
	(0.010)	(0.011)	(0.009)	(0.009)	(0.005)	(0.005)	(0.004)	(0.004)
CTL Mean	0.052		0.029		0.014		0.005	
Test Composite (std)	-	-	-	-	-	-	-0.081	-0.079
							(0.045)	(0.047)
CTL Mean							0.000	
Controls								
Site F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.

Note. See Table 5 Note. For each outcome measure, estimates shown in the left-hand column are identical to the fully-controlled estimates from Table 5. In the even-numbered columns, we adjusted estimates for attrition by running a series of logistic regression models where students present at each respective wave were coded as "1" and students missing to attrition were coded as "0." We then used the estimated probabilities from these regressions to "weight-up" students who were more likely to be missing due to attrition in the treatment impact models.

Table 8

TNVPK 4th Grade Impacts for Cohort 1

	4th Grade			
	FE Only	Full Controls	2SLS	Attrition Adjusted
	(1)	(2)	(3)	(4)
Special Ed	0.030	0.030	0.082	0.029
	(0.018)	(0.018)	(0.049)	(0.019)
CTL Mean		0.091		
Gifted	-0.021	-0.022	-0.059	-0.019
	(0.009)	(0.010)	(0.027)	(0.009)
CTL Mean		0.024		
Discipline Off.	0.011	0.012	0.032	0.015
	(0.010)	(0.009)	(0.025)	(0.012)
CTL Mean		0.035		
Absences (std)	0.072	0.077	0.210	0.081
	(0.055)	(0.055)	(0.158)	(0.062)
CTL Mean		0.002		
Test Composite (std)	-0.167	-0.165	-0.449	-0.146
	(0.053)	(0.053)	(0.168)	(0.060)
CTL Mean		0.000		
Controls				
Site F.E.	Inc.	Inc.	Inc.	Inc.
State Controls		Inc.	Inc.	Inc.
Observations		1594		

Note. See Table 5 note. Estimates shown here only included children in the first cohort. Estimates in the "Attrition Adjusted" column were weighted by the inverse probability of having 4th grade wave data.

Table 9

Comparisons Between Students Included in the ISS Study and Excluded Students

	ISS	Excluded	P-Value of Difference	P-Value (w/ FE)
	M (SD)	M (SD)		
Age at Preschool Entry (months)	53.19 (3.39)	53.28 (3.51)	0.422	0.516
Female	0.53	0.50	0.177	0.060
White	0.55	0.45	0.022	0.115
Black	0.24	0.29	0.072	0.957
Hispanic	0.19	0.24	0.170	0.077
English Primary Lang	0.80	0.74	0.138	0.089
Study Design Elements				
Cohort 1 (2009-2010)	0.29	0.75	0.001	NA
	F- Test Results			
F (7, 110); no FE's =	1.56		$p = 0.170$	
F (7, 110); w/ FE's =	1.82		$p = 0.106$	
Observations	1065	1925		

Note. See Table 2 note.

Table 10

Impacts for the ISS Sample on Composite Woodcock-Johnson Scores

	ITT (State Controls)	ITT (ISS Controls)	OLS	2SLS
	End of Preschool			
	(1)	(2)	(3)	(4)
n= 1065	0.158	0.117	0.255	0.266
	(0.048)	(0.029)	(0.032)	(0.056)
	Kindergarten			
	(5)	(6)	(7)	(8)
n= 947	-0.001	-0.021	0.010	-0.048
	(0.052)	(0.032)	(0.037)	(0.073)
	First Grade			
	(9)	(10)	(11)	(12)
n= 950	-0.024	-0.046	-0.054	-0.105
	(0.051)	(0.036)	(0.036)	(0.082)
	Second Grade			
	(13)	(14)	(15)	(16)
n= 918	-0.087	-0.083	-0.087	-0.188
	(0.048)	(0.039)	(0.038)	(0.089)
	Third Grade			
	(17)	(18)	(19)	(20)
n= 927	-0.050	-0.061	-0.059	-0.145
	(0.055)	(0.047)	(0.045)	(0.111)

Note. Woodcock-Johnson scores were normed to national averages for each respetive grade-level. See Table 6 note for description of OLS and 2SLS models.

+ p<0.10 * p< 0.05 ** p < 0.01 *** p < 0.001

Table 11

Impacts for the ISS Sample on Behavioral Outcomes

	Feelings About School Scale (std)				Behavioral Composite (std)			
	ITT (State Controls)	**ITT (ISS Controls)**	**OLS**	**2SLS**	**ITT (State Controls)**	**ITT (ISS Controls)**	**OLS**	**2SLS**
				Kindergarten				
	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
n= 947	0.033	0.034	0.056	0.078	0.011	0.034	0.264	0.078
	(0.072)	(0.070)	(0.085)	(0.156)	(0.071)	(0.070)	(0.068)	(0.153)
				First Grade				
	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
n= 950	-0.103	-0.085	-0.064	-0.193	-0.115	-0.098	0.014	-0.223
	(0.080)	(0.076)	(0.095)	(0.172)	(0.085)	(0.073)	(0.095)	(0.169)
				Second Grade				
	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)
n= 918	-0.194	-0.172	0.012	-0.389	-0.194	-0.165	0.056	-0.372
	(0.052)	(0.060)	(0.086)	(0.146)	(0.072)	(0.073)	(0.089)	(0.174)
				Third Grade				
	(29)	(30)	(31)	(32)	(33)	(34)	(35)	(36)
n= 927	-0.090	-0.076	0.085	-0.180	-0.007	0.018	0.079	0.043
	(0.078)	(0.084)	(0.087)	(0.202)	(0.082)	(0.078)	(0.088)	(0.179)

Note. All outcome variables were transformed to z-scores using the control group mean and standard deviation. Both the "Feelings About School Scale" and the "Behavioral Composite" scale were positively rated, so better behavior (or better feelings about school) is indicated by higher scores. See Table 6 for explanation of OLS and 2SLS models.

Supplementary information for

"A Reanalysis of Impacts for the Tennessee Voluntary Prekindergarten Program"

Supplementary Figures

Figure S1 presents the distribution of total days of TNVPK attended for children assigned the treatment and control groups, respectively. These distributions have some degree of overlap because children who attended zero days are not included. As Figure S1 reflects, if students attended any days at all, most attended for the full year.

Figures S2.A through S2.C present distributions of the predicted probability of having non-missing outcome data for each respective wave. These predicted probabilities were generated from a series of logistic regressions in which we regressed an indicator for having non-missing outcome data on treatment status, r-list fixed effects, and baseline covariates.

Finally, Figure S3 presents a CONSORT diagram for students included in the intensive sub-study sample (ISS).

Supplementary Results for the Full Sample

Treatment outcome models for students who were never retained. Recall that our study design includes all students with non-missing data at each respective follow-up wave. Thus, our first grade "wave" includes data for students taken from the 2011-2012 school year for Cohort 1 and the 2012-2013 school year for Cohort 2 (i.e., 2 years following pre-kindergarten). If a student from Cohort 1 repeated kindergarten in 2011-2012, they would also be included in this wave. Further, due to missing data issues in our administrative dataset, our primary third grade treatment impact models shown in Table 5 exclude children who were retained prior to grade 3 (i.e., our dataset did not have test scores and retention data for these children).

To establish the comparability of our results across models, and to examine whether retention patterns might have influenced impact estimates on other outcomes, we present models in Table S1 that only included children who had never been retained. As Table S1 reflects, these estimates were very similar to the estimates shown in Table 5.

Coefficients and standard errors for control variables. Table S2 contains the coefficients and standard errors for all variables included in our primary treatment impact models (i.e., Table 5). In this table, we present results for the "outcome summation" measures and the grade 3 test score composite.

