Effects of a Statewide Prekindergarten Program on Children's Achievement and Behavior through Sixth Grade

Kelley Durkin[1], Mark W. Lipsey[2], Dale C. Farran[1], and Sarah E. Wiesen[3]

[1] Department of Teaching and Learning, Vanderbilt University

[2] Department of Human and Organizational Development, Vanderbilt University

[3] Peabody Research Office, Vanderbilt University

Published: March 2022

Citation

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology*, *58*(3), 470-484. http://dx.doi.org/10.1037/dev0001301

Author Note

Kelley Durkin https://orcid.org/0000-0002-9439-5386

Study materials are available at https://my.vanderbilt.edu/tnprekevaluation/ (Evaluating the Effectiveness of Tennessee's Voluntary Pre-K Program, n.d.) and deidentified datasets are available by request from Kelley Durkin. This study was not preregistered. This research was supported by the U.S. Department of Education Institute of Education Sciences Grants R305E090009 and R305A210130 and the U.S. Department of Health and Human Services National Institute of Child Health and Human Development Grant R01HD079461-01 to Vanderbilt University. The opinions expressed are those of the authors and do not represent the views of the Institute of Education Sciences, the U.S. Department of Education, or the National Institute of Child Health and Human Development. This work would not have been possible without the assistance of the Tennessee Department of Education and Tennessee Education Research Alliance. Special thanks to Stone Dawson, Jane Hughart, and Ilknur Sekmen, for their hard work collecting, cleaning, and analyzing data for this project. We are also grateful for the support of multiple school districts and school administrators throughout Tennessee.

Correspondence concerning this article should be addressed to Kelley Durkin, Department of Teaching and Learning, Vanderbilt University, 230 Appleton Place, PMB 230, Nashville, TN 37203, United States. Email: kelley.durkin@vanderbilt.edu

Abstract

As state-funded prekindergarten programs expand, it is critical to investigate their short- and long-term effects. This paper presents the results through 6th grade of a longitudinal randomized control study of the effects of a scaled-up, state-supported pre-k program. The analytic sample includes 2990 children from low-income families who applied to oversubscribed pre-k program sites across the state and were randomly assigned to offers of admission or a wait list control. Data through 6th grade from state education records showed that the children randomly assigned to attend pre-k had lower state achievement test scores in third through 6th grades than control children, with the strongest negative effects in 6th grade. A negative effect was also found for disciplinary infractions, attendance, and receipt of special education services, with null effects on retention. The implications of these findings for pre-k policies and practices are discussed.

Keywords: public pre-k, randomized control trial, longitudinal, early childhood education, achievement, policy

Effects of a Statewide Prekindergarten Program on Children's Achievement and Behavior at Sixth Grade

Prior to 1980, only two states funded pre-k programs; currently 43 states fund such programs. There has been a corresponding increase in the enrollment of 4-year-old children with about 1.5 million currently enrolled in state-funded pre-k programs—nearly one-third of all 4-year-old children in the U.S. (Friedman-Krauss et al., 2018) and a larger enrollment than the federal Head Start program (Administration for Children and Families, 2019). The objectives of the advocacy groups and state legislatures promoting this expansion vary, but one common theme is enhancing the school readiness of children prior to kindergarten entry, especially children from economically disadvantaged families. Alongside that short-term goal, however, is an expectation that pre-k will have longer-term effects on academic and behavioral outcomes and help close the achievement gap (Phillips et al., 2017).

The expectations for long-term pre-k effects are largely based on the results of the now classic early longitudinal studies – the Perry Preschool and Abecedarian demonstration projects (Bailey et al., 2017). Both projects not only showed positive effects on later academic performance but, after high school, on such life outcomes as employment and income, further education, criminal behavior, and health. However, these were small demonstration projects that involved intensive programs for highly disadvantaged African American children in contexts that provided few intellectually stimulating alternatives. Whether similar results can be produced by less intensive scaled-up contemporary statewide programs for more diverse populations is a critical question for both policy and practice.

Unfortunately, questions about both short-term and, especially, long-term state pre-k effects are difficult to address in a methodologically rigorous way. A randomized study akin to

those used in the Perry Preschool and Abecedarian projects would require assembling a sample of pre-k eligible children prior to the beginning of the pre-k school year whose parents consented to having them randomly assigned to attend or not attend the state program. Few parents can be expected to allow the equivalent of a coin flip to make such an important decision for their child. Absent random assignment, the inherent parental discretion involved in choosing between options creates considerable potential for selection bias in any comparison of outcomes for state pre-k vs. any or all of the available alternatives. Similar concerns apply to evaluations of other publicly funded pre-k programs such as place-based locally funded programs (e.g., Tulsa and Boston) and the federally funded Head Start program.

One credible approach to evaluating the immediate effects of state pre-k programs under these circumstances is the age-cutoff regression-discontinuity (RD) design developed by Gormley and colleagues in their study of the Tulsa pre-k program (Gormley et al., 2005). This design capitalizes on a strict age cutoff for pre-k eligibility and compares age-adjusted outcomes at the beginning of kindergarten for children who attended pre-k the prior year with those for younger children just beginning the pre-k program and thus had no pre-k exposure that prior year. These evaluations have focused on early literacy, language, and math and, almost without exception, have found positive effects at kindergarten entry (e.g., Pion & Lipsey, *in press*; Hustedt et.al. 2021; Montrosse-Moorhead et al., 2019; Wong et al., 2008). A limitation of this design is that it cannot be used for longer term follow-up—by the end of kindergarten both the treatment and control groups have experienced pre-k.

Because of the practical difficulties of implementing random assignment, studies of the sustained effects of state pre-k past the beginning of kindergarten almost universally use

nonrandomized designs. One approach to avoiding selection bias associated with parental decisions about attending pre-k has been to use aggregate data that includes all the children in different cohorts whether they attended state pre-k or not. Longer-term outcomes are then compared for cohorts that vary in their exposure to state pre-k. For example, Ladd, Muschkin, and Dodge (2014) examined 3rd grade test scores as a function of differences across counties in the onset and magnitude of state financial investments in the North Carolina pre-k program. They found positive effects in reading and math with especially large estimates if the aggregate effects were assumed to come entirely from the children in each county who actually attended pre-k. Similarly, other studies (e.g., Bartik & Hershbein, 2018; Cascio & Schanzenbach, 2013; Rosinsky, 2014) have analyzed 4th grade NAEP test scores in relation to variation in pre-k enrollment across time and, in some instances, across states. The results of these studies have been mixed and generally show sensitivity to differences in the statistical models used.

Attributing such aggregate effects to the influence of pre-k, however, assumes that the aggregate units most successful in increasing exposure to pre-k were not also more likely to implement or experience any other developments that affected educational performance co-occurring with the pre-k expansion or in any year between then and the time at which the longer-term outcomes were measured. This assumption that pre-k variation is independent of other educational influences is difficult to confirm and not entirely plausible in the context of interconnected educational innovations and policy.

The most common approach to investigating longer-term effects of locally funded and statewide pre-k programs is a post hoc comparison of later outcomes for children who did vs. did not attend the pre-k program years earlier (e.g., Hill et al., 2015; Preskitt et al., 2020). These

studies generally attempt to address the inherent potential for selection bias by matching the children in each group on a set of variables available at the time of outcome measurement. A recent example is an investigation of the long-term effects of Georgia's universal pre-k program (Early et al., 2019). Using demographic data from the state database, kindergarten eligibility for free or reduced-price lunch, and kindergarten school attended, Early et al. created propensity score matches for pre-k attenders and non-attenders. The results showed positive pre-k effects on 3rd grade reading and math scores for children from low-income families, but negative effects for pre-k participants from higher income families. Notably, however, no pretest measures taken before the beginning of the pre-k year for early reading and math skills or for family circumstances that might influence pre-k enrollment were available to ensure initial equivalence between those groups. Indeed, the authors caution that variables such as these might have accounted for the differences found. The child and family characteristics associated with enrollment in pre-k generally favor enrollees (Ansari & Crosnoe, 2015; Coley et al., 2016) in ways that undermine the validity of such post hoc comparisons.

Despite the challenges that have motivated the large body of nonrandomized studies of the longer-term effects of state and locally funded pre-k, there are circumstances amenable to randomized designs. For example, when a sizeable number of pre-k sites receive more applicants than they have capacity to serve, some applicants must of necessity be turned away. Randomization provides an equitable and transparent way to allocate the available seats at the respective sites. One of the two studies of state or locally funded pre-k that has randomized in such a situation capitalized on the lottery procedure used to select applicants for popular pre-k program sites in Boston (Weiland et al., 2020). This naturally occurring lottery allowed Weiland

and colleagues to compare children randomly selected to attend their first choice program with a control group not selected for that first choice. This contrast was diminished, however, by the ultimate attendance of 72% of the control group in another Boston pre-k program, 24% in a center-based preschool other than a Boston pre-k, and only about 3% who did not attend any preschool. The direct intent-to-treat comparison between the overall treatment and control groups found no differences on 3rd grade English language arts and math tests nor on retention in grade and special education placements through 3rd grade. An instrumental variable analysis that compared children attending any Boston pre-k program with those attending other preschools or none similarly found no difference on any of these outcomes. It is notable, however, that the effect estimates for the 3rd grade math scores and kindergarten special education placements were negative, though short of statistical significance.

While a state pre-k program is the focus of this paper, a large body of research has focused on the Head Start program. However, there is only one randomized study of longer-term Head Start effects (Puma et al., 2012), one that also randomized applicants to oversubscribed programs. Head Start children had larger gains than controls on literacy and language measures (but not math) prior to kindergarten entry, but these effects disappeared by the end of kindergarten. Focusing on earlier Head Start programs, Deming (2009) conducted a study comparing siblings within the same family born between 1976 and 1986 who did or did not attend Head Start, and found long-term positive Head Start effects on adult outcomes even though test score differences faded. In a similar analysis, Pages et al. (2020) found that using the Deming sample but extending the measurement period decreased the adult effects, and data for children attending more recent Head Start programs showed mostly negative effects.

Siblings who attended Head Start were *less* likely to be employed or enrolled in school compared to their siblings who mostly received home care. These later Head Start programs occurred within the same time window as the implementation of the Tennessee Voluntary Pre-K (TN-VPK) program that is the topic of the current paper.

This paper reports results from a randomized longitudinal study of TN-VPK that began with the 2009 and 2010 pre-k cohorts. As in the Head Start Impact Study, the Tennessee research team implemented randomization at oversubscribed program sites and followed the resulting sample afterwards to investigate how well the pre-k effects were sustained. The results through 6th grade are reported here and those for earlier periods are summarized.

TN-VPK and Effects through Third Grade

While state pre-k programs vary, the Tennessee program is relatively typical. Pilot programs began in 1996 with full statewide implementation in 2005. TN-VPK is organized and overseen by the state department of education and serves more than 18,000 4-year-old children from low-income families statewide with local program sites in all but a few of the school districts in the state. The state requires a minimum instructional time of 5.5 hours per day, five days a week during the school year, classes of no more than 20 students staffed by a state-licensed teacher endorsed for early childhood education and paid at public school teacher rates, an educational assistant in each room, and a curriculum selected from a state-approved list (Tennessee Department of Education, 2019).

When the TN-VPK program began, it met 9 of the 10 standards advocated until recently revised by the National Institute of Early Education Research (NIEER: Barnett et al., 2009). The current study began with the 2009 and 2010 pre-k cohorts. In 2015 (Farran & Lipsey) we

reported a separate related study of a representative sample of TN-VPK programs in the state that found that quality as measured by the Early Childhood Environment Rating Scale (Harms & Clifford, 1980) matched or exceeded that reported in evaluations of other state pre-k programs. More recently, Pion & Lipsey (in press) used a regression-discontinuity design with that statewide sample to investigate end-of-pre-k effects on a battery of commonly used cognitive measures. The TN-VPK results compared favorably with those found in similar designs for more than a dozen other statewide pre-k programs.

The findings of the TN-VPK study through 3rd grade were described in prior research reports (Lipsey et al., 2013a and 2013b) and a published summary article (Lipsey et al., 2018). The design involves over-subscribed program sites across Tennessee in which applicants were randomized to offers of admission or waitlist status (*N* = 2,990 children; referred to as the *RCT analytic sample*). In addition, parental consent was obtained for 1,076 children to allow the research team to collect additional data from the children and their teachers at the beginning and end of the pre-k year and annually through 3rd grade (referred to as the *intensive substudy (ISS) sample*). We attempted to obtain parental consent for all children in the RCT full sample, but different procedures for obtaining consent were used with the two cohorts of students due to logistical reasons. For the 2009 cohort, the Tennessee Department of Education officials interpreted the confidentiality requirement for FRPL eligible children in a way that only allowed parents to be contacted via a mailing sent from their central office. Almost all parents who responded consented, but many did not respond (consent rate of 24.4%). For the 2010 cohort, arrangements were negotiated to allow parents to be approached about consent as part of the VPK application paperwork, and a member of the research team was available to respond to

questions (consent rate of 67.8%). Again, most of the remainder did not respond and very few actively declined to consent. The interaction between cohort and condition was included in the ISS analyses, and the interaction was not ever significant. The ISS sample also showed strong baseline equivalence on achievement pre-tests and a wide range of family characteristics. A battery of Woodcock-Johnson achievement outcome measures showed significant positive immediate effects of TN-VPK at the end of pre-k. Those effects were especially pronounced for children who entered pre-k with lower baseline scores on the achievement measures and those for whom English was not their native language.