Disaggregated state test outcomes. At grade 3, children took state-administered TCAP tests in mathematics, reading and science. Because scores on these tests were highly correlated (r ranged from 0.71 to 0.75), we averaged the tests together and presented treatment impact results on the average of the three TCAP assessments in the main text. In Table S3, we present unique TNVPK estimates for the mathematics, reading and science score, respectively. As Table S3 reflects, the TNVPK impact was generally similar across the three tests.

Site heterogeneity. In Table S4, we present treatment impact estimates for our "outcome summation" measures and the third grade test score composite estimated at the site level. To generate these estimates, we calculated an r-list-level average for each outcome variable and baseline characteristic for both the control children and the treatment. This process created two observations for each r-list (i.e., a treatment observation and a control observation). We then regressed each respective "r-list averaged" outcome on treatment status and the aggregated baseline characteristics and included a fixed effect adjustment for site. These models treat each r-list as a unique observation, ignoring the number of students present on each r-list.

As Table S4 reflects, we found that these r-list level estimates were largely similar to the primary treatment impact estimates shown in Table 5. However, the TNVPK impact on third grade achievement was smaller when estimated at the site level ($\beta = -0.036$), and the effect on absences was larger ($\beta = 0.119$). This indicates that weighting the treatment effect for each r-list based on the number of student applicants on the list had some influence on the estimates, but estimating the treatment effects at the site level would not have led to substantially different conclusions regarding the long-term effects of TNVPK.

Main treatment impacts with site-level fixed effects. Table S5 presents alternative specifications of our main treatment impact results (i.e., Table 5) that included site-level fixed effects ($g = 79$) instead of r-list fixed effects. As with the primary estimates shown in Table 5, these models also included robust standard errors adjusted for clustering at the site level.

Cohort effects. In Table S6, we present impacts for the "outcome summation" measures and the third grade test score composite, and these models included an interaction between cohort and treatment status. Because models that include r-list fixed effects preclude the inclusion of a cohort control variable, these models include the site-level fixed effect specification shown in Table S5. Across the models, we found statistically significant interactions between cohort and treatment status.

In Table S7, we display our main ITT estimates for children only included in Cohort 1. As these estimates reflect, although we observed no statistically significant cohort interactions, point estimates for Cohort 1 tended to reflect slightly higher placement in special education due to TNVPK assignment and slightly more negative third grade test score impacts when compared with the full sample estimates.

Fourth grade models. Table S8 presents fourth grade results for children in Cohort 1, with children who had ever been retained removed from the sample. As Table S8 reflects, these estimates were similar to those presented in the text, again suggesting that retention rates did not affect our overall impact estimates.

Measurement Information for the Intensive Sub-Sample (ISS)

Outcome measures

For children included in the ISS, measures of cognitive ability were collected via the Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew & Mather, 2001) during the fall of prekindergarten (i.e., study baseline), spring of prekindergarten (i.e., post-treatment), and during the spring of kindergarten, first grade, second grade, and third grade. Teacher ratings of children's behavioral adjustment were collected during the spring semester of each grade point between kindergarten and grade 3. In the following sections, we describe these measures in further detail.

Cognitive skills. For the ISS, children in both conditions were administered the Woodcock Johnson III Tests of Achievement (Woodcock, McGrew & Mather, 2001), a widely-used assessment of cognitive functioning and academic achievement. The WJ-III subtests measure cognitive ability in math as well as reading. For math, the sample was given two subtests at the beginning of preschool, *Applied Problems* and *Quantitative Concepts*. *Applied Problems* measures a child's ability to apply mathematical knowledge, use calculation, and reason quantitatively. The *Quantitative Concepts* subtest measures symbol recognition and manipulation of points on a number line. Beginning during the kindergarten wave of data

collection, the *Calculation* subtest was added, which measures a child's ability to complete visually-presented math problems.

The reading subtests include four tests given at each measurement wave beginning in preschool and another subtest added during the kindergarten year. *Letter-Word Identification* measures ability to read letters and words, which tests recognition of known words as well as phonetic ability to read new words. The *Spelling* subtest measures the child's ability to draw and trace letters, as well as ability to spell orally presented words. The *Oral Comprehension* subtest measures the student's ability to understand short orally presented passages, and requires students to give a missing word at the end of a sentence or sentences. The *Picture Vocabulary* is a test of word knowledge and requires students to name presented pictures. Finally, the *Passage Comprehension* subtest was during the kindergarten measurement wave and tests a child's ability to understand written text. It requires students to match short phrases to pictures, and more complex items require students to fill in missing words from sentences of increasingly complex paragraphs.

For the present analysis, we generated a composite measure of achievement across the WJ-III subtests by taking the average of each child's non-missing WJ-III subtests. For each subtest, we used the WJ-III standard score. Thus, each test was nationally normed to a mean of 100 and standard deviation of 15, allowing us to compare performance on the measure between our sample and a national population at each respective wave.

Behavioral measures. Children in the ISS received behavioral ratings from teachers during the spring of kindergarten, grade 1, grade 2 and grade 3. These behavioral ratings included two measures, the Cooper-Farran Behavioral Ratings Scales (Cooper & Farran, 1991) and the Academic Classroom and Behavior Record (Farran, Bilbrey, & Lipsey, 2003). The Cooper-Farran Behavioral Ratings Scales included two subscales. The *Work-Related Skills* subscale is a report of the child's ability to work independently, listen to and comply with instructions, complete tasks, and generally behave appropriately in the classroom. The *Social Behavior* subscale is a report of the child's interactions with peers including appropriate group and play behavior, expression of feelings and response to others. The Academic Classroom and Behavior Record, also a teacher rating scale, includes three scales. *Readiness for Grade Level Work* measures how well the child is generally prepared for grade level work in math, reading, and social behavior. The *Feelings About School* subscale is a teacher report of how much the child appears to like school. The *Behavior Problems* subscale (the only one that is negatively scaled) measures whether the child exhibits behavioral problems, including physical or relational aggression, social withdrawal or anxiety, and motor difficulties. Lastly, the *Peer Relations* subscale measures the degree to which the child appears to be liked by other students in the classroom.

For the current analysis, the *Behavior Problems* subscale was rescaled so that positive scored indicated less behavior problems. To test the specific hypothesis that the TNVPK program might have led to negative adjustment in school by causing burnout, we tested impacts on the *Feelings About School* scale individually, and averaged together the remaining behavioral measures to create a positively-scaled behavioral composite.

Supplementary Results for the ISS Sample

ISS Attrition. Table S9 contains attrition information for the ISS sample. Overall, we did not see high rates of attrition for the treatment or control groups, and attrition rates were similar for treatment and control groups at all waves.

ISS Baseline Equivalence. Table S10 presents tests of baseline comparability between treatment and control children in the ISS sample. Compared with the few administrative measures available for students in the full sample, the ISS sample participated in a much more extensive data collection effort at study baseline. Consequently, Table S10 shows baseline information for the state administrative controls, as well as information regarding parenting behaviors and parental background characteristics taken from a parenting survey, as well as cognitive skills taken from the WJ-III test. Study administrators encountered some difficulties collected timely baseline measures of student achievement via the WJ-III, which meant that students were administered the WJ-III test an average of 43 days into the school year. This differed between the treatment and control group by 10 days ($p < 0.001$), with the control group being tested later, on average.

As with the tests for baseline balance shown in Table 2 of the main text, we again regressed each characteristic on the dummy variable for treatment and adjusted standard errors for site-level clustering. The p-values shown in the first column came from bivariate regressions that did not include r-list fixed effects, and the p-values shown in the second column came from regressions included the r-list fixed effect adjustment. Similar to the pattern of results observed in Table 2, we again saw that once r-list fixed effects were taken into account, most differences between the treatment and control group were not close to statistically significant. However, even with the fixed effect adjustment included in the model, the F-test assessing whether the entire set of baseline characteristics jointly differed produced a statistically significant result ($F(20,57)= 3.85$, $p = 0.014$), indicating some degree of difference between the treatment and control groups at baseline.

TNVPK state-data impacts for ISS. To help establish the comparability between the ISS and the full sample of children who participated in the TNVPK lottery, we present the same TNVPK impact estimates shown in Table 5 for only children included in the ISS. Across the estimates shown in Table S11, assignment to TNVPK tended to produce slightly worse outcomes for children in the ISS and slightly higher placement in special education. In Table S12, we present interactions between ISS status and our TNVPK assignment variable. Reflecting the point estimates shown in Table S11, we found marginally statistically significant interactions for our measure of absences and the 3rd grade test score- again indicting worse TNVPK impacts for the ISS. However, TNVPK lowered the overall retention rate for children in the ISS when compared to children not included in the ISS.

ISS Outcome Descriptives. Table S13 presents descriptive information for the set of unique outcome measures collected for the ISS sample. For the WJ-III measures of cognitive skills, we created composite measures that were generated by taking the average of all non-missing WJ-III subtests at a given time point. We also created subject-specific composite scores for mathematics and reading. All WJ-III subtests were scaled to have a mean of 100 and a standard deviation of 15 at a given time-point.

For the behavioral outcome measures, we transformed each respective subscale into a z-score, using the mean and standard deviation of the control group. We then considered the

Feelings About School score individually and averaged the other behavioral measures together to create a positively-scaled composite measure of teacher-rated behavioral adjustment.

ISS Disaggregated Treatment Impacts. In Table 10 of the main text, we present TNVPK impacts on the overall composite measure of WJ-III cognitive skills (i.e., the measure that included both mathematics- and reading-focused subtests). In Table S14, we present analogous findings for WJ-III disaggregated measures of mathematics and reading, respectively. As with the TCAP state achievement tests, we again found similar impacts on measures of both mathematics and reading achievement.

Mediation Model with Special Education. In Table S15, we tested whether placement in special education accounted for the negative impact observed in grade 3 on the state achievement test. Column 1 displays the same fully-controlled treatment impact model shown in Table 5 of the main text. In Column 2, we included an indicator of whether the student was ever placed in special education between kindergarten and grade 3, and Column 3, we included an indicator of special education placement for each individual wave. In both models the negative treatment impact on the grade 3 composite test score was slightly reduced, but only to 0.06 SD's, indicating that special education placement alone does not account for the negative TNVPK impact on third grade test scores.

References

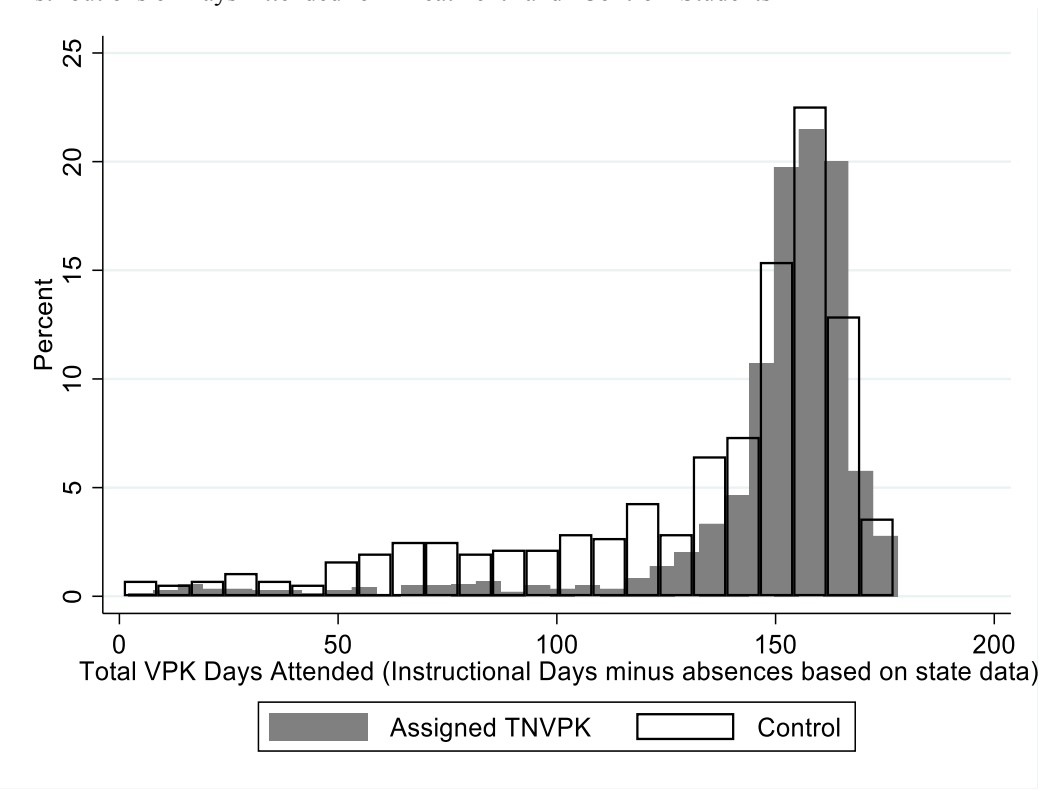
Cooper, D. & Farran, D.C. (1991). Cooper-Farran Behavioral Rating Scale. Clinical Psychology Publishing Company, Inc.

Farran, D.C., Bilbrey, C. & Lipsey, M. (2003). Academic and Classroom Behavior Record. Unpublished scale available from D.C. Farran, Peabody Research Institute, Vanderbilt University, Nashville, TN.

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

Figure S1

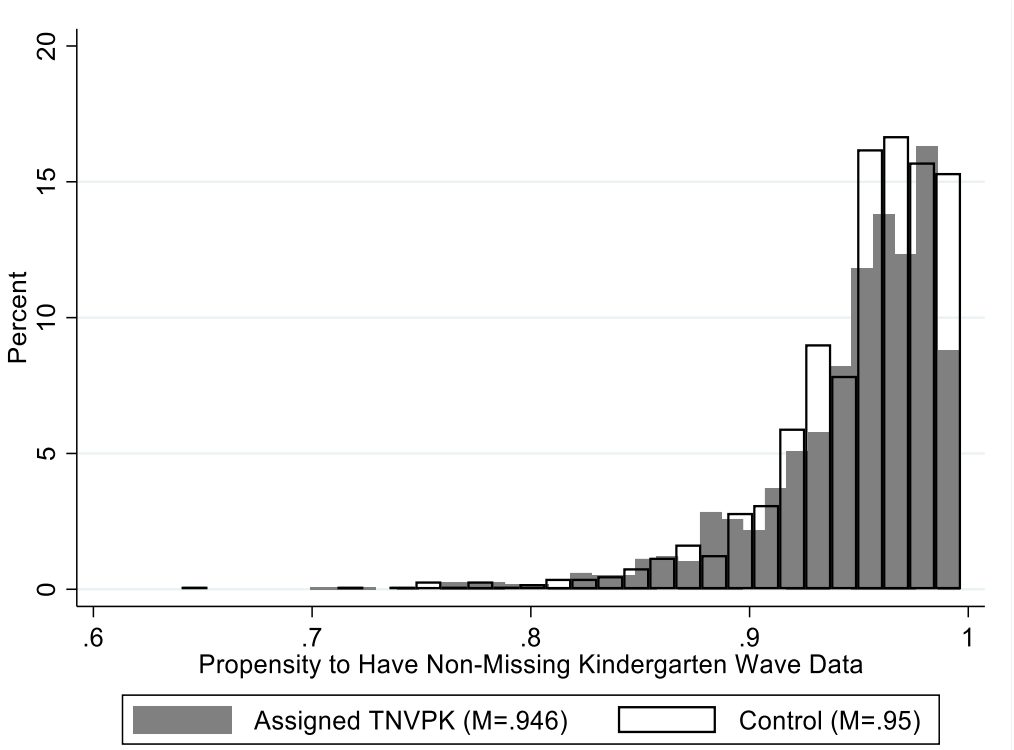
Distributions of Days Attended for "Treatment" and "Control" Students