By the end of kindergarten, most of the effects on achievement were no longer statistically significant and, in later years, nearly all had turned at least slightly negative, although generally short of statistical significance (Lipsey et al., 2018). By the end of 3rd grade, state achievement test scores for the RCT analytic sample echoed the achievement results from the ISS subsample with TN-VPK participants scoring lower than nonparticipants, significantly so in math and science. No effects were found on either attendance or grade retention rates through 3rd grade. However, children who attended TN-VPK had marginally significant higher rates of school rule violations and a significantly greater proportion of TN-VPK participants had special education placements (Lipsey et al., 2018). There were no significant effects on the quality of the schools the students subsequently attended or their exposure to higher quality teachers in those schools (Pearman et al., 2020). However, positive TN-VPK effects were found on the 3rd grade state achievement tests for the small proportion (12%) of children who attended higher quality schools *and* were exposed to higher quality teachers.

TN-VPK Effects through Sixth Grade

The current paper reports the next phase of this study, which extended the investigation of longer-term TN-VPK effects through 6th grade. Several issues were of interest for this follow-up period. Paramount was the question of whether the negative effects found on the 3rd grade state achievement tests diminished or continued in the later grades. Similar questions apply to the further development of the null or negative findings on later attendance, retention, disciplinary infractions, and special education placements.

Method

RCT Sample

This study involves 79 over-subscribed TN-VPK program sites with two cohorts of pre-k applicants randomized to offers of admission or a waitlist, one cohort entering pre-k in 2009-10, the other in 2010-11. This resulted in randomization of 111 site-level applicant lists (R-Lists). To be included in the RCT analytic sample, students had to: (1) be eligible for free or reduced price lunch, (2) be four years old by September 30 of their pre-k year, (3) be applicants to an oversubscribed TN-VPK program site that successfully randomized admission decisions, (4) not have applied for out-of-classroom special education services prior to pre-k enrollment, and (5) have a record in the state education database for at least one year of attendance in a Tennessee public school between pre-k and 3rd grade. Of the students who met criteria 1 through 4, there were 141 students who did not have subsequent state data in any year from kindergarten through 3rd grade (criterion 5). Omitting those students left a total of 2,990 eligible children in the sample used for analysis. Criterion 5 kept the RCT analytic sample the same as in previous reports; extending that criterion to include students with data from fourth

through 6th grade would have added only one student to the sample.

We report all results using two definitions of treatment and control conditions: intent-to-treat (ITT) and treatment-on-treated (TOT). ITT differentiates students according to whether they were randomly assigned to receive offers of admission. TOT differentiates students according to whether they actually attended TN-VPK or not. Figure 1 provides a visual representation of the ITT and TOT conditions and the number of children in each. The children enrolled in TN-VPK attended an average of 143.8 days (*SD* = 31.6) during the school year. All participants were treated ethically, and the Vanderbilt University Institutional Review Board (IRB #090666, "Evaluating the Effectiveness of Tennessee Voluntary Pre-K Program") approved this study.

Counterfactual Conditions

While we do not have information on the alternative care arrangements for students in the RCT analytic sample who did not attend TN-VPK, we do have that information via parent interviews for the 306 non-attending children in the ISS sample described earlier. Overall, 63% received home-based care by a parent, relative, or other person; 13% attended Head Start; 16% were in private center-based childcare; 5% had some combination of Head Start and private childcare; and childcare for 3% was not reported. Characteristics of the programs and students contributing to the ISS were very similar to those in the RCT analytic sample (Lipsey et al., 2018).

Representativeness of the RCT Analytic Sample

Another component of the Tennessee Pre-K Study not otherwise discussed here involved a statewide probability sample of the 942 TN-VPK classrooms operating at the time

the study began (Pearman et al., 2020). That sample included 155 TN-VPK classrooms purposely selected to be representative of the state program and 2,093 children in those classrooms with data on the same demographic characteristics available for the RCT analytic sample. To allow estimation of TN-VPK effects for students with the statewide demographic profile, children in the RCT analytic sample were identified with exact or very similar demographic profile matches to children in the statewide sample. Weighting functions were then created to up-weight or down-weight the children in the analytic sample to match the proportions with corresponding profiles in the statewide probability sample (see Lipsey et al., 2018 for details about the statewide probability sample demographics). This procedure allowed us to generalize the findings from oversubscribed programs in the RCT analytic to the statewide program as a whole.

RCT Analytic Sample Outcome Measures

Data providing outcome variables for the RCT analytic sample were drawn from the state database each year through 6th grade and included the following:

- Tennessee Comprehensive Assessment Program (TCAP): State achievement test during students' 3rd grade year; scaled scores for reading, math, and science.
- TNReady: TNReady replaced TCAP in 2015-16 as the state standardized achievement test. A major breakdown in the testing program when switching to TNReady resulted in a loss of 4th grade test scores for Cohort 2 and 5th grade scores for Cohort 1. We analyzed the 6th grade scaled scores for both cohorts in English language arts, math, and science.
- Violation of school rules: Any recorded violations of school policy such as attendance-related issues, dress code violations, cheating, and the like.

- Major disciplinary infractions: Possession of drugs, alcohol, or weapons, theft, vandalism, violence or threats of violence, bullying, fighting, assault, and sexual harassment.
- IEP other than for gifted or physical disability: Children with special education programming for a specific learning disability, intellectual disability, speech or language impairment, emotional disturbance, autism, functional delay, or developmental delay.
- Attendance rate: The number of days attended divided by the number of days enrolled.
- Grade level: Enrollment below vs. at or above expected grade indicating retention in grade.

Some children were not enrolled in a TN public school in some years and did not have state records those years. In kindergarten, 98.5% had state data, 96.5% in 1st grade, 95.0% in 2nd grade, 93.7% in 3rd grade, 92.5% in 4th grade, 91.3% in 5th grade, and 90.3% in 6th grade.

Analysis

Missing Data

Missing values on the variables used in analysis fell into two categories. First, 141 children in the initial randomization did not enroll in TN public schools after the pre-k year (with one late exception) and only 40 of those enrolled in TN-VPK during the pre-k year. Most of these children had no baseline data, none had outcome data, and they were dropped from the analytic sample. Second, there was missing outcome data in the analytic sample (ranging from 9.7% to 13.3%), though none on baseline variables.

The potential for differential attrition to bias treatment effect estimates was explored for the N=2990 analytic sample and the N=3131 initial sample that included the 141 cases without outcome data. First, missing data rates for outcome variables were compared for the ITT treatment vs. control conditions. The differences for the analytic sample were all less than 2

percentage points; for the initial sample, all were less than 3 percentage points. None of these differences were statistically significant at alpha=.05, but significance at $p<.10$ was found for major disciplinary infractions and the combined measure of major infractions and school rule violations in the analytic sample and for all three disciplinary measures in the initial sample.

Although missing data rates for the treatment and control conditions were generally comparable, consideration was also given to the possibility that the missing outcome data were different for each condition, which also could bias effect estimates. For this purpose, we imputed estimates of the missing values and compared effect estimates with and without the imputed data. Twenty-five imputed data sets were generated separately for the ITT treatment and control conditions and combined for analysis. The analysis models described in the next section were then used to estimate treatment effects for each outcome variable with the observed and imputed values for the analytic and initial samples.

With two exceptions, the effect estimates for the observed and imputed data from the analytic sample were closely comparable for direction, magnitude, and statistical significance. The exceptions were the proportion of children with an IEP in 6th grade and the proportion with any major disciplinary offense over the K-6th grades. In both cases, the estimates with imputed values were smaller and nonsignificant compared to larger significant ones with observed values. However, these effect estimates were very small, ranging from .014 to .033. For the initial sample, the direction, magnitude, and statistical significance of the effect estimates for the observed and imputed data were also comparable across outcomes with the exception of IEP and major disciplinary offenses, but also for attendance. Again, the estimates were very small, ranging from .001 to .033, and of dubious practical significance. Notably, the critical

achievement test effect estimates were quite comparable across these various comparisons.

Details about these analyses and their results are in Supplemental Materials 1. Given how little evidence was found of consequential differentials in the proportions of missing values or the effects found with and without imputation, we have chosen to focus this report on the results found with the observed data in the analytic sample.

Analysis Models

Analysis of TN-VPK effects used hierarchical linear models (HLM) with students nested in the R-List program sites from which applicants were randomized and those nested in their respective school districts. Equations for the HLM models and the corresponding SPSS syntax are in Supplemental Materials 2. A reanalysis of TN-VPK effects through 3rd grade using block fixed effects models as an alternative approach resulted in findings comparable to those with HLM (Watts et al., 2019). Many of the outcome variables of interest are binary (e.g., whether a student is on grade level). For consistency, the HLM results are presented here. Confirmation that they are substantially similar to the results of more technically appropriate multilevel logistic regression analyses are in Supplemental Table S1.

All the analyses of TN-VPK effects (unless otherwise indicated) incorporated what we will refer to as the "standard set" of covariates to adjust for baseline differences, improve statistical power, and provide a basis for moderator analysis. These included age at the beginning of the pre-k year, gender (male), race/ethnicity (White, Black, Hispanic), and native language other than English. At Level 2, to aid statistical power, we included TN-VPK program site characteristics identified by representatives of the Tennessee Department of Education as relevant to program performance: Urbanicity (urban vs. rural), Partner programs (operated by

community organizations vs. schools), Priority schools (operated in the lowest performing schools), Pilot programs (funded in 1996 as pilot pre-k programs), and Region (west, central west, central east, and east parts of the state)[1]. Further analyses tested interactions between treatment condition and the student demographic variables to determine whether any of those variables moderated the effects of TN-VPK. These analyses involved a large number of statistical tests and Benjamini-Hochberg corrections for false positive rates were applied.

The primary analyses involved ITT comparisons with the original observed data and repeated with the weighting functions derived from the statewide probability sample described earlier. A principal stratification strategy modeled on that used in the Head Start Impact Study (Puma et al., 2010) was used to generate complier average causal effect (CACE) estimates from the ITT estimates. This procedure recognizes that the ITT treatment and control groups include four distinct subgroups differentiated by how they react to the randomization: *Compliers* who accept the condition to which they are randomized; *Always Takers* who obtain treatment whether randomized to it or not; *Never Takers* who do not participate in treatment whether randomized to it or not; and *Defiers* who choose the opposite of their assigned condition.

Some reasonable assumptions are that there are no or trivially few Defiers (unlikely that parents who apply for TN-VPK will reject admission if offered but obtain it if not offered), and that the expectation from randomization will yield the same proportion and characteristics for Always Takers in both the ITT treatment and control groups (latter called Crossovers). Along

[1] This set of covariates is more extensive than those used in analyses reported earlier (Lipsey et al., 2018) and, by using more information, resulted in somewhat better fitting models. As a consequence of this change, third grade results reported here do not agree exactly with those reported earlier, although the patterns remain the same.

with a few technical assumptions, these allow derivation of a multiplier that scales the ITT effect estimates into CACE estimates (Gennetian et al., 2005; Puma et al., 2010, p. 5–53). This procedure also rescales the standard errors of the ITT estimates with the same multiplier so the statistical significance of the CACE estimates is the same as for the corresponding ITT estimates. These scaled up estimates have been shown to be equivalent to the CACE estimates derived by an alternative approach using two-stage least-squares instrumental variable analysis with random assignment as the instrument (Angrist et al., 1996).

As effect estimates for Compliers, the CACE estimates omit the Always Taker subgroups that also participated in TN-VPK and would thus need to be included in full TOT estimates. While effects for Always Takers cannot be estimated directly, their outcomes on the state achievement tests were compared with those for control condition Compliers and for treatment condition No Shows to provide some general indication of their potential magnitude. Those estimates fell well within the confidence intervals for the CACE estimates. On that basis, we assume the effects for the Always Takers are comparable to the CACE estimates and have interpreted the CACE estimates as TOT estimates. Details for the CACE and TOT derivations are provided in Supplemental Materials 3.

Baseline Equivalence between Conditions

The state administrative data system does not collect data on TN-VPK applicants at the beginning of the pre-k year. Descriptive baseline variables for the RCT analytic sample were thus limited to demographic and program site characteristics. Baseline equivalence comparisons for the ITT treatment and control groups on student demographic characteristics in the observed data are presented in Table 1 (this comparison does not apply to program sites,

only within sites). There were no statistically significant differences on any of the demographic variables and the effect sizes were small, indicating substantial similarity between the respective treatment and control groups. These analyses were repeated with the weighting function applied and with multilevel logistic regression, which also revealed no significant baseline differences (Supplemental Tables S2 and S3). As a further check, baseline differences on student demographics were examined for the data contributing to each outcome taking attrition into account. None of these differences were significant (Supplemental Table S4).

Study materials are available at https://my.vanderbilt.edu/tnprekevaluation/ and deidentified datasets are available by request from the first author. This study was not preregistered.