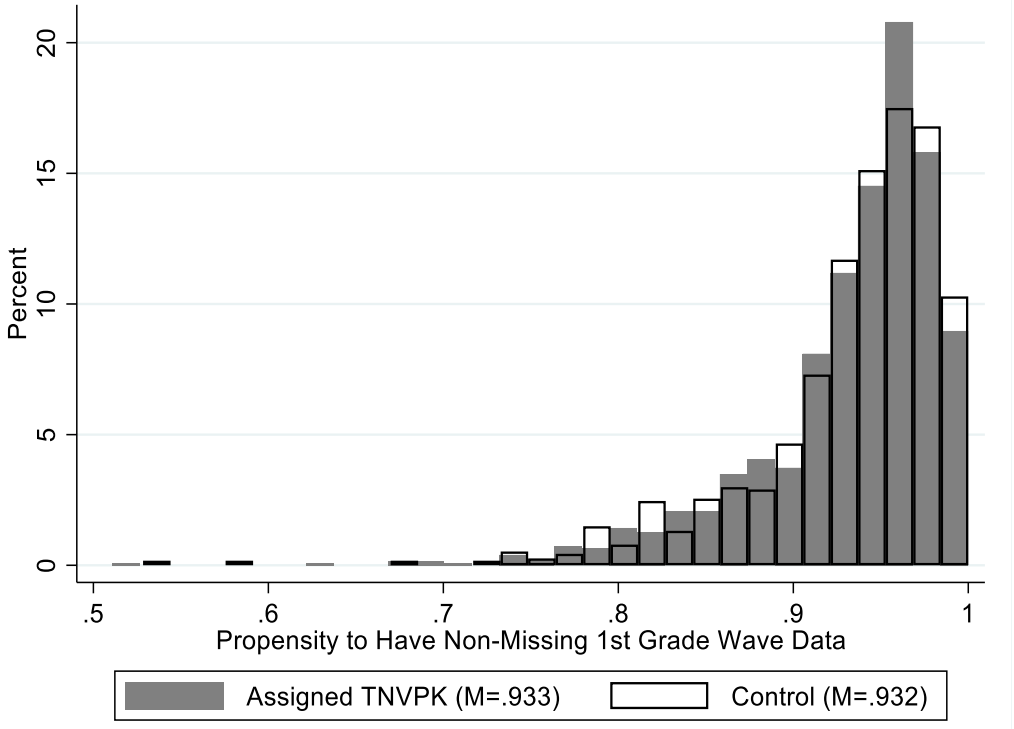


Note. These distributions do not include children who attended 0 days.

Figure S2.A

Distributions of the Propensity to Have Non-Missing Data at Each Follow-Up Wave

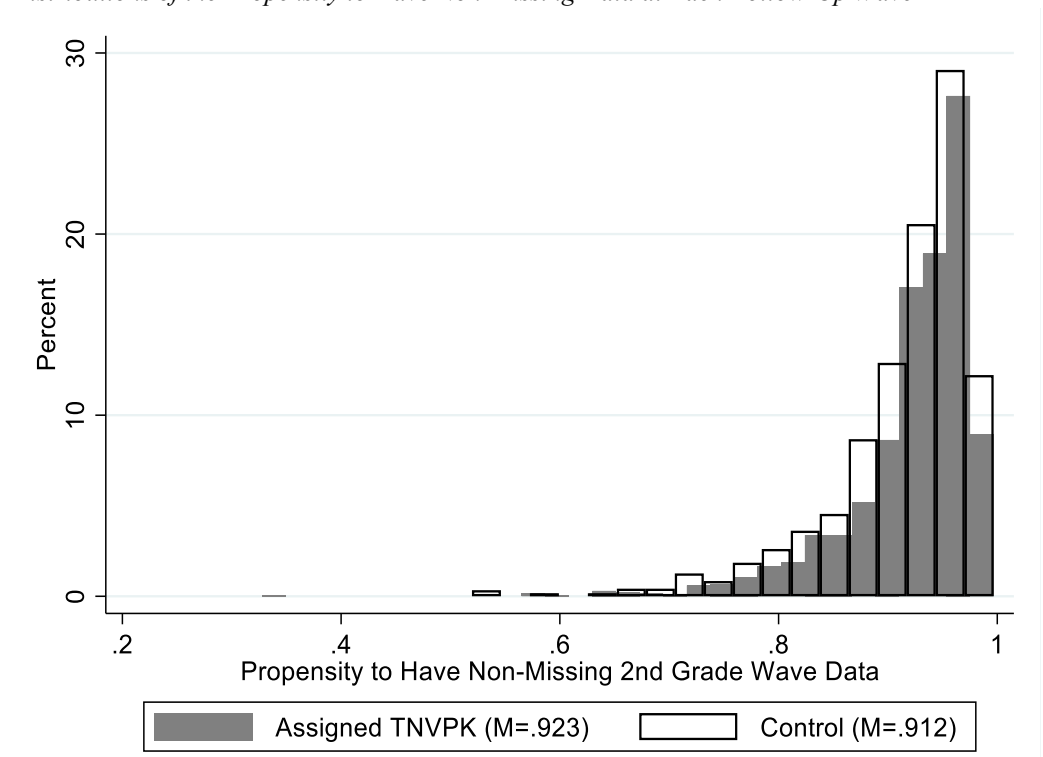


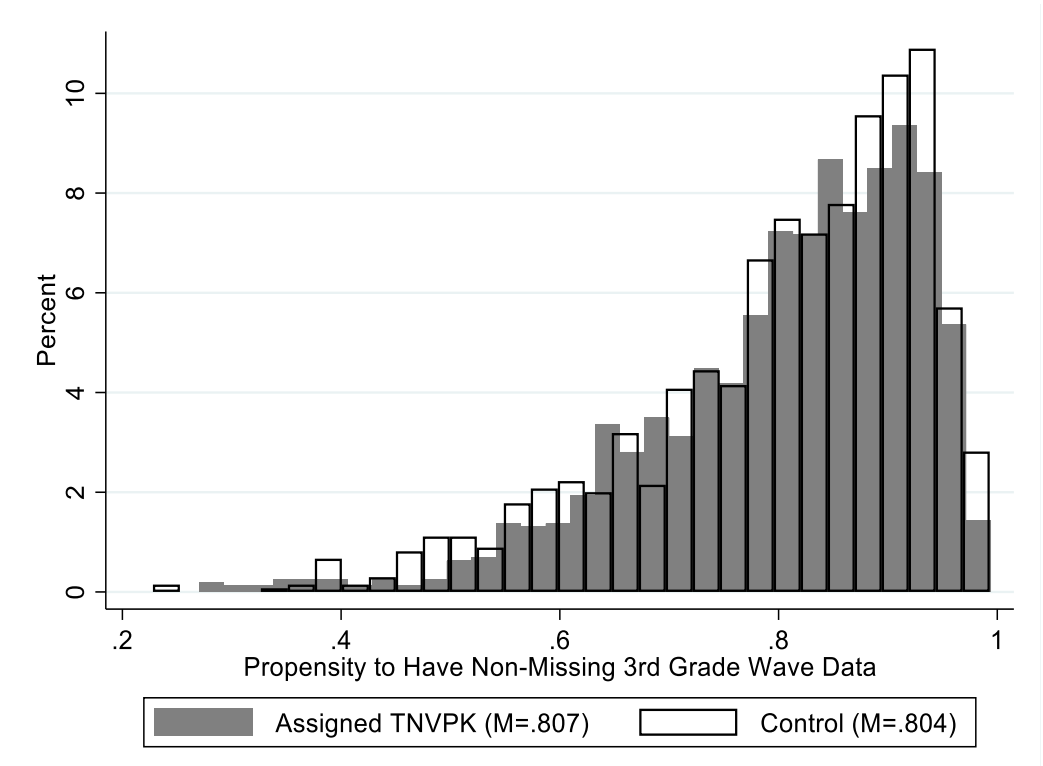


Note. See Figure S2.C for full figure note.

Figure S2.B

Distributions of the Propensity to Have Non-Missing Data at Each Follow-Up Wave

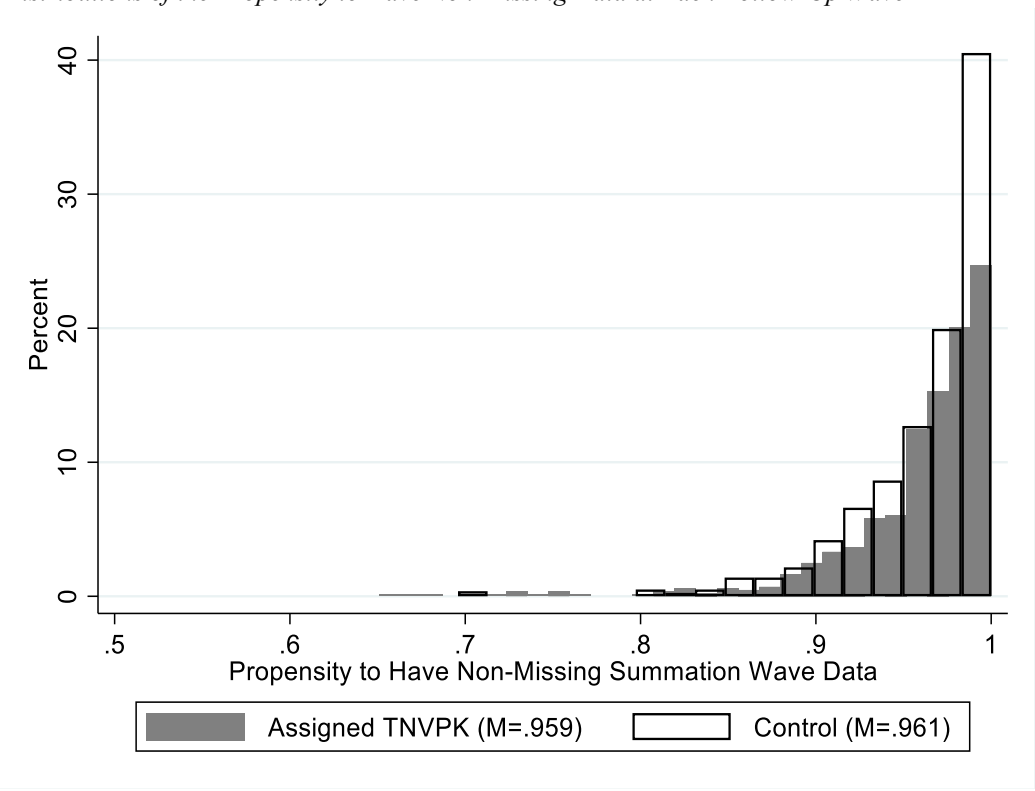




Note. See Figure S2.C for full figure note.

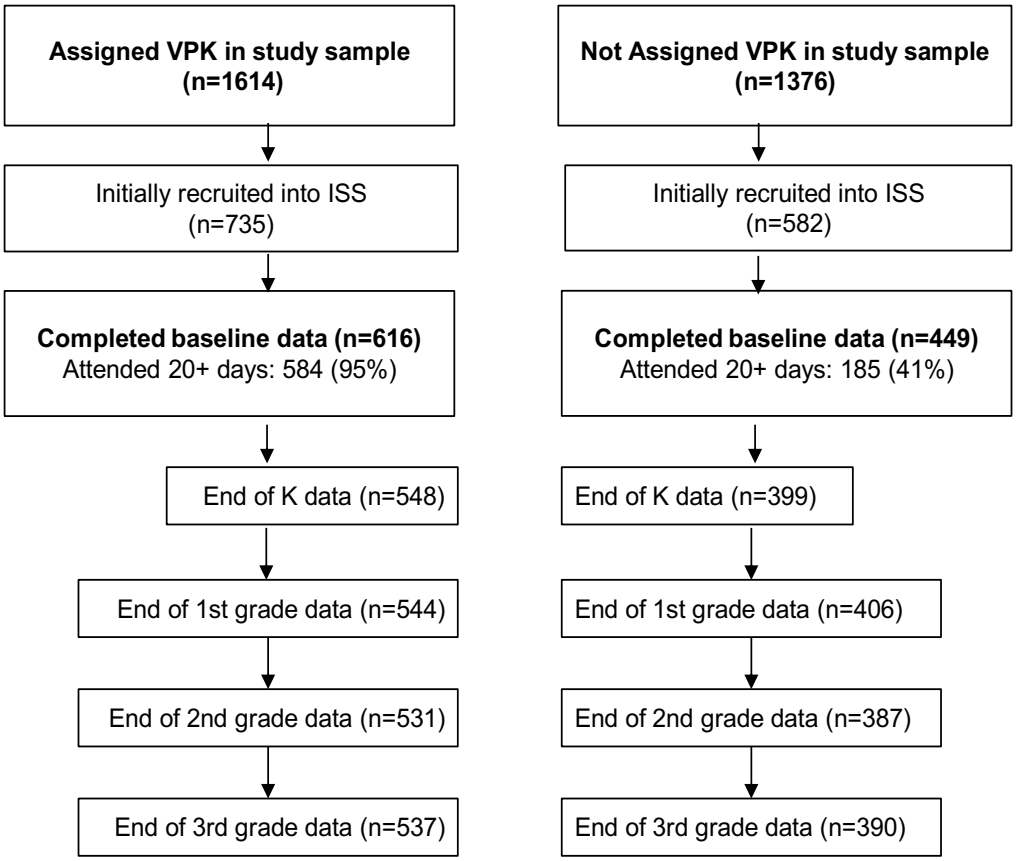
Figure S2.C

Distributions of the Propensity to Have Non-Missing Data at Each Follow-Up Wave



Note. Each figure presents the predicted probability of having full outcome data at each respective wave. These probabilities were generated by running a series of logistic regressions that modeled a binary indicator of remaining in the sample as a function of treatmetn status, r-list fixed effects, and baseline covariates.

Figure S3
CONSORT diagram describing process that determined inclusion in intensive sub-sample (ISS).



Note. Sub-sample of students were recruited from the broader study to complete more intensive testing. For our analyses, we include students with complete data on all measures for each respective wave. Thus, children can leave the sample at one wave and re-enter at a later wave if data for the later wave are non-missing.

Table S1

Impact Estimates for the TNVPK Program- Excluding Children Ever Retained Between K and Grade 3

	Kindergarten		Grade 1		Grade 2		Grade 3		Outcome Summation	
	Full Sample	None Retained	Full Sample	None Retained	Full Sample	None Retained	Full Sample	None Retained	Full Sample	None Retained
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Special Ed	0.032	0.023	0.030	0.024	0.030	0.021	0.021	0.019	0.044	0.037
	(0.011)	(0.011)	(0.015)	(0.014)	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)
Gifted	-0.000	-0.000	-0.004	-0.005	-0.008	-0.009	-0.009	-0.009	-0.009	-0.009
	(0.002)	(0.002)	(0.007)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.006)	(0.007)
Discipline Off.	0.004	0.006	0.002	0.000	-0.010	-0.008	0.016	0.018	0.004	0.013
	(0.004)	(0.005)	(0.006)	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)	(0.012)	(0.013)
Absences (std)	0.083	0.073	0.029	0.037	0.049	0.047	0.047	0.049	0.056	0.062
	(0.044)	(0.041)	(0.043)	(0.041)	(0.039)	(0.040)	(0.042)	(0.042)	(0.041)	(0.040)
Test Composite (std)	-	-	-	-	-	-	-0.081	-0.074	-	-
							(0.045)	(0.045)		
Controls										
R-List F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Observations	2876	2561	2828	2509	2783	2472	2417	2399	2925	2601

Note. See Table 5 note. All models included baseline controls. Models shown in the odd columns are identical to the specifications shown in Table 5. In the even columns, only students who were never retained between kindergarten and grade 3 were included.