Results: RCT Analytic Sample

Academic Performance

State Achievement Tests

Most students in the RCT analytic sample had scores on the state achievement tests first administered in 3rd grade. TN-VPK effects on those scores were reported previously (Lipsey et al., 2018), but we include them here to allow comparison with the later 6th grade scores and because the weighting function and covariates were updated from previous reports. The first two panels of Table 2 present ITT and TOT results for 3rd and 6th grade; the second two panels present analogous results with the observed values weighted to match the demographic profile of the statewide TN-VPK population. The 3rd grade results show that control children outperformed TN-VPK children across the three subject areas, with those differences statistically significant for mathematics and science in the unweighted analyses and for all three

tests in the weighted analyses. On the 6th grade TNReady tests, control children continued to outperform the TN-VPK children in reading, mathematics, and science with statistically significant differences larger than those observed in 3rd grade. These effects were similar when examining only the ISS sample (Table 3, see analysis details in Supplemental Table S5).

The loss of 4th grade state achievement test scores for Cohort 2 and 5th grade scores for Cohort 1 resulting from the lapse in testing when the TCAP was replaced by the TNReady tests precluded year-to-year comparisons for the analytic sample. Figure 2 charts the available reading and math scores for the cohorts with state data at each grade level. Supplemental Table S6 provides more detail for these estimates. These findings do not change when students who were retained or promoted a grade level are excluded (Supplemental Table S7).

Exploration of differential effects on the state achievement tests in 3rd and 6th grade found no statistically significant interactions of ITT treatment condition with age, gender, White, Black, or non-native English language. However, there were significant interactions for Hispanic children in 3rd grade for the weighted analysis for reading (B = 14.63, SE = 4.57, t = 3.20, p = .020) and marginally for mathematics and science after Benjamini-Hochberg adjustments (B = 13.26, SE = 4.74, t = 2.79, p = .098 and B = 12.50, SE = 4.74, t = 2.64, p = .125, respectively). Results for the unweighted models were similar. In 3rd grade, Hispanic students who did not participate in TN-VPK performed better than those who did participate; non-Hispanic students who did not participate in TN-VPK performed the same as those who did participate. In 6th grade, there were no significant interactions for any of the achievement tests.

Retention in Grade and Special Education

Whether students were in the expected grade level in 6th grade (had not been retained)

was represented with a binary variable, 0 for below the expected level and 1 if at that level or (rarely) above. As Table 4 reports, 87.2% of the TN-VPK participants in the analytic sample and 88.1% of the nonparticipants were at grade level. There were no significant differences between these groups in retention with either the weighted or unweighted analysis, a finding confirmed with multilevel logistic regression (Supplemental Table S1).

Students were coded as 1 if they had any IEP except for gifted or physical disabilities in a given year, and 0 if they did not. As Table 4 reports, more TN-VPK participants (11.7%) had an IEP compared to nonparticipants (8.4%) in 6th grade. These differences were significant in both the weighted and unweighted analysis, and in parallel multilevel logistic regression (Supplemental Table S1). These effects were similar when examining the ISS sample compared to the RCT analytic sample (Table 3, see analysis details in Supplemental Table S8).

Although our focus is on 6th grade outcome data, trends across the years are also informative. From kindergarten through 6th grade, the treatment and control groups in the analytic sample had similar retention rates (Supplemental Table S9 and Supplemental Figure S1). One exception was 1st grade when more TN-VPK participants were at expected grade level (fewer retained in kindergarten). However, by 2nd grade the conditions were similar again.

Because the state department of education administers TN-VPK and most of the classrooms are in elementary schools, TN-VPK participants have an early extra year in which to be screened and identified for special education services, and once identified, generally maintain that status for several years. More TN-VPK participants than nonparticipants had an IEP in kindergarten, continuing into 6th grade (Supplemental Table S10 and Supplemental Figure S2). Moreover, the proportion of students in the control group with IEPs began trending down

in 2nd grade while that for TN-VPK participants was more stable over time.

Behavioral Outcomes

Attendance and Attendance Trends

Attendance rates in 6th grade (proportion of instructional days without a recorded absence) were high for both TN-VPK participants and nonparticipants. Nonetheless, the difference between groups was statistically significant with a slightly higher rate for nonparticipants (97.5% vs. 97.1%, p = .013 for the ITT analysis with observed values). Supplemental Table S11 provides model details for each year (see also Supplemental Figure S3). Sixth grade was the first academic year with a significant attendance difference between conditions, although there were marginally significant effects in kindergarten and 1st grade.

Disciplinary Infractions

The frequency of expulsions, in-school suspensions, and out-of-school suspensions increased across the school years but was relatively low in any one school year. To summarize, we created outcome variables that indicated whether any such events were recorded across the kindergarten to 6th grade years (1 if any offenses, 0 in none). We also differentiated events classified as minor or major offenses. All analyses revealed higher rates of recorded disciplinary events for TN-VPK participants than nonparticipants, and these differences were statistically significant except for the weighted analysis for major offenses (Table 5; multilevel logistic results in Supplemental Table S1). These effects were also similar when examining the ISS sample (Table 3, see analysis details in Supplemental Table S12). There were no significant interactions between ITT condition and demographics for these outcomes. The offense rates are graphed across grades in Figure 3 (detail in Supplemental Table S13). The differences

between conditions grew larger each year, particularly for school rule violations.

Discussion

We have presented the results through 6th grade of the effects of a scaled-up, state-funded prekindergarten program, the only randomized control study of a statewide pre-k program to date. As reported in prior papers (Lipsey et al., 2018; Pion & Lipsey, in press), the effects of TN-VPK on individually assessed early achievement measures at the end of the pre-k school year were strong, especially on literacy measures. Those results are thus similar to the findings of multiple age-cutoff regression-discontinuity studies that have become the most common research model for assessing end of pre-k effects (e.g., Hustedt et al., 2021; Montrosse-Moorhead et al., 2019; Wong et al., 2008).

Followed over time, however, the TN-VPK effects disappeared by the end of kindergarten and turned negative by the end of 3rd grade (Lipsey et al., 2018). Subsequently the achievement effects have increased in negative magnitude across the years and been moderate to strong. We found the same increasingly negative trend for disciplinary infractions and, by 6th grade, for attendance. For retention in grade we found no effect, and for the need for special education services we found a negative effect.

Quality of the TN-VPK Program

If TN-VPK is quite different from those implemented in other states, it could mean our results are limited to Tennessee. As we report, the statewide scale up of TN-VPK began in 2005 after nearly 10 years of pilot testing and met 9 of the 10 NIEER benchmarks (Barnett et al., 2009). A recent review of statewide programs by the NIEER group (Friedman-Krauss et al., 2019) praised the program in Tennessee for being among those in 27 states that paid its pre-k

teachers at parity with elementary teachers, one of only 26 states to offer pre-k teachers retirement benefits, health care and paid time off, and one of only 25 to require its teachers to have a bachelor's degree plus certification. Among state-funded pre-k programs, the TN program is above average and arguably in the top tier on characteristics many believe mark high quality (Sharpe et al., 2017).

However, as Bassok and Engel noted (2019), "there is surprisingly little consensus on the specific characteristics or combination of programmatic features that are most essential for ensuring the effectiveness of ECE programs" (p. 4). Judged alternatively by its performance in producing student gains on commonly measured cognitive outcomes, an age-cutoff regression-discontinuity substudy has also found that TN-VPK ranks as a top tier program when compared to the results of similar studies in other states (Pion & Lipsey, in press). While it is an open question whether results similar to those found for TN-VPK would be found with a similarly rigorous long-term evaluation of any other state program, no distinctive characteristics of the Tennessee program have yet been identified that are a likely explanation for the disappointing findings. It is important therefore to explore other potential explanations of a more general sort.

Reversal of Initial Positive Pre-K Effects

Our results are stronger than, but not dissimilar to, those from the Head Start Impact study (Puma et al., 2010) and other long-term follow up assessments without random assignment (see Bailey et al., 2017, Bailey et al., 2020 for reviews). Almost all early childhood interventions show initial positive effects and almost all show substantial fade out of effects, some immediate as for TN-VPK, others taking somewhat longer to emerge. As the only

longitudinal evaluation of a statewide program with random assignment, this TN-VPK study is also the only one to date to show long-term *negative* effects. When we reported the 3rd grade results, there was concern about this unexpected finding. We anticipate that these results through 6th grade will heighten those concerns. While it is always speculative to explain unanticipated findings, it is important to offer some possible avenues for consideration. As mentioned earlier, there were no significant differences in the quality of the schools and teachers that VPK and control students subsequently experienced after pre-k (Pearman, 2020), so differences in the quality of schools treatment and control children attended is not a possible explanation for these findings.

Constrained versus Unconstrained Academic Skills

One contributor to the fade out of pre-k effects may involve the content focus of the instruction children receive, an idea recently gaining traction. Evaluating eight statewide pre-k programs, Barnett and colleagues (2018) found, as we did, that the largest immediate effect was in concrete literacy skills, with much smaller effects on language and math skills. They urge pre-k programs to broaden their scope of instruction. These early concrete literacy skills include directly teachable skills in a finite domain (e.g., 26 letters of the alphabet): "constrained skills have a ceiling; the learner can achieve perfect performance" (Snow & Matthews, 2016, p. 58).

Unconstrained skills in literacy (vocabulary, listening comprehension, and background knowledge) and in numeracy (problem solving and mathematical reasoning) are not typically the focus of instruction in early childhood classrooms (Montrosse-Moorhead et al., 2019; Valentino, 2017), perhaps because they are not the usual content of assessments amid the increasing emphasis on "school readiness." Over time, these skills become increasingly

important in school, but they are more difficult to teach and assess (Snow & Matthews, 2016).

A consistent finding across recent studies is that children who attend pre-k enter kindergarten scoring higher on concrete school readiness skills, skills that are then mastered by non-attenders over the course of the kindergarten year or shortly thereafter. The early childhood field has not been successful so far in identifying classroom characteristics and interactions linked to improvements in unconstrained skills (e.g., Guerrero-Rosada et al., 2021) although practices linked to gains in school readiness skills have been identified (e.g., Farran et al., 2017). In 2020, Bailey and colleagues addressed this perplexing finding by stressing the importance of targeting "trifecta" skills – "ones that are malleable, fundamental, and would not have developed in the absence of intervention" (Bailey et al., 2020, p. 66-67). They argue that the early childhood field must first answer the question about which fundamental and malleable outcomes pre-k should aim to improve if longer-term effects are to be attained.

Attention and Working Memory – Two Important Unconstrained Skills

The interest in unconstrained skills has thus far focused almost exclusively on specific academic outcomes like vocabulary and certain math skills; other more fundamental skills may be equally or more important. Many statewide programs target children of low-income families. Studies over the last 20 years or so have demonstrated the devastating effects of poverty on the developing brain (Brito & Noble, 2015; Yaple & Yu, 2020), particularly in the areas of language and executive function. Among the executive function skills that appear most affected are working memory and attention (Lupina & Posner, 2012). Moreover, differences in working memory continue to be associated with SES well into adolescence (Judd et al., 2020). Despite the focus of targeted pre-k programs on children from high poverty families, with rare

exceptions the early childhood field has not taken these neuroscience findings into account. Even a pre-k curriculum focused on the development of executive function skills failed to show short or long-term effects on any of those skills that were measured (Nesbitt & Farran, 2021).

The possible benefits from developing good strategies for affecting children's development in working memory and attention are illustrated in recent research. For example, improvements in working memory have been identified as a critical factor in children moving from a reliance on "reactive" to one of "proactive" cognitive control (Troller-Renfree et al., 2020), the latter associated with more planful learning strategies. Further, in a longitudinal study of mathematics achievement, Geary and colleagues (2017) demonstrated that working memory in early childhood emerged as the most important domain-general ability associated with performance in later grades.

In addition to working memory, early attention skills are related to SES and important for long-term development. A review by Duncan and colleagues (2007) of five major longitudinal studies identified early measures of attention as one of three key predictors of long-term outcomes. More recent research has shown that attention skills in early childhood appear to be composed of two factors: selective-sustained attention and an executive factor. Only sustained-selective attention related to gains in pre-literacy and math skills (Shannon et al., 2020; Steele et al., 2012). "The current findings therefore provide a direct demonstration that cognitive building blocks to early numeracy and literacy depend on effortful control in early childhood" (Steele et al., 2012, p. 2039).

Working memory and attention may indeed be the relevant building blocks among the unconstrained skills that underlie development of more academic outcomes. However, little is

known about how to facilitate their development in a classroom context. This issue of which skills are foundational and thus the most important focus for instruction is a critical one as we examine long-term effects from pre-k attendance (Green, 2020). State legislatures believe pre-k will positively affect 3rd grade reading scores and thus long-term school achievement, closing the gap. If those are indeed the desired outcomes, the early childhood field must identify the fundamental skills that relate to these outcomes and determine when and how those skills can be positively affected. The rush to implement statewide programs and the focus on initial school readiness concrete skills have meant that these important steps were not carried out.

Negative Behavioral Outcomes Associated with Attending Pre-K

Apart from the lack of positive effects on achievement, an unexpected finding important to explore further is the negative behavioral outcomes. Here also the results for TN-VPK are not at odds with findings from other studies of children who experience group care in early childhood. One outcome reported for the Abecedarian program was more aggressive behavior for program participants in the early grades (Haskins, 1985). Similar findings have emerged in the two ECLS-K samples where both cohorts were found to exhibit more externalizing behaviors and less self-control if they had any type of formal care before kindergarten (Bassok et al., 2015). These findings were replicated and extended through age 15 in the NICHD Study of Early Childcare and Youth Development where children who experienced more care outside the family prior to school entry were greater risk takers and more impulsive (Vandell et al., 2010). Lest we think that these negative outcomes are a function of earlier outdated versions of pre-k, a study from the current IES Early Learning Network reported that children who attended pre-k had higher rates of kindergarten teacher-reported conflict and lower rates of task orientation

(Ansari et al., 2021). Our findings of higher rates of school disciplinary infractions for pre-k participants provide further support for this as an issue that warrants serious attention.