Table S2

Coefficients and Standard Errors for Baseline Characteristics

	Special Ed	Gifted	Discipline Off	Absences	Retention	3rd Gr. Test Composite
	(1)	(2)	(3)	(4)	(5)	(6)
Assigned TNVPK	0.044	-0.009	0.004	0.056	-0.001	-0.081
	(0.018)	(0.006)	(0.012)	(0.041)	(0.013)	(0.045)
Controls						
Age at PreK Entry	-0.005	0.002	0.001	-0.009	-0.061	0.083
	(0.008)	(0.002)	(0.005)	(0.019)	(0.008)	(0.025)
Female	-0.097	0.001	-0.077	0.014	-0.054	0.111
	(0.014)	(0.005)	(0.011)	(0.038)	(0.013)	(0.037)
White	0.016	-0.051	0.029	0.427	0.022	-0.471
	(0.045)	(0.040)	(0.031)	(0.100)	(0.036)	(0.154)
Black	-0.033	-0.047	0.088	0.077	-0.015	-0.773
	(0.039)	(0.039)	(0.036)	(0.106)	(0.037)	(0.158)
Hispanic	0.020	-0.046	0.020	0.199	-0.018	-0.433
	(0.037)	(0.033)	(0.034)	(0.090)	(0.038)	(0.168)
English Primary Lang.	0.086	0.013	0.050	0.333	0.045	0.231
	(0.027)	(0.008)	(0.015)	(0.051)	(0.020)	(0.099)
R-List Fixed Effects	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Observations	2925	2925	2925	2925	2925	2417

Note. See Table 5 note. For models 1 through 5, we used the "outcome summation" measures shown in Column 9 of Table 5. The control measure of baseline age was standardized across the sample. All other control variable were binary (i.e., dummy coded) indicators, with males serving as the reference group for gender and "ethnicity- other" serving as the reference group for race and ethnicity.

Table S3

Treatment Impacts on Disaggregated 3rd Grade Test Scores

	Math	Reading	Science
Treatment	-0.084	-0.066	-0.070
	(0.044)	(0.044)	(0.045)
Controls			
Age at PreK Entry	0.052	0.084	0.090
	(0.027)	(0.023)	(0.023)
Female	0.052	0.258	-0.009
	(0.039)	(0.037)	(0.038)
White	-0.503	-0.423	-0.342
	(0.137)	(0.160)	(0.162)
Black	-0.759	-0.595	-0.733
	(0.151)	(0.156)	(0.167)
Hispanic	-0.413	-0.384	-0.368
	(0.143)	(0.169)	(0.179)
English Primary Lang	0.102	0.249	0.270
	(0.117)	(0.081)	(0.094)
R-List F.E.	Inc.	Inc.	Inc.
Observations	2416	2414	2415

Note. See Table 5 note. All continuous variables were standardized, so coefficients can be interpreted as effect sizes.

Table S4

Effects of TNVPK Estimated at the R-List Level

	Special Ed	Gifted	Discipline Off.	Absences	Retention	3rd Grade Test Composite
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.039	-0.006	0.018	0.119	0.016	-0.036
	(0.024)	(0.006)	(0.017)	(0.070)	(0.018)	(0.067)
Controls						
Age at PreK Entry	0.081	0.007	0.067	0.230	-0.049	0.118
	(0.049)	(0.012)	(0.033)	(0.139)	(0.036)	(0.120)
Female	0.052	0.025	0.039	0.664	-0.053	-0.187
	(0.084)	(0.020)	(0.057)	(0.240)	(0.063)	(0.203)
White	0.315	-0.206	0.211	0.826	0.145	-1.552
	(0.340)	(0.083)	(0.232)	(0.975)	(0.254)	(0.885)
Black	0.390	-0.207	0.220	0.103	0.167	-2.481
	(0.331)	(0.080)	(0.226)	(0.950)	(0.247)	(0.891)
Hispanic	0.420	-0.190	0.156	-0.201	0.175	-2.358
	(0.322)	(0.078)	(0.220)	(0.923)	(0.240)	(0.847)
English Primary Lang.	0.228	0.017	-0.033	0.052	-0.014	-0.775
	(0.196)	(0.048)	(0.134)	(0.561)	(0.146)	(0.512)
Number of R-Lists	111	111	111	111	111	111

Note. Outcome variables included the "outcome summation" measures for Columns 1 through 5 and the 3rd grade test score composite for Column 6. Models were estimated by caluclating a treatment and control mean for each baseline control variable and outcome variable at the r-list level. This aggregation essentially created two observations for each r-list (i.e., average outcomes and baseline characteristics for treatment students and average outcomes and baseline characteristics for control children). We then regressed the respective outcome variable on treatment status and the baseline variables while including r-list fixed effects. These estimates produce r-list level treatment impacts that do not take into account the number of students present at each site, essentially treating

Table S5

Impact Estimates for the TNVPK Program with Site Fixed Effects

	Kindergarten		Grade 1		Grade 2		Grade 3		Outcome Summation
	FE Only	Full Controls	FE Only	Full Controls	FE Only	Full Controls	FE Only	Full Controls	Full Controls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Special Ed	0.029	0.03	0.027	0.029	0.027	0.029	0.017	0.020	0.043
	(0.011)	(0.011)	(0.014)	(0.015)	(0.016)	(0.016)	(0.018)	(0.018)	(0.017)
CTL Mean	0.07		0.10		0.11		0.10		0.12
Gifted	-0.000	-0.000	-0.004	-0.004	-0.008	-0.008	-0.010	-0.010	-0.009
	(0.002)	(0.002)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)	(0.008)	(0.006)
CTL Mean	0.00		0.01		0.01		0.01		0.01
Discipline Off.	0.004	0.004	0.001	0.001	-0.008	-0.008	0.015	0.016	0.005
	(0.004)	(0.004)	(0.006)	(0.005)	(0.008)	(0.008)	(0.009)	(0.008)	(0.011)
CTL Mean	0.01		0.02		0.03		0.03		0.02
Absences (std)	0.081	0.087	0.040	0.046	0.041	0.048	0.042	0.043	0.062
	(0.043)	(0.044)	(0.040)	(0.040)	(0.038)	(0.038)	(0.045)	(0.045)	(0.040)
CTL Mean	0.00		0.00		0.00		0.00		0.00
Retention	-0.016	-0.016	0.015	0.016	-0.008	-0.007	0.007	0.007	-0.001
	(0.010)	(0.009)	(0.008)	(0.009)	(0.006)	(0.006)	(0.004)	(0.004)	(0.013)
CTL Mean	0.05		0.03		0.01		0.01		0.01
Test Composite	-	-	-	-	-	-	-0.09	-0.082	-
							(0.045)	(0.043)	
CTL Mean							0.00		
Controls									
Site F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls		Inc.		Inc.		Inc.		Inc.	Inc.
Observations	2876		2828		2783		2417		2925

Note. See Table 5 note. Estimates came from models identical to the models shown in Table 5, except that we included site-level fixed effects ($g = 79$) instead of r-list fixed effects ($g = 111$). Robust standard errors were again adjusted for site-level clustering. Models include a control for cohort.

Table S6

ISS: TNVPK Impacts by Cohort Interactions

Special Ed	0.031
	(0.024)
x Cohort interaction	0.018
	(0.025)
Gifted	-0.012
	(0.010)
x Cohort interaction	0.006
	(0.008)
Disciplinary Offense	0.014
	(0.019)
x Cohort interaction	-0.013
	(0.025)
Retention	-0.008
	(0.016)
x Cohort interaction	0.012
	(0.020)
Absences	0.054
	(0.060)
x Cohort interaction	0.014
	(0.062)
3rd Grade TCAP Composite	-0.011
	(0.070)
x Cohort interaction	-0.120
	(0.077)

Note. See Table 5 note. This presents results indicating whether the outcome ever occured over the measurement period. The 3rd Grade TCAP is a composite of all state tests. The cohort interaction coefficient indicates whether results differred by cohort for each respective outcome. Models included site level fixed effects and standard errors clustered at the site level.

Table S7

Impact Estimates for the TNVPK Program- Cohort 1 Only

	Kindergarten		Grade 1		Grade 2		Grade 3		Outcome Summation	
	Both Cohorts	Cohort 1	Both Cohorts	Cohort 1	Both Cohorts	Cohort 1	Both Cohorts	Cohort 1	Both Cohorts	Cohort 1
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Special Ed	0.032	0.035	0.030	0.028	0.030	0.032	0.021	0.025	0.044	0.051
	(0.011)	(0.014)	(0.015)	(0.016)	(0.017)	(0.018)	(0.018)	(0.020)	(0.018)	(0.021)
Gifted	-0.000	-0.003	-0.004	-0.008	-0.008	-0.010	-0.009	-0.013	-0.009	-0.012
	(0.002)	(0.003)	(0.007)	(0.006)	(0.007)	(0.006)	(0.008)	(0.007)	(0.006)	(0.006)
Discipline Off.	0.004	0.007	0.002	0.008	-0.010	-0.015	0.016	0.006	0.004	-0.014
	(0.004)	(0.005)	(0.006)	(0.007)	(0.008)	(0.012)	(0.009)	(0.011)	(0.012)	(0.012)
Absences (std)	0.083	0.050	0.029	0.024	0.049	0.047	0.047	0.035	0.056	0.049
	(0.044)	(0.060)	(0.043)	(0.055)	(0.039)	(0.049)	(0.042)	(0.053)	(0.041)	(0.050)
Retention	-0.015	-0.010	0.014	0.017	-0.006	-0.011	0.006	0.009	-0.001	0.006
	(0.010)	(0.012)	(0.009)	(0.009)	(0.005)	(0.008)	(0.004)	(0.005)	(0.013)	(0.018)
Test Composite (std)	-	-	-	-	-	-	-0.081	-0.115	-	-
							(0.045)	(0.054)		
Controls										
R-List F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Observations	2876	1681	2828	1653	2783	1637	2417	1421	2925	1709

Note. See Table 5 note. All models included baseline controls. Models shown in the odd columns are identical to the specifications shown in Table 5. In the even columns, only students in "Cohort 1" were included.

Table S8

TNVPK 4th Grade Impacts for Cohort 1- No Previously Retained Students

	4th Grade			
	FE Only	Full Controls	2SLS	Attrition Adjusted
	(1)	(2)	(3)	(4)
Special Ed	0.028	0.027	0.073	0.025
	(0.019)	(0.019)	(0.049)	(0.020)
CTL Mean		0.071		
Gifted	-0.021	-0.021	-0.056	-0.021
	(0.011)	(0.011)	(0.029)	(0.011)
CTL Mean		0.025		
Discipline Off.	0.012	0.013	0.035	0.015
	(0.009)	(0.009)	(0.024)	(0.012)
CTL Mean		0.034		
Absences (std)	0.101	0.100	0.267	0.113
	(0.059)	(0.058)	(0.165)	(0.068)
CTL Mean		-0.026		
Test Composite (std)	-0.161	-0.160	-0.427	-0.132
	(0.060)	(0.061)	(0.175)	(0.066)
CTL Mean		0.000		
Controls				
Site F.E.	Inc.	Inc.	Inc.	Inc.
State Controls		Inc.	Inc.	Inc.
Observations		1594		

Note. See Table 5 note. Only students in "Cohort 1" who had no record of having been retained in a previous grade were included. Estimates in the "Attrition Adjusted" column were weighted by the inverse probability of having 4th grade wave data.

Table S9

ISS: Attrition at Each Wave

	TX	Control	P-Value of Difference	P-Value of Difference (w/FE)
Kindergarten	0.110	0.111	0.968	0.750
Grade 1	0.117	0.096	0.392	0.149
Grade 2	0.138	0.138	0.996	0.844
Grade 3	0.128	0.131	0.878	0.588
Observations	616	449		

Note. For each outcome wave, the proportion of students missing all tests scores and behavioral ratings is presented. The first p-value column was generated without considering school fixed effects, and the second p-value column included school fixed effects. Unfortunately, we could not assess attrition on end-of-preschool measures, because study developers only provided us with data on the ISS children who had non-missing test score data in preschool.

Table S10

ISS: Baseline Characteristics Drawn from State Data and ISS Measures

	TX	Control	P-Value of Difference	P-Value (w/ FE)
Cohort 1 (2009-2010)	0.33	0.22	0.044	NA
State Data Controls				
Age at PreK Entry	53.17	53.27	0.626	0.456
	(3.43)	(3.43)		
Female	0.53	0.52	0.590	0.289
White	0.61	0.49	0.015	0.499
Black	0.21	0.26	0.120	0.530
Hispanic	0.15	0.23	0.082	0.936
English Primary Lang	0.86	0.73	0.005	0.504
Additional ISS Controls				
Library Card Use (0-2)	0.93	0.89	0.457	0.032
Newspaper Subscriptions (0-3)	0.37	0.36	0.839	0.739
Magazines Subscriptions (0-2)	0.29	0.26	0.405	0.848
Mother's Education				
Did not graduate H.S.	0.15	0.17	0.481	0.766
Graduated H.S.	0.68	0.65	0.467	0.381
Some college	0.11	0.11	0.913	0.647
BA or higher	0.07	0.07	0.832	0.349
One Parent Works	0.52	0.55	0.364	0.495
Two Parents Work	0.36	0.35	0.635	0.417
Time Between R.A. and Baseline Test (Days)	39.47	49.63	0.002	0.001
	(27.71)	(32.65)		
Woodcock Johnson				
Letter Word Identification	93.64	92.47	0.320	0.205
	(12.75)	(14.51)		
Spelling	86.20	87.77	0.190	0.592
	(14.56)	(14.95)		
Oral Comprehension	93.99	91.64	0.044	0.917
	(12.64)	(14.10)		
Picture Vocabulary	96.65	91.94	0.006	0.380
	(16.69)	(21.11)		
Applied Problems	97.70	95.69	0.054	0.867
	(13.24)	(14.26)		
Quantitative Concepts	89.21	89.42	0.815	0.323
	(11.87)	(12.52)		
Composite	92.90	91.48	0.139	0.918
	(10.60)	(11.99)		
F- Test Results				
F (20, 57); no FE's =		2.55		*p=0.003*
F (20, 57); w/ FE's =		2.12		*p=0.014*
Observations	616	449		

Note. See Table 2 note. The F-statistic tested whether the entire set of baseline characteristics (excluding cohort and the "Time between R.A. and Baseline Test" measure) jointly differed between groups. The first F-test p-value was taken from a model without r-list fixed effects, whereas the second p-value included fixed effects.