Searching for possible explanations of this common outcome, however, has not been immediately fruitful. Moffitt and colleagues (2011) found that early measures of self control were predictive of health and financial outcomes when individuals were in their 30's. The finding was robust after controlling for social class and early measures of IQ. Some children developed more self control in early childhood with subsequent better outcomes via what Moffitt calls a "natural history change." Whether an intervention-induced change would yield the same positive outcomes is an open question.

One possibility is that center-based care (the common denominator for studies with negative behavioral outcomes) could be preventing children from developing the internal self-control necessary for long-term development. In particular, classrooms of 20 4-year-olds require behavioral control exerted by adults. Studies demonstrate that teachers in these circumstances often display a flat to negative affect (Coelho et al., 2021; Farran et al., 2017), one that could lead to children developing negative attentional biases. Negative attentional biases have been associated with increased reactivity to later stressors (Todd et al., 2012).

The long-term negative outcomes on behavior for children in group care have been found in both small experimental studies and broad-based population studies. Determining their etiology and creating classroom practices that yield different outcomes is critical for programs that serve children from low-income families, but efforts so far have proven unsuccessful (Morris et al., 2014). Several large-scale studies and a meta-analysis have demonstrated the long-term negative effects from the early school suspensions we found in

TN-VPK (Mendez, 2003; Mowen et al., 2020; Noltemeyer et al., 2015). School suspensions, even though most are for nonviolent infractions, are associated with lower academic achievement in later grades and eventually dropping out of school.

Conclusion

The randomized control study of the effects of a scaled-up statewide pre-k program reported here provides results through the end of 6th grade that should lead, at minimum, to questions about the content and pedagogical strategies currently employed in pre-k classrooms nationally. Kindergarten readiness on constrained skills was demonstrated in this pre-k program as it has been in many others. Longer-term effects are not so sanguine. Our results are robust and contrary to the claims made by many advocates for the universally positive effects of pre-k participation. Children from poor families who attended a state pre-k program did not, for the most part, become proficient readers in 3rd grade. On the contrary, their performance on all measures of achievement through 6th grade was significantly below that of comparable children who did not attend. Children who attended pre-k were not less likely to be retained and had a greater likelihood of being referred for special education services from pre-k through 6th grade – both of these in opposition to savings promised to states (Council of Economic Advisers, 2015). Given prior research, our findings of more disciplinary infractions for children in 6th grade who attended pre-k should not have been so unexpected but are nonetheless worrisome.

The whole package of outcomes we have found is disconcerting. The intent of everyone who has advocated for expansion of state pre-k programs is well meaning and reflects a commitment to improving the life outcomes for children from impoverished circumstances. If the programs we have created do not produce the desired effects, the findings themselves

should not be dismissed simply because they were unanticipated and unwelcome. Rather, they should stimulate creative research into both policies and practices with potential to have the desired effects. The goal remains the same. If we are serious about the goal, the means to attain it may have to change.

References

Administration for Children and Families [ACF] (2019). *Head Start federal funding and funded enrollment history.* https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/hs-federal-funding-enrollment-history.pdf

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-455. https://doi.org/10.2307/2291629

Ansari, A., & Crosnoe, R. (2015). Immigration and the interplay of parenting, preschool enrollment, and young children's academic skills. *Journal of Family Psychology, 29(3),* 382-393. https://doi.org/10.1037/fam0000087

Ansari, A., Pianta, R. C., Whittaker, J. E., Vitiello, V., & Ruzek. E. (2021). Enrollment in public-prekindergarten and school readiness skills at kindergarten entry: Differential associations by home language and program characteristics. *Early Childhood Research Quarterly, 54,* 60-71. https://doi.org/10.1016/j.ecresq.2020.07.011

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest. 21*(2), 55–97. https://doi.org/10.1177/1529100620915848

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. https://doi.org/10.1080/19345747.2016.1232459

Barnett, S., Epstein, D. J., Friedman, A. H., Sansanelli, R. A., & Hustedt, J. T. (2009). *The state of preschool 2009: State preschool yearbook* (Tennessee pp. 132-133). The National Institute for Early Education Research [NIEER]. Rutgers University. http://nieer.org/wp-

content/uploads/2016/10/200920yearbook-1.pdf

Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., Howes, C., & Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open, 4*(2), 1-16. https://doi.org/10.1177/2332858418766291

Bartik, T. J., & Hershbein, B. (2018). *Pre-k in the public schools: Evidence from within U.S. states*. Upjohn Institute Working Paper 18-285. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/wp18-285

Bassok, D., & Engel, M. (2019). Early childhood education at scale: Lessons from research for public policy and practice. AERA Open, 5(1), 1-7. https://doi.org/10.1177/2332858419828690

Bassok, D., Gibbs, C., & Latham, S. (2015). Do the benefits of early childhood interventions systematically fade? Exploring variation in the association between preschool participation and early school outcomes. EdPolicyWorks Working Paper Series No. 36. https://curry.virginia.edu/working-paper-preschool-fade-out

Brito, N. H., & Noble, K. G. (2015). Socioeconomic status and structural brain development. *Frontiers in Neuroscience, 8,* 276. https://doi.org/10.3389/fnins.2014.00276

Cascio, E. U., & Schanzenbach, D. W. (2013). The impacts of expanding access to high quality preschool education. *Brookings Papers on Economic Activity*, Fall, 127-192. https://doi.org/10.3386/w19735

Coelho, V., Aström, F., Nesbitt, K., Sjöman, M., Farran, D., Björck-Äkesson, E., Christopher, C., Granlund, M., Almqvist, L., Grande, C., & Pinto, A. (2021). Preschool practices in Sweden, Portugal, and the United States. *Early Childhood Research Quarterly, 55,* 79-96. https://doi.org/10.1016/j.ecresq.2020.11.004

Coley, R. L., Votruba-Drzal, E., Collins, M., & DeMeo Cook, K. (2016). Comparing public, private, and informal preschool programs in a national sample of low-income children, *Early Childhood Research Quarterly, 36*, 91-105. http://dx.doi.org/10.1016/j.ecresq.2015.11.002

Council of Economic Advisers. (2015). *The economics of early childhood investments*. https://obamawhitehouse.archives.gov/sites/default/files/docs/early_childhood_report_update_final_non-embargo.pdf

Deming, D. (2009). Early childhood intervention and life cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134. https://doi.org/10.1257/app.1.3.111

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A.C., Klebanov, P., . . . Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Early, D. M., Li, W., Maxwell, K. L., & Ponder, B. D. (2019). Participation in Georgia's pre-k as a predictor of third-grade standardized test scores. *AERA Open, 5*(2), 1-16. https://doi.org/10.1177/2332858419848687

Evaluating the Effectiveness of Tennessee's Voluntary Pre-K Program. (n.d.). https://my.vanderbilt.edu/tnprekevaluation/

Farran, D. C., & Lipsey, M. W. (2015). Expectations of sustained effects from scaled up pre-K: Challenges from the Tennessee study. *Evidence Speaks Reports*, *1*(3).

Farran, D. C., Meador, D., Christopher, C., Nesbitt, K. & Bilbrey, L. (2017*).* Data-driven improvement in prekindergarten classrooms: Report from a partnership in an urban district. *Child Development, 88,* 1466-1479. DOI: 10.1111/cdev.12906

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G., & DiCrecchio,

N. (2019). *The state of preschool 2018: State preschool yearbook* (NIEER]. Rutgers University. http://nieer.org/wp-content/uploads/2019/08/YB2018_Full-ReportR3wAppendices.pdf

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*. The National Institute of Early Education Research [NIEER]. Rutgers University. http://nieer.org/state-preschool-yearbooks/yearbook2017

Geary, D. C., Nicholas, A., Li, Y., & Sun, J. (2017). Developmental change in the influence of domain-general abilities and domain-specific knowledge on mathematics achievement: An eight-year longitudinal study. *Journal of Educational Psychology, 109*(5), 680-693. https://doi.org/10.1037/edu0000159

Gennetian, L. A., Morris, P. A., Bos, J. M., & Bloom, H. S. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H. Bloom (Ed). *Learning more from social experiments: Evolving analytic approaches* (pp. 75-114). Russell Sage Foundation.

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

Green, C. S. (2020). Interventions to do real-world good: Generalization and persistence. *Psychological Science in the Public Interest, 21*(2), 43–49. https://doi.org/10.1177/1529100620933847

Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children's language, math, and executive function skills: A replication and extension study. *Early Childhood Research Quarterly, 54,* 1-

12. https://doi.org/10.1016/j.ecresq.2020.07.009

Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York, NY: Teachers College Press.

Haskins, R. (1985). Public school aggression among children with varying day-care experience. *Child Development, 56*(3), 689-703. https://doi.org/10.2307/1129759

Hill, C. J., Gormley, W. T., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly, 32*, 60-79. http://dx.doi.org/10.1016/j.ecresq.2014.12.005

Hustedt, J., Jung, K., Friedman-Krauss, A., Barnett, S. & Slicker, G. (2021). Impacts of the New Mexico pre-k initiative by children's race/ethnicity. *Early Childhood Research Quarterly, 54,* 194–203. https://doi.org/10.1016/j.ecresq.2020.09.006

Judd, N., Sauce, B., Wiedenhoeft, J., Tromp, J., Chaarani, B., Schliep, A., van Noort,B., Penttilä, J., Grimmer, Y., Insensee, C., Becker, A., Banaschewski, T., Bokde, A., Quinlan, E., Desrivières, S., Flor, H., Grigis, A., Gowland, P., Heinz, A., . . . Klingberg, T. (2020). Cognitive and brain development is independently influenced by socioeconomic status and polygenic scores for educational achievement. *PNAS, 117,* 12411-12418. www.pnas.org/cgi/doi/10.1073/pnas.2001228117

Ladd, H. F., Muschkin, C. G., & Dodge, K. A. (2014). From birth to school: Early childhood initiatives and third-grade outcomes in North Carolina. *Journal of Policy Analysis and Management, 33*(1), 162-187. https://doi.org/10.1002/pam.21734

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*, 155-176. https://doi.org/10.1016/j.ecresq.2018.03.005

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013a). Evaluation of the Tennessee Voluntary Prekindergarten Program: End of pre-K results from the randomized control design. Research report. Nashville, TN: Vanderbilt University, Peabody Research Institute. https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/May2013_PRI_EndofPK_TN-VPK_RCT_ProjectResults.pdf

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013b). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and first grade follow-up results from the randomized control design. Research report. Nashville, TN: Vanderbilt University, Peabody Research Institute. https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

Lupina, S. J., & Posner, M. I. (2012). The impact of poverty on the development of the brain. *Frontiers in Human Neuroscience. 6,* 1-12. https://doi.org/10.3389/fnhum.2012.00238

Mendez, L. M. R. (2003). Predictors of suspension and negative school outcomes: A longitudinal investigation. *New Directions for Youth Development, 2003*(99), 17-33. https://doi.org/10.1002/yd.52

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS, 108(7),* 2693-2698. https://doi.org/10.1073/pnas.1010076108

Montrosse-Moorhead, B., Dougherty, S., La Salle, T., Weiner, J., & Dostal, H. (2019). The overall and differential effects of a targeted prekindergarten program: Evidence from Connecticut. *Early Childhood Research Quarterly, 48,* 134-145. https://doi.org/10.1016/j.ecresq.2019.02.006

Morris, P., Mattera, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). *Impact Findings from the Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence.* OPRE Report 2014-44. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Mowen, T. J., Brent, J. J., & Boman IV, J. H. (2020). The effect of school discipline on offending across time. *Justice Quarterly, 37*(4), 739-760. https://doi.org/10.1080/07418825.2019.1625428

Nesbitt, K. & Farran, D. C. (2021). Effects of prekindergarten curricula: *Tools of the Mind* as a case study. *Monographs of the Society for Research in Child Development*, *86*(1)*,* 7-119. https://doi.org/10.1111/mono.12425

Noltemeyer, A. L., Ward, R. M., & Mcloughlin, C. (2015). Relationship between school suspension and student outcomes: A meta-analysis. *School Psychology Review, 44*(2), 224–240. https://doi.org/10.17105/spr-14-0008.1

Pages, R., Lukes, D. J., Bailey, D. H., & Duncan, G. J. (2020). Elusive longer-run impacts of Head Start: Replication within and across cohorts. *Educational Evaluation and Policy Analysis, 42*(4), 471-492. https://doi.org/10.3102/0162373720948884

Pearman, F. A., Springer, M. P., Lipsey, M., Lachowicz, M., Swain, W., & Farran, D. (2020). Teachers, schools, and pre-K effect persistence: An examination of the sustaining environment hypothesis. *Journal of Research on Educational Effectiveness*, *13*(4), 547-573. https://doi.org/10.1080/19345747.2020.1749740

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects: A consensus statement*. Washington, DC: The

Brookings Institution.

Pion, G. M., & Lipsey, M. W. (in press). Impact of the Tennessee Voluntary Prekindergarten Program on children's literacy, language, and mathematics skills: Results from a regression-discontinuity design. *AERA Open*. https://doi.org/10.1177/23328584211041353

Preskitt, J., Johnson, H., Becker, D., Ernest, J., Fifolt, M., Adams, J., Strichik, T., Ross, J., & Sen, B. (2020). The persistence of reading and math proficiency: The benefits of Alabama's pre-kindergarten program endure through elementary and middle school. *International Journal of Child and Education Policy, 14,* 1-12. https://doi.org/10.1186/s40723-020-00073-3

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study. Technical report.* Washington DC: US Department of Health and Human Services, Administration for Children and Families.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, D., Mashburn, A., & Downer, J. (2012). *Third Grade Follow-up to the Head Start Impact Study Final Report,* OPRE Report # 2012-45, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Rosinsky, K. (2014). *The relationship between publicly funded preschool and 4th grade math test scores: A state-level analysis*. Georgetown University, Washington, DC. https://repository.library.georgetown.edu/handle/10822/709852

Shannon, K. A., Scerif, G., & Raver, C. C. (2020). Using a multidimensional model of attention to predict low-income preschoolers' early academic skills across time. *Developmental Science, 24*(2), 1-16. https://doi.org/10.1111/desc.13025

Sharpe, N., Davis, B. & Howard, M. (2017). *Indispensable policies and practices for high-quality pre-k: Research and pre-k standards review.* New America Foundation.

https://www.newamerica.org/education-policy/policy-papers/indispensable-policies-practices-high-quality-pre-k/

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *The Future of Children, 26*(2), 57-74. https://www.jstor.org/stable/43940581

Steele, A., Karmiloff-Smith, A., Cornish, K., & Scerif, G. (2012). The multiple subfunctions of attention: Differential developmental gateways to literacy and numeracy. *Child Development, 83*(6), 2028-2041. https://doi.org/10.1111/j.1467-8624.2012.01809.x

Tennessee Department of Education [TDOE]. (2019). *Scope of Services for Voluntary Pre-K 2019-20.* https://www.tn.gov/content/dam/tn/education/early-learning/pre-k/prek_scope_of_services.pdf

Todd, R. M., Cunningham, W. A., Anderson, A. A., & Thompson, E. (2012). Affect-biased attention as emotion regulation. *Trends in Cognitive Sciences, 16*(7), 365-372. https://doi.org/10.1016/j.tics.2012.06.003

Troller-Renfree, S. V., Buzzell, G. A., & Fox, N. A. (2020). Changes in working memory influence the transition from reactive to proactive cognitive control during childhood. *Developmental Science, 23*(6), 1-9. https://doi.org/10.1111/desc.12959

Valentino, R. (2017). Will public pre-k really close achievement gaps? Gaps in prekindergarten quality between students and across states. *American Educational Research Journal, 55*(1), 79-116. https://doi.org/10.3102/0002831217732000

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 childcare and youth development. *Child Development, 81*(3), 737-56. https://doi.org/10.1111/j.1467-8624.2010.01431.x

Watts, T., Duncan, G. J., and Rivas, M. (2019). *A reanalysis of impacts of the Tennessee Voluntary Prekindergarten Program*. (EdWorkingPaper: 19-28). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/mzk4-jk96

Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development*.*91,* 1401-1422. https://doi.org/10.1111/cdev.13308

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*(1), 122–154. https://doi.org/10.1002/pam.20310

Yaple, Z. A., & Yu, R., (2020). Functional and structural brain correlates of socioeconomic status. *Cerebral Cortex, 30*(1), 181-196. https://doi.org/10.1093/cercor/bhz080

Table 1

Intent-to-Treat (ITT) Treatment and Control Comparisons on Baseline Variables (RCT Analytic Sample, Observed Data)

Variable	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
Age (months)	53.21	53.30	3.47	-.087	-.025	.507
Gender (male)	.50	.49	.50	.006	.012	.752
White	.67	.68	.50	-.010	-.019	.578
Black	.20	.20	.45	.001	.003	.941
Hispanic	.14	.13	.41	.007	.018	.639
Non-native English	.14	.13	.41	.008	.019	.617
	N = 1852	*N* =1138				

[a] Estimated marginal means from multilevel analysis models.

[b] Pooled treatment and control group standard deviations.

[c] Coefficients for the treatment-control differences from multilevel models predicting each baseline variable with ITT condition as the only predictor.

[d] Effect size: Coefficient for the treatment-control difference divided by the pooled standard deviation.

Table 2

Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Third and Sixth Grade State Achievement Tests (RCT Analytic Sample)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p* - value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
				Third Grade TCAP (Observed Values)				
Reading	746.1	748.2	34.34	-2.13	-.062	.146	-4.05	-.118
Math	755.9	760.2	35.56	-4.22*	-.119	.006	-8.02*	-.225
Science	748.6	752.2	35.33	-3.58*	-.101	.016	-6.80*	-.192
	N = 1505-1506	*N* = 935-936		*N* = 2440-2442				
				Sixth Grade TNReady (Observed Values)				
ELA	321.2	325.0	29.86	-3.83*	-.128	.002	-7.18*	-.240
Math	317.1	323.6	36.31	-6.46*	-.178	.000	-12.12*	-.333
Science	750.4	755.6	39.37	-5.18*	-.132	.002	-9.83*	-.249
	N = 1615-1630	*N* = 976-996		*N* = 2591-2626				
				Third Grade TCAP (Weighted Observed Values)				
Reading	746.9	750.1	33.59	-3.26*	-.097	.027	-6.19*	-.184
Math	755.6	761.0	34.84	-5.40*	-.155	.000	-10.24*	-.293
Science	750.0	754.1	35.48	-4.03*	-.114	.008	-7.64*	-.215
	N = 1505-1506	*N* = 935-936		*N* = 2440-2442				

Table 2 (continued).

	Sixth Grade TNReady (Weighted Observed Values)							
ELA	320.5	325.1	30.26	-4.56*	-.151	.000	-8.56*	-.282
Math	316.8	324.5	36.14	-7.70*	-.213	.000	-14.44*	-.399
Science	750.0	756.4	39.09	-6.35*	-.163	.000	-12.06*	-.308
	N = 1615-1630	*N* = 976-996		*N* = 2591-2626				

$*p < .05$, †p < .10 for coefficients

[a] Covariate-adjusted means generated by multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients are between 1.8965-1.8990 with third grade and 1.8751-1.8972 for sixth grade.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation. Negative signs indicate a less favorable outcome for the treatment group.

[e] The *p*-values for statistical significance that are the same for the ITT and TOT coefficients.

Table 3

Effect Sizes[a] for the RCT and ISS Samples for Sixth Grade Outcomes

	RCT (N = 2591-2700)		ISS (N = 914-965)	
	ITT	TOT	ITT	TOT
Achievement Tests				
English	-.128	-.240	-.091	-.185
Math	-.178	-.333	-.113	-.227
Science	-.132	-.249	-.075	-.156
On Grade	-.025	-.047	.063	.125
IEP	-.107	-.203	-.135	-.270
School Rules	-.119	-.222	-.158	-.316
Major Offenses	-.083	-.157	-.073	-.146
Any Offenses	-.090	-.170	-.140	-.278

[a] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation. Negative signs indicate a less favorable outcome for the treatment group.

Table 4

Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Grade Level and Special Education Status at the End of Sixth Grade (RCT Analytic Sample)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
				Observed Values				
On grade	.872	.881	.329	-.008	-.025	.531	-.016	-.047
IEP	.117	.084	.304	.033*	-.107	.010	.062*	-.203
				Weighted Observed Values				
On grade	.851	.860	.354	-.009	-.026	.528	-.017	-.049
IEP	.126	.081	.310	.045*	-.144	.001	.085*	-.272
	N = 1678-1679	*N* = 1021		*N* = 2699-2700				

*$p < .05$

Note. On grade is a binary variable: 1=at or above expected grade level, 0 = below expected grade level. IEP = Individualized Educational Program as the formal special education designation coded 1 as yes and 0 as no.

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

Table 4 (continued).

[c] Coefficients for the treatment-control differences from multilevel multiple models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for ITT coefficients that estimate TOT coefficients are 1.8907 for expected grade level and 1.8904 for IEP.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance are the same for the ITT and TOT coefficients.

Table 5

Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Cumulative Disciplinary Events through Sixth Grade (RCT Analytic Sample)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p* - value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
				Observed Values				
School Rules	.231	.185	.396	.047*	-.119	.004	.088*	-.222
Major Offenses	.137	.109	.331	.028*	-.083	.043	.052*	-.157
All Offenses	.273	.234	.429	.039*	-.090	.025	.073*	-.170
				Weighted Observed Values				
School Rules	.249	.194	.409	.055*	-.135	.001	.103*	-.253
Major Offense	.139	.117	.339	.022	-.066	.121	.042	-.123
All Offenses	.287	.250	.440	.037*	-.084	.041	.070*	-.159
	N = 1618-1626	*N* = 974-980		*N* = 2592-2606				

**p* < .05 for coefficients

Note. School rules: violations of school rules or other administrative issues; major offenses: fighting, bullying, weapon in school, and the like; all offenses: total across school rule and major offenses categories. These are coded for whether there is any infraction recorded in school records cumulatively from K through the sixth-grade year (1 = yes, 0 = no).

[a] Covariate-adjusted means generated by the multilevel analysis models.

Table 5 (continued).

[b] Pooled treatment and control group standard deviations. There were minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes were computed on the exact values.

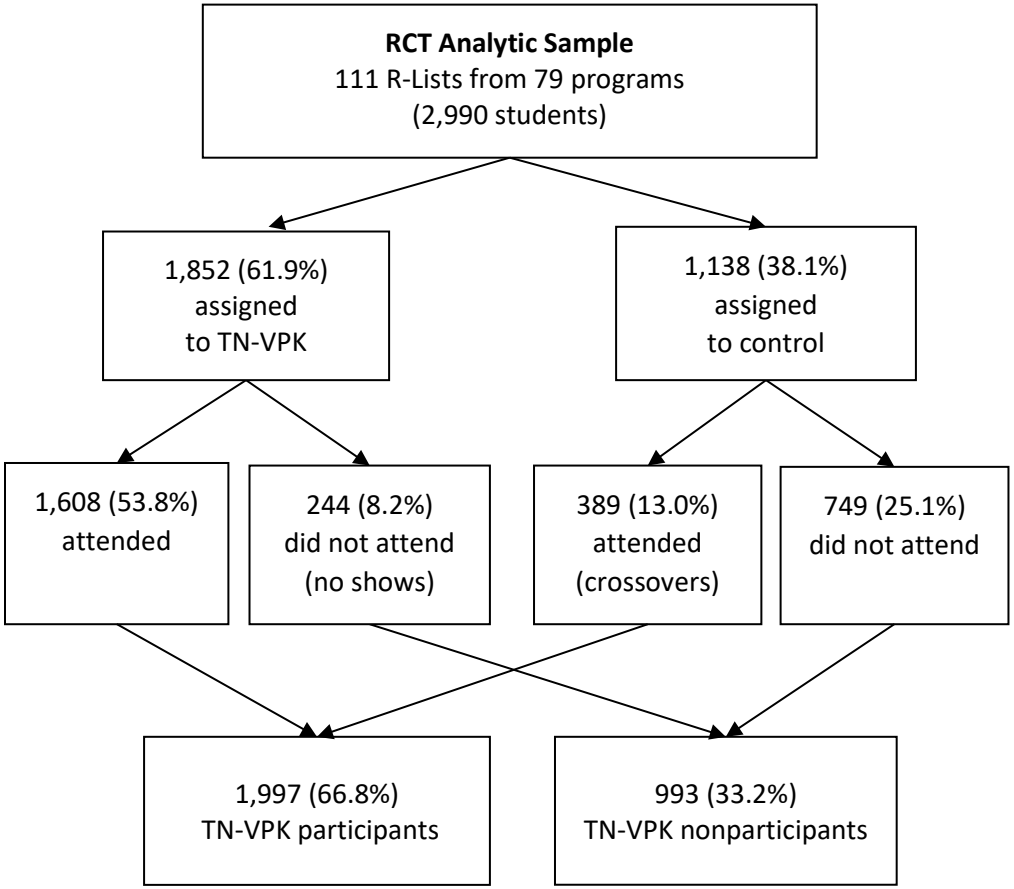
[c] Coefficients for the treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multiplier for ITT coefficients that estimates TOT coefficients is 1.8790 for school rule violations, 1.8811 for major offenses, and 1.8847 for all offenses.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients.

Figure 1

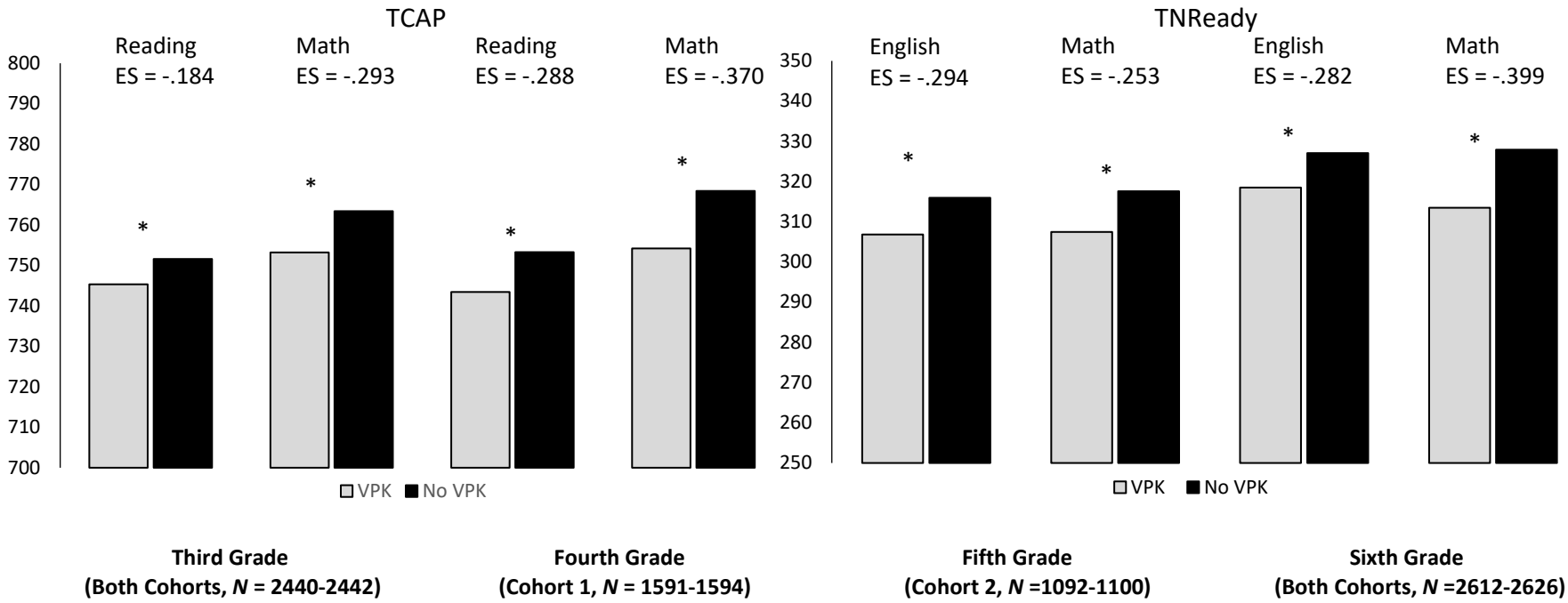
Composition of the Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Groups in the RCT Analytic Sample



Note: Percentages read across each row.

Figure 2

Standardized Tests TOT Weighted Covariate-Adjusted Means in Third through Sixth Grades (RCT Analytic Sample)



Note. Asterisks indicate significant differences (p <.05). These graphs include students who have been promoted or retained. The grades noted above refer to students' expected grade levels. More detailed ITT and TOT results for each grade with observed and weighted data are provided in Supplemental Table S6. Analyses were also performed including only students who were at or above expected grade level with similar results (Supplemental Table S7).

Figure 3

Disciplinary Offenses in Kindergarten through Sixth Grade (RCT Analytic Sample)

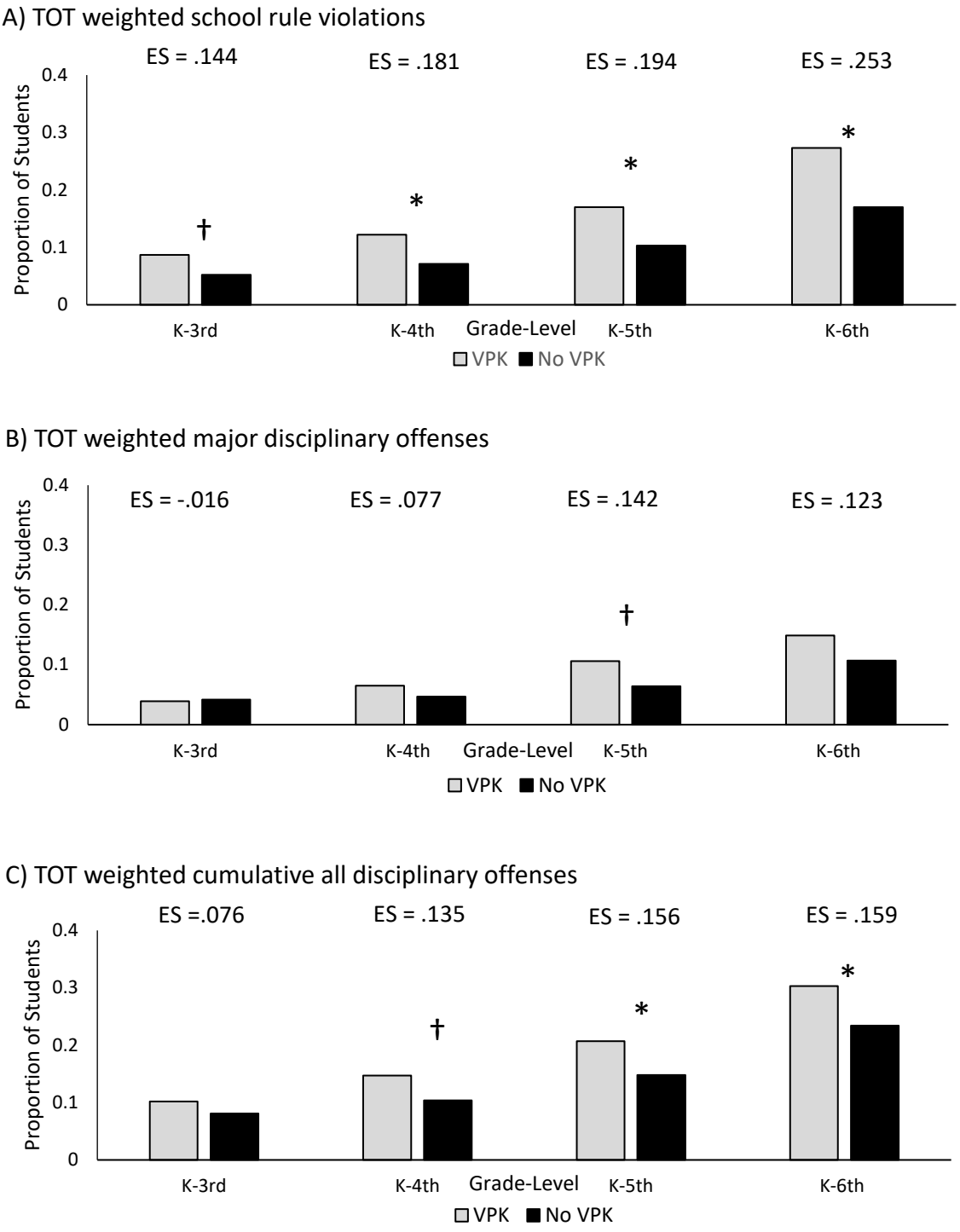


Figure 3 (continued).

Note. Percentage of students with A) one or more school rule violation disciplinary offenses, B) one or more major disciplinary offenses, and C) any type of disciplinary offense across grades. Asterisks indicate $p < .05$ and obelisks indicate $p < .10$. Cumulative disciplinary analyses are located in Supplemental Table S13.

Supplemental Material to Accompany the Manuscript
Effects of a Statewide Prekindergarten Program on Children's Achievement and Behavior through Sixth Grade

Table of Contents

Supplemental Materials 1: The Influence of Attrition on Estimated VPK Effects in 6th Grade

The RCT involves 3131 eligible children randomized via eligible site level R-Lists. Of those, 141 children were not enrolled in TN public schools after the pre-k year through 6th grade (with one exception who emerged in 5th and 6th grade) and thus did not appear in DOE data. These 141 also had very little data for the pre-k year, especially those in the control group, most of whom did not enroll in VPK. These 141 were excluded from the RCT analytic sample, leaving N=2990.

The question of the influence of attrition on the outcome findings is one of whether the missing outcome data are nonrandomly distributed between the treatment and control groups in ways that bias effect estimates based on the cases for which data are available. There are two parts to this question. One involves the N=141 children who did not have any outcome data. The other involves the children in the N=2990 analytic sample who were missing data on any given outcome measure.

What we know about the N=141 cases

<u>The proportions of the 141 in the ITT treatment and control conditions are not significantly different</u>

The 141 are .045 of the 3131 initially randomized children; 79 (.041) of those are in the ITT treatment condition (N=1931); 62 (.052) are in the ITT control condition (N=1200). A test of the difference between these proportions (.011) found SE=.0078, Z=1.411, *p*=.158 using the Wald Z test statistic with variance estimates under the null hypothesis (Wald H_0 in the SPSS 28 options; *p*-values for all other test options were slightly larger). A multilevel logistic regression testing this difference that takes into account the nesting of children within R-Lists and school districts (Mixed Generalized Linear in SPSS) produced a t-value of .925, *p*=.355.

<u>Most of the 141 children did not actually enroll in VPK during the pre-k year</u>

34 of 79 (43.0%) children in the ITT treatment condition enrolled in VPK and attended for at least some instructional days: mean of 202.1 days enrolled; 115.6 instructional days attended. The remaining 45 children (57.0%) did not enroll or attend at all.

6 of 62 (9.7%) children in the ITT control condition enrolled and attended for some instructional days: mean of 207.0 days enrolled; 115.2 instructional days attended. The remaining 56 children (90.3%) did not enroll or attend VPK during the pre-k year.

Thus of the total of 141 children, 101 (71.6%) did not enroll or attend VPK during the pre-k year.

<u>Very few of the 141 were included in the N=1076 ISS subsample</u>

Only 5 of the 79 children in the ITT treatment condition (6.3%) were in the ISS subsample; only 6 of the 62 children in the ITT control condition (9.7%) were in that subsample. Thus 11 of 141 (7.8%) overall; these 11 were not included in the full RCT N=2990 analytic sample because of the lack of post pre-k data. These 11 do not provide sufficient representation of the N=141 for their data to be helpful in assessing differences between those in the ITT treatment and control conditions as an indication of differential attrition.

Attrition in the RCT sample with (N=3131) and without (N=2990) the 141 included

Attrition on Outcome Variables for N=3131 Initial Randomization Sample (Tx=1931, Ctr=1200)

6th Grade Outcome	Overall		Treatment		Control		Tx-Ctr Difference[a]
	N	Missing	N	Missing	N	Missing	
TNReady, ELA scores	2612	.166	1624	.159	988	.177	.018
TNReady, Math scores	2626	.161	1630	.156	996	.170	.014
TNReady, Science scores	2591	.172	1615	.164	976	.187	.023
Attendance	2696	.139	1675	.133	1021	.149	.016
Expected Grade Level	2699	.138	1678	.131	1021	.149	.018
IEP (no Gifted or Physical)	2700	.138	1679	.131	1021	.149	.018
School Rule Violations, K-6	2595	.171	1619	.162	976	.187	.025
Major Offenses, K-6	2592	.172	1618	.162	974	.188	.026
Any Offenses, K-6	2606	.168	1626	.158	980	.183	.025

[a] Absolute value of the difference. Proportion missing for control is larger than for treatment for all outcomes.

Statistical Tests of the ITT Treatment-Control Attrition Differences for the N=3131 Initial Sample

6th Grade Outcome	Tx-Ctr Difference	Diff Between Proportions[a]			ML Logistic Regression[b]	
		SE	z-value	*p*-value	t-value	*p*-value
TNReady, ELA scores	.018	.014	1.294	.196	1.262	.207
TNReady, Math scores	.014	.014	1.045	.296	1.084	.278
TNReady, Science scores	.023	.014	1.658	.097	1.044	.297
Attendance	.016	.013	1.305	.192	1.132	.258
Expected Grade Level	.018	.013	1.431	.152	1.250	.211
IEP (no Gifted or Physical)	.018	.013	1.474	.141	1.284	.199
School Rule Violations, K-6	.025	.014	1.812	.070	1.781	.075
Major Offenses, K-6	.026	.014	1.891	.059	1.900	.058
Any Offenses, K-6	.025	.014	1.848	.065	1.832	.067

[a] Test of the difference in proportions using the Wald Z test statistic with variance estimates under the null hypothesis (Wald H_0 test in the SPSS 28 Compare Means/Independent-Samples Proportions; *p*-values for the other test options were larger).

[b] Multilevel logistic regression testing the difference in proportions that takes into account the nesting within R-Lists and Districts (SPSS Mixed Models/Generalized Linear).

Attrition on Outcome Variables for N=2990 Analytic Sample (Tx=1852, Ctr=1138)

6th Grade Outcome	Overall		Treatment		Control		Tx-Ctr Difference[a]
	N	Missing	N	Missing	N	Missing	
TNReady, ELA scores	2612	.126	1624	.123	988	.132	.009
TNReady, Math scores	2626	.122	1630	.120	996	.125	.005
TNReady, Science scores	2591	.133	1615	.128	976	.142	.014
Attendance	2696	.098	1675	.096	1021	.103	.007
Expected Grade Level	2699	.097	1678	.094	1021	.103	.009
IEP (no Gifted or Physical)	2700	.097	1679	.093	1021	.103	.010
School Rule Violations, K-6	2595	.132	1619	.126	976	.142	.016
Major Offenses, K-6	2592	.133	1618	.126	974	.144	.018
Any Offenses, K-6	2606	.128	1626	.122	980	.139	.017

[a] Absolute value of the difference. Proportion missing for control is slightly larger than for treatment for all outcomes.

Statistical Tests of the ITT Treatment-Control Attrition Differences for the N=2990 Analytic Sample

6th Grade Outcome	Tx-Ctr Difference	Diff Between Proportions[a]			ML Logistic Regression[b]	
		SE	z-value	*p*-value	t-value	*p*-value
TNReady, ELA scores	.009	.013	.695	.487	1.037	.300
TNReady, Math scores	.005	.012	.399	.690	.807	.419
TNReady, Science scores	.014	.013	1.123	.261	.784	.433
Attendance	.007	.011	.646	.519	.895	.371
Expected Grade Level	.009	.011	.794	.427	1.038	.299
IEP (no Gifted or Physical)	.010	.011	.843	.399	1.079	.281
School Rule Violations, K-6	.016	.013	1.297	.195	1.637	.102
Major Offenses, K-6	.018	.013	1.388	.165	1.780	.075
Any Offenses, K-6	.017	.013	1.334	.182	1.679	.093

[a] Test of the difference in proportions using the Wald Z test statistic with variance estimates under the null hypothesis (Wald H_0 test in the SPSS Compare Means/Independent-Samples Proportions; *p*-values for the other test options were larger).

[b] Multilevel logistic regression testing the difference in proportions that takes into account the nesting within RLists and Districts (SPSS Mixed Models/Generalized Linear).

Summary: There are modest differences between the ITT treatment and control conditions in the proportions of missing values on the outcome variables that are somewhat larger for the control group for both the initial and analytic sample. None of those differences are statistically significant at alpha=.05 although some are marginal ($p<.10$) for disciplinary outcomes.

Potential for differences in the characteristics of the children without outcome data in the treatment and control conditions to bias effect estimates

Even though there are only relatively small and nonsignificant differences between the ITT treatment and control conditions in the proportions of missing outcome data, it is possible that the children with missing data in those conditions are different in the outcomes they would have shown if their data were available.

There's no definitive way to know what the missing outcome values would be if we had them, but an informative approach is to impute the missing values with a strategy that predicts based on the data we do have on each of these children. This is especially tenuous for the 141 who have little presence in DOE data during the pre-k year and none thereafter. The only descriptors we have for most of them are program level ones—the R-list they are on (program site) and the descriptive variables available for those program sites. These include Urbanicity (urban vs. rural areas), Partner programs (operated by community organizations vs. schools), Priority schools (operated in the lowest performing schools), Pilot programs (funded in 1996 as pilot pre-k programs), and Region (west, central west, central east, and east parts of the state).

The multiple Imputation routine in SPSS 28 was used to generate 25 imputed data sets for the initial 3131 cases in the initial randomization sample. This was done separately for the ITT treatment and control conditions with the two datasets generated then combined for analysis. The imputation method used by SPSS is a fully conditional iterative Markov Chain Monte Carlo procedure described as follows: "For each iteration and for each variable in the order specified in the variable list, the fully conditional specification (FCS) method fits a univariate (single dependent variable) model using all other available variables in the model as predictors, then imputes missing values for the variable being fit."

The imputed values generated by this procedure were then examined for outliers. For binary categorical variables, the imputed values generally maintained the native 0/1 coding. For the scaled achievement test variables, there was a relatively modest number of outliers at both the lower and upper end. These were recoded to match the smallest and largest scores respectively that were found in the observed data.

The table on the next page shows the treatment effect coefficient estimates for the observed values in the analytic sample (these are the ones reported in the paper) and for the observed values in the initial randomization sample (these are identical because the addition of the 141 cases included in the initial sample, none with any of the outcome data, did not change the observed data, only the number of missing cases excluded from the analysis).

The more informative results are from the analysis of the multiply imputed values (pooled estimates over the 25 imputed datasets). Those for the analytic sample are testing the influence of the relatively few missing values in the outcome data for that sample, i.e., whether what we reported using only observed values could be biased because of the missing data within that analytic sample. These imputations should be relatively solid because of the amount of other data on these children that were used in the prediction of the missing values.

The coefficient estimates for the initial randomization sample then also include the imputed values for the 141 cases with almost no data at all. This tests whether those 141 cases that we chose to omit from the analytic sample show any potential to have biased our effect estimates. In all these analyses, the same multilevel models with the same covariates that generated the effect estimates reported in the paper were used.

ITT Treatment Effect Estimates (Coefficients from Multilevel Models) and Their Statistical Significance from Analyses with the Analytic and Initial Samples with and without Imputation of Missing Outcome Data

6th Grade Outcome	Analytic Sample Observed Values		Initial Sample Observed Values		Analytic Sample Imputed Values		Initial Sample Imputed Values	
	B	*p* -value	B	*p* -value	B	*p* -value	B	*p* -value
TNReady, ELA scores	-3.83*	.002	-3.83*	.002	-3.28*	.025	-4.12*	.020
TNReady, Math scores	-6.46*	<.001	-6.46*	<.001	-6.46*	<.001	-7.21*	.002
TNReady, Science scores	-5.18*	.002	-5.18*	.002	-4.44*	.041	-4.82†	.094
Attendance	-.003*	.013	-.003*	.013	-.003*	.048	-.001	.474
Expected Grade Level	-.008	.531	-.008	.531	-.005	.831	-.002	.921
IEP (no Gifted or Physical)	.033*	.010	.033*	.010	.016	.381	.011	.572
School Rule Violations, K-6	.047*	.004	.047*	.004	.043*	.011	.046*	.014
Major Offenses, K-6	.028*	.043	.028*	.043	.014	.335	. 010	.583
Any Offenses, K-6	.039*	.025	.039*	.025	.035†	.058	.034†	.096

The critical achievement test scores show a high level of consistency across these analyses with similar effect estimates that are all statistically significant. School Rule Violations and Any Offense Violations show a high degree of consistency in the coefficient estimates, though the statistical significance for Any Offenses is marginal for the analyses with imputed values. There is less consistency in both the effect estimates and their statistical significance for Attendance, IEP, and Major Offenses. However, in all cases the direction of effects is the same, i.e., either negative or positive across the board.

Supplemental Materials 2: Analysis Model Details

Analyses of treatment control differences were conducted with hierarchical linear models (HLM) with eligible child TN-VPK applicants nested in the program sites that participated in the randomization (R-Lists) and those R-List program sites nested in the districts where they were located.

The mixed models subroutine of SPSS version 27 was used to implement these analyses. The syntax for main effects analyses took the following form:

```
MIXED DV BY Tx WITH Cov1  Cov2  Cov3 . . .
/CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1) SINGULAR(0.000000000001)
  HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=Tx  Cov1  Cov2  Cov3 . . .  | SSTYPE(3)
/METHOD=REML
/PRINT= SOLUTION
/RANDOM=INTERCEPT | SUBJECT(District.ID) COVTYPE(VC)
/RANDOM=INTERCEPT | SUBJECT(District.ID*RList.ID) COVTYPE(VC).
/EMMEANS=TABLES(Tx).
```

With DV=dependent variable; Tx=treatment condition; and Cov1, Cov2, Cov3 etc.= to the covariates included in the model. The EMMEANS command generates the estimated marginal means for each group defined by Tx. When interactions with treatment condition were examined, the FIXED command represented the terms needed for the interaction test as follows:

```
/FIXED=Tx Cov Tx*Cov | SSTYPE(3)
```

The formal model represented in this syntax is as follows:

(1) Level 1, fixed effects for children

$$DV_{ijk} = \alpha_{0jk} + \beta_0 Tx_{ijk} + \beta X_{ijk} + e_{ijk} \quad i=1 \text{ to } I,\ j=1 \text{ to } J,\ k=1 \text{ to } K$$

Where DV_{ijk} is the dependent variable score for child i in the sample of I children, with each nested in a j R-List and a k district; α_{0jk} is the intercept within the J R-Lists and K districts; β_0 is the coefficient for the treatment variable Tx_{ijk}; β is the coefficient for a representative covariate X; and e_{ijk} is the error term at Level 1.

(2) Level 2, random effects for R-Lists

$$\alpha_{0jk} = \gamma_{00k} + e_{0jk} \quad j=1 \text{ to } J,\ k=1 \text{ to } K$$

Where γ_{00k} is the R-List intercept in each k District; and e_{0jk} is the error term at Level 2.

(3) Level 3, random effects for Districts

$$\gamma_{00k} = \lambda_{000} + e_{00k} \quad k=1 \text{ to } K;$$

Where λ_{000} is the District intercept and e_{00k} is the error term at Level 3.

Supplemental Materials 3: Derivation of the CACE and TOT Effect Estimates

In the analytic sample of N=2990, 86.8% of the children offered VPK admission actually participated and 34.2% of the children not offered admission managed to enroll in VPK anyway.

	Participation		
Randomization	Enrolled in VPK	Did not enroll	
Assigned to Tx	1608 (.868)	244 (.132) [no shows]	1852
Assigned to Ctr	389(.342) [crossovers]	749 (.658)	1138
	1997	993	2990

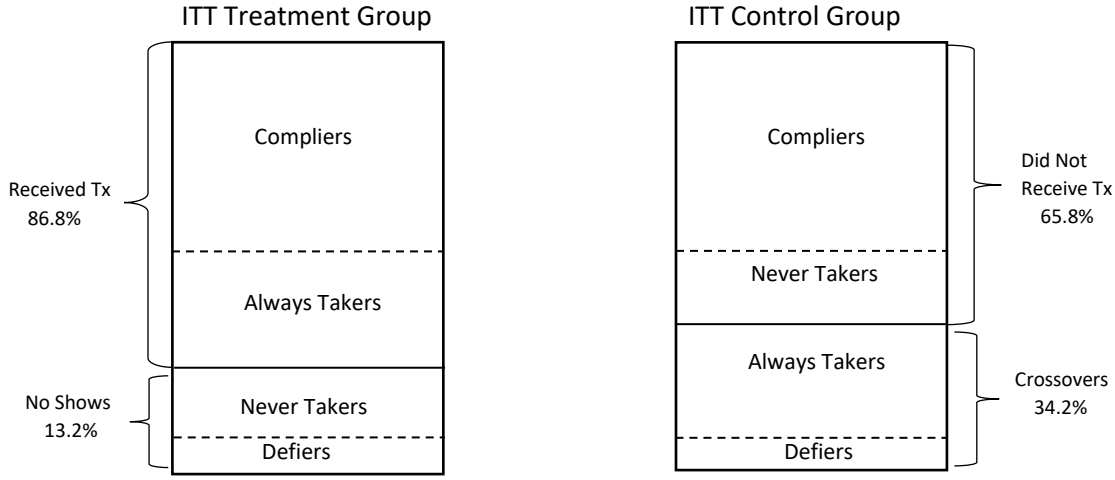
The ITT effect estimates compared outcomes for children assigned to the VPK treatment condition with those assigned to the control condition irrespective of actual participation. In addition, we want TOT estimates of the effect of VPK on the children who actually participated.

We modelled our derivation of the TOT estimates on the principal stratification procedure used in the Head Start impact study (Puma, Bell, Cook, & Heid, 2010) and discussed more generally by others (e.g., Gennetian, Morris, Bos, & Bloom, 2005).

In this procedure the ITT treatment group is recognized as consisting of four subgroups defined in terms of how they react to the randomization:

- *Compliers* who accept treatment when assigned to the treatment condition and do not participate in treatment when assigned to the control condition.
- *Always Takers* who participate in the treatment whether assigned to the treatment or control condition. Those assigned to the control who nonetheless obtain treatment are *Crossovers*.
- *Never Takers* who do not participate in the treatment irrespective of which group they are assigned to. Those assigned to treatment who do not then participate are referred to as *No Shows*.
- *Defiers* who respond in opposition to the assignment, failing to participate if assigned to treatment and managing to participate anyway if assigned to control.

While these subgroups are assumed to exist in the ITT treatment group, the individuals in each subgroup cannot necessarily be identified. However, because of randomization to ITT conditions, the ITT control group is assumed to include equivalent subgroups in the same proportions as in the ITT treatment group. This situation can be depicted as follows for the N=2990 analytic sample.



Notation

M=mean for a group, subscript t if in ITT treatment, c if in ITT control; the overall ITT effect estimate is thus ITT = $M_t - M_c$.

P_1 is the proportion of the ITT treatment group that participates in treatment (P_1=.868) and the mean for that group is M_{t1}. P_0 is the proportion of the ITT treatment group that does not participate in treatment (No Shows; P_o=.132) and the mean for that group is M_{t0}.

A second subscript identifies subgroups

c for Compliers; M_{tc} for the ITT treatment subgroup mean, M_{cc} for the ITT control subgroup mean for the equivalent individuals, P_c for the subgroup proportion in the full ITT treatment group.

a for Always Takers; M_{ta} for the ITT treatment subgroup mean, M_{ca} for the ITT control subgroup mean for the equivalent individuals, P_a for the subgroup proportion in the full ITT treatment group.

n for Never Takers; M_{tn} for the ITT treatment subgroup mean, M_{cn} for the ITT control subgroup mean for the equivalent individuals, P_n for the subgroup proportion in the full ITT treatment group.

d for Defiers; M_{td} for the ITT treatment subgroup mean, M_{cd} for the ITT control subgroup mean for the equivalent individuals, P_d for the subgroup proportion in the full ITT treatment group.

Using this notation, the ITT treatment effect can be represented as a proportional combination of the effects for those receiving treatment and the No Shows:

(1) $ITT = M_t - M_c = P_1 (M_{t1} - M_{c1}) + P_0 (M_{t0} - M_{c0})$ $\quad P_1 + P_o = 1$

The effect for those participating in the treatment can be divided into effects for Compliers and Always Takers

$M_{t1} - M_{c1} = P_c/P_1 (M_{tc} - M_{cc}) + P_a/P_1 (M_{ta} - M_{ca})$ $\quad P_c/P_1 + P_a/P_1 = 1$ *

The effect for those not receiving treatment (No Shows) can be further divided into effects for Never Takers and Defiers

$M_{t0} - M_{c0} = P_n/P_0 (M_{tn} - M_{cn}) + P_d/P_0 (M_{td} - M_{cd})$ $\quad P_n/P_0 + P_d/P_o = 1$

Substituting into Equation (1) yields

(2) $ITT = M_t - M_c = P_1 [P_c/P_1 (M_{tc} - M_{cc}) + P_a/P_1 (M_{ta} - M_{ca})] + P_0 [P_n/P_0 (M_{tn} - M_{cn}) + P_d/P_0 (M_{td} - M_{cd})]$

Some key assumptions:

- There are no Deniers or, at most, a trivial number. It's not plausible that there are parents who would apply for VPK then respond to the randomization by refusing admission if assigned to an offer of enrollment but make an effort to obtain admission if randomized to the control. Thus P_d=0 and the term $P_d/P_0 (M_{td} - M_{cd})$ drops out of Equation 2.
- Neither the Never Takers in the ITT treatment group or the equivalent individuals in the ITT control group participate in the treatment, so they experience no treatment effect. Therefore $M_{tn} - M_{cn} = 0$ and the term $P_n/P_0 (M_{tn} - M_{cn})$ drops out of Equation 2.
- The Crossovers from the ITT control who participate in VPK have the same mean outcome as the equivalent Always Takers in the ITT treatment who participate in VPK. Note that crossovers from the ITT control come from the same program-level RLists as the comparable children in those RLists embedded in the ITT treatment and thus have essentially the same VPK program options. Thus $M_{ta} = M_{ca}$, $M_{ta} - M_{ca} = 0$, and the term $P_a/P_1 (M_{ta} - M_{ca})$ drops from Equation 2.

Equation (2), therefore, reduces to

(3) $ITT = M_t - M_c = P_1 [P_c/P_1 (M_{tc} - M_{cc})]$

$P_c/P_1 = 1 - P_a/P_1$ [see * above] so Equation (3) can be written as

$ITT = M_t - M_c = P_1(1- P_a/P_1) (M_{tc} - M_{cc})] = P_1((P_1 - P_a)/P_1) (M_{tc} - M_{cc})] = (P_1 - P_a) (M_{tc} - M_{cc})$

Rearranging terms yields

(4) $M_{tc} - M_{cc} = (M_t - M_c)/ (P_1 - P_a) = ITT / (P_1 - P_a) = ITT (1/(P_1 - P_a))$

$M_{tc} - M_{cc}$ in Equation (4) is the effect estimate for Compliers, known as the Complier Average Causal Effect (CACE) (or the Local Average Treatment Effect, LATE). This is the effect for those who react to the randomization by complying with their respective assignment to the treatment or control condition.

In this formulation, P_1 is the proportion of the ITT treatment group that participated in VPK (.868); P_a is the proportion of the ITT treatment group equivalent to the Crossovers in the ITT control group that also participated in VPK (.342). For the N=2990 analytic sample, therefore, $P_1 - P_a$ = .868 - .342 = .526 and 1/.526 = 1.901. (Note: These proportions will vary in analyses of outcomes with attrition that changes the proportions of P_1 or P_a).

The complier effect estimate (CACE) therefore can be estimated by rescaling the ITT effect estimate, in this case multiplying it by 1.901. It applies to the ITT effect estimate when it is adjusted by baseline covariates as well as when it is not; improvements in the ITT estimate also improve the complier effect estimate as well. Moreover, the standard errors are scaled by the same factor so the statistical significance for the ITT estimate and the complier effect estimate is the same.

Another method for estimating CACE is via a two-stage least squares instrumental variables analysis with randomization as the instrumental variable (Angrist, Imbens, and Rubin, 1996). The stratification procedure described here has been shown to yield the same estimates as this instrumental variable method (Gennetian et al., 2005; Puma, et al., 2010).

The Complier Effect Estimate as a TOT Effect Estimate

The CACE Complier effect estimate compares outcomes for a group of participants to the outcomes for an equivalent counterfactual group of nonparticipants and thus focuses on the effects of the treatment on some of those who actually participated in the treatment. However, it is limited to Compliers, those who react to the randomization according to the randomized assignment to conditions. It does not include all treatment participants, in particular, the Crossovers in the ITT control group who received treatment or their Always-Taker counterparts in the ITT treatment group.

A full TOT effect estimate would include these additional subgroups in proportion to their respective numbers. The Crossovers in the ITT control group can be readily identified. In the analytic sample 389 (34.2%) of that group are Crossovers (control Always Takers). The expectation from randomization is that there will be the same proportion of Always Takers in the ITT treatment group, i.e., .342 x 1852 = 633. Of the 1608 in the ITT treatment group who participated in VPK, that leaves 1608 – 633 = 975 ITT treatment group Compliers. The total, 389+633+975=1997, thus includes all those in the analytic sample who participated in VPK and should be represented in TOT effect estimates in proportion to their respective subgroup sizes as follows.

TOT effect =

[(389/1997) x Crossover effect]+[(633/1997) x Tx Always-Taker effect]+[(975/1997) x Complier effect]

= (.195 x Crossover effect) + (.317 x Tx Always-Taker effect) + (.488 x Complier effect)

The expectation from the randomization is that the ITT treatment Always-Taker effect will be the same as the ITT control Crossover effect. The TOT effect thus reduces to:

TOT effect = (.512 x AlwaysTaker-Crossover effect) + (.488 x Complier effect)

The Complier effect can be estimated using the procedure described above. If the Always-Taker and the equal Crossover effects are the same as the Complier effect or very close, then the Complier effect is itself a good estimate of the TOT effect. The Always-Taker and Crossover effects cannot be directly

estimated from the data available, but some exploration of their potential to be notably larger or smaller than the Complier effect is possible.

For this, we use the achievement test scores that are especially important outcome variables. A first step is to compare the outcomes on these variables for the children in the ITT treatment group who participated in VPK and the Crossovers in the ITT control group who also participated. This comparison was made using multilevel models to take account of any design effects associated with the nesting of students in RLists and school districts. The only predictor variable was ITT treatment condition applied to a sample that included only treatment participants. Those multilevel models (SPSS Mixed Models) generate estimated marginal means that take account of any influence from the nesting.

While the Always Takers in the ITT treatment group cannot be individually identified (mixed in with the Compliers), the expectation from randomization is that they would be there in the same proportion and with the same characteristics as their identifiable counterparts in the ITT control group (Crossovers). With the means for Crossovers and ITT treatment participants (marginal means estimated in the multilevel analyses), and the expected proportions, it is possible to decompose the ITT Tx participant group into the Always taker and Complier Subgroups.

Applying these procedures, we find the following for the comparison of outcomes for ITT treatment participants and ITT control crossovers (with the mean values being the marginal means reported by the multilevel models). Note that sample sizes vary as a result of the attrition on these outcome variables.

Achievement Test 6th Grade Marginal Means

Variable	Tx Participants Mean (N)	Crossovers Mean (N)	t-value	p-value
TNReady English	321.0 (1420)	327.4 (337)	3.44	<.001
TNReady math	318.6 (1422)	324.8 (338)	2.77	.006
TNReady science	751.2 (1409)	759.6 (337)	3.46	<.001

Inferring ITT Tx Always Takers vs Compliers Given That Always Takers Should Match Crossovers

		Tx Participant Subgroups	
Variable	Crossovers Mean (N)	Tx Always Takers Mean (N)	Tx Compliers Mean (N)
TNReady English	327.4 (337)	327.4 (484)	317.7 (936)
TNReady math	324.8 (338)	324.8 (483)	315.5 (939)
TNReady science	759.6 (337)	759.6 (487)	746.9 (922)

As the table above reveals, the mean outcomes for the Crossovers and the matched Tx Always Takers are consistently larger than those for the ITT treatment Complier groups. This shows that the children who crossover from the ITT control group tend to be higher performing than the average VPK participants in the ITT treatment group. However, this does not necessarily mean that the Crossover and Tx Always-Taker groups experience larger VPK effects. The children in those groups may also be higher performing before the VPK experience and not gain more from the VPK experience than other children.

It is not possible to identify children who did not participate in VPK who are fully equivalent to the Crossover/Always-Taker participants to serve as a credible control for estimating VPK effects for those subgroups. However, some idea of the possible magnitude of those effects can be obtained by comparing their outcomes with the various nonparticipating subgroups that can be identified.

One such comparison was made *within* the ITT control group. We have assumed there are no Defiers, so that group is composed of Crossovers and ITT control nonparticipants. This comparison, analyzed with

the set of covariates used in the main ITT analysis, provides a Crossover effect estimate, but one almost certainly biased by unobserved differences between those who crossover and those who remain behind in the ITT control group.

Another comparison can be made *between* subgroups of the ITT treatment and control groups. The outcomes for the Crossovers in the ITT control group can be compared with the No Shows in the ITT treatment group, again with the full set of covariates. And again, despite the covariates, the result is likely to be a biased estimate with No Shows expected to perform more poorly than Crossovers.

A third comparison was made between the outcomes for the Crossover subgroup and the outcomes for the Complier control condition. This comparison supposes that if the Crossover subgroup had not participated in VPK, its outcomes might be the same as those for the ITT control Compliers.

6th Grade Effect Estimates

		Complier Effect		Crossover Comparisons with Nonrandomized Nonparticipant Subgroups					
	Pooled ITT SD	CACE Estimate[a]	Effect Size	(1) w/in ITT Ctr Estimate[b]	(1) Effect Size	(2) Btwn ITT T & C Estimate[c]	(2) Effect Size	(3) Crossover vs Complier Control[d]	(3) Effect Size
English	29.86	-7.18	-.240	-4.35	-0.146	-7.13	-0.239	2.53	0.085
Math	36.31	-12.12	-.333	-1.23	-0.034	-10.50	-0.289	-2.82	-0.078
Science	39.37	-9.83	-.249	-4.70	-0.119	-7.79	-0.198	2.86	0.073

[a] From principal stratification estimates.

[b] ITT control participants (Crossovers) compared with ITT control nonparticipants.

[c] Crossovers compared with No Shows in the ITT treatment group.

[d] Crossover outcomes compared with the inferred outcomes for the Complier control group.

Details for Crossover outcome means compared to complier control means, (3) above

	(a) Inferred Complier Treatment Outcome	(b) Complier Effect Estimate	Implied Complier Control Mean (a)-(b)	Crossover Outcome	Crossover minus Complier Control
English	317.7	-7.18	324.9	327.4	2.53
Math	315.5	-12.12	327.6	324.8	-2.82
Science	746.9	-9.83	756.7	759.6	2.86

Combining Complier, ITT Tx Always Takers, and Crossovers into a combined estimate (TOT?)

TOT = (.488 x complier effect) + (.317 x Tx always-taker effect) + (.195 x crossover effect]

	.488	.512				
	Complier effect estimate	(1) Crossover vs Complier Control	(2) w/in ITT Ctr	(3) Btwn ITT T & C	TOT Combined Estimate (1)	TOT Combined Estimate (2)	TOT Combined Estimate (3)
English	-7.18	2.53	-4.35	-7.13	-2.21	-5.73	-7.15
Math	-12.12	-2.82	-1.23	-10.50	-7.36	-6.54	-11.29
Science	-9.83	2.86	-4.70	-7.79	-3.33	-7.20	-8.79

All these combined effect estimates are negative but less negative than the CACE. There is no obvious basis for selecting any one as a good TOT estimate. But though they vary widely, it is within a fairly restricted range. The key question is whether it is plausible that the differences with the Complier effects are small enough to consider the Complier effect estimate the equivalent of a full TOT estimate.

It is relevant in this regard that the Complier effects are not estimated very precisely. The table below

shows the confidence intervals for those estimates. All the estimates in the table above fall within the confidence intervals for the CACE estimates. Indeed, those confidence intervals are so broad that it is unlikely that the Crossover and ITT treatment Always Taker effects would fall appreciably far outside of them if we were, in fact, able to get good estimates of those effects. To fall outside of those confidence intervals, the Crossover/AlwaysTaker effects for English would have to be ±63% larger or smaller than the CACE estimate; for math, ±45% larger or smaller; and for science, ±62% larger or smaller.

Confidence Intervals for CACE Effect Estimates

	CACE effect	Multiplier	ITT SE	SE x multiplier	Multiplied SE x 1.96	CACE lower CI	CACE upper CI
English	-7.18	1.875	1.233	2.311	4.530	-11.710	-2.650
Math	-12.12	1.876	1.498	2.810	5.508	-17.628	-6.612
Science	-9.83	1.898	1.643	3.117	6.109	-15.939	-3.721

Conclusion

The principal stratification approach used in our analyses is expected to provide valid estimates of the VPK effect on Compliers (CACE) that are equivalent to those that would be obtained using the alternative