Table S11

Impact Estimates for the TNVPK Program- ISS Sample Only

	Kindergarten		Grade 1		Grade 2		Grade 3		Outcome Summation	
	Both Cohorts	ISS Only	Both Cohorts	ISS Only	Both Cohorts	ISS Only	Both Cohorts	ISS Only	Both Cohorts	ISS Only
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Special Ed	0.032	0.044	0.030	0.041	0.030	0.035	0.021	0.056	0.044	0.062
	(0.011)	(0.023)	(0.015)	(0.033)	(0.017)	(0.032)	(0.018)	(0.028)	(0.018)	(0.034)
Gifted	-0.000	0.000	-0.004	-0.007	-0.008	-0.011	-0.009	-0.010	-0.009	-0.011
	(0.002)	(0.000)	(0.007)	(0.010)	(0.007)	(0.012)	(0.008)	(0.014)	(0.006)	(0.011)
Discipline Off.	0.004	0.005	0.002	-0.003	-0.010	0.005	0.016	0.009	0.004	0.019
	(0.004)	(0.006)	(0.006)	(0.010)	(0.008)	(0.011)	(0.009)	(0.012)	(0.012)	(0.019)
Absences (std)	0.083	0.205	0.029	0.125	0.049	0.147	0.047	0.154	0.056	0.168
	(0.044)	(0.088)	(0.043)	(0.086)	(0.039)	(0.068)	(0.042)	(0.057)	(0.041)	(0.078)
Retention	-0.015	-0.029	0.014	0.006	-0.006	0.000	0.006	-0.005	-0.001	-0.029
	(0.010)	(0.021)	(0.009)	(0.019)	(0.005)	(0.007)	(0.004)	(0.005)	(0.013)	(0.021)
Test Composite (std)	-	-	-	-	-	-	-0.081	-0.210	-	-
							(0.045)	(0.073)		
Controls										
R-List F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
State Controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Observations	2876	1032	2828	1014	2783	992	2417	856	2925	1039

Note. See Table 5 note. All models included baseline controls. Models shown in the odd columns are identical to the specifications shown in Table 5. In the even columns, only students in the ISS were included

Table S12

ISS: TNVPK Impacts by ISS Status Interactions

Special Ed	0.032
	(0.022)
x ISS interaction	0.032
	(0.035)
Gifted	-0.008
	(0.005)
x ISS interaction	-0.002
	(0.011)
Disciplinary Offense	-0.001
	(0.013)
x ISS interaction	0.010
	(0.026)
Retention	0.021
	(0.018)
x ISS interaction	-0.057
	(0.027)
Absences	0.006
	(0.052)
x ISS interaction	0.154
	(0.089)
3rd Grade TCAP Composite	-0.038
	(0.060)
x ISS interaction	-0.169
	(0.096)

Note. See Table 5 note. For administrative outcomes, we used the "outcome summation" measures. The ISS interaction coefficient indicates whether results differed for students included in the ISS sample.

Table S13

ISS: Descriptive Information for Unique ISS Outcome Variables

	Spring of Pre-K	Kindergarten	Grade 1	Grade 2	Grade 3
Woodcock-Johnson Composite					
Tx	96.91	101.31	101.52	98.87	98.04
	(9.80)	(9.34)	(10.08)	(9.97)	(10.27)
Control	92.71	100.25	100.77	98.81	97.79
	(12.63)	(9.95)	(9.85)	(9.49)	(10.02)
Math Composite					
Tx	97.59	101.08	101.32	98.73	97.81
	(11.52)	(10.67)	(11.74)	(11.78)	(12.33)
Control	94.06	100.36	101.16	99.20	98.36
	(13.49)	(11.33)	(11.38)	(11.39)	(12.40)
Reading Composite					
Tx	96.56	101.43	101.65	98.96	98.19
	(9.72)	(9.55)	(9.98)	(9.97)	(10.19)
Control	92.04	100.19	100.53	98.58	97.43
	(12.94)	(10.19)	(9.95)	(9.36)	(9.79)
Behavior Composite					
Tx	-	0.07	-0.08	-0.10	-0.07
	-	(0.76)	(0.83)	(0.83)	(0.81)
Control	-	-0.00	0.00	-0.00	-0.00
	-	(0.80)	(0.79)	(0.81)	(0.81)
Feelings About School Scale					
Tx	-	0.06	-0.05	-0.16	-0.10
	-	(1.01)	(1.00)	(1.03)	(1.04)
Control	-	0.00	0.00	0.00	0.00
	-	(1.00)	(1.00)	(1.00)	(1.00)

Note. Means are presented, with standard deviations in parentheses. The *Woodcock Johnson Composite* is an average of all WJ-III subtests available at a given timepoint. The *Math Composite* is the average of the math-specific WJ-III tests at a given timepoint (i.e., *Applied Problems, Calculation,* and *Quantitative Concepts)*. The *Reading Composite* is the average of all reading-specific WJ-III subtests at each timepoint (i.e., *Letter Word Identification, Spelling, Picture Vocabulary,* and *Passage Comprehension*). Finally, the *Behavior Composite* variable is the average of a set of standardized (i.e., $M = 0$; $SD = 1$) teacher ratings conducted in each grade between kindergarten and grade 3 that asked teachers to assess the child on their classroom behavior (positive scores indicate better-behaved children). The Feelings About School scale is considered separately.

Table S14

Impacts for the ISS Sample on Woodcock-Johnson Math and Reading Scores

	ITT (State Controls)	ITT (ISS Controls)	OLS	2SLS	ITT (State Controls)	ITT (ISS Controls)	OLS	2SLS
	Math End of Preschool				**Reading End of Preschool**			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
n= 1065	0.103	0.076	0.242	0.172	0.185	0.138	0.261	0.313
	(0.056)	(0.037)	(0.043)	(0.076)	(0.047)	(0.031)	(0.033)	(0.060)
	Math Kindergarten				**Reading Kindergarten**			
	(5)	(6)	(7)	(8)	(5)	(6)	(7)	(8)
n= 947	-0.020	-0.015	0.004	-0.035	0.008	-0.024	0.013	-0.054
	(0.064)	(0.047)	(0.040)	(0.105)	(0.051)	(0.032)	(0.040)	(0.072)
	Math First Grade				**Reading First Grade**			
	(9)	(10)	(11)	(12)	(9)	(10)	(11)	(12)
n= 950	-0.051	-0.069	-0.044	-0.157	-0.008	-0.032	-0.060	-0.074
	(0.060)	(0.047)	(0.048)	(0.107)	(0.051)	(0.036)	(0.038)	(0.079)
	Math Second Grade				**Reading Second Grade**			
	(13)	(14)	(15)	(16)	(13)	(14)	(15)	(16)
n= 918	-0.132	-0.123	-0.106	-0.279	-0.060	-0.059	-0.076	-0.133
	(0.063)	(0.057)	(0.047)	(0.132)	(0.047)	(0.037)	(0.039)	(0.081)
	Math Third Grade				**Reading Third Grade**			
	(17)	(18)	(19)	(20)	(17)	(18)	(19)	(20)
n= 927	-0.132	-0.123	-0.100	-0.248	-0.016	-0.033	-0.032	-0.079
	(0.063)	(0.057)	(0.057)	(0.149)	(0.054)	(0.044)	(0.044)	(0.102)

Note. Woodcock-Johnson scores were normed to national averages for each respetive grade-level. See Table 6 note for description of OLS and 2SLS models.

Table S15

Special Education Placement as a Mediator of the TNVPK Test Score Impact

	3rd Grade Test Composite		
	(1)	(2)	(3)
TNVPK	-0.082	-0.061	-0.067
	(0.045)	(0.045)	(0.044)
Ever Placed in Special Ed		-0.574	
		(0.081)	
Single Year Special Ed Placement			
Kindergarten			0.003
			(0.132)
Grade 1			-0.153
			(0.155)
Grade 2			0.225
			(0.138)
Grade 3			-0.867
			(0.116)
Controls			
R-List F.E.	Inc.	Inc.	Inc.
State Controls	Inc.	Inc.	Inc.
Observations	2417	2417	2376

Note. The treatment impact estimate shown in Column 1 is identical to Column 3 of Table 5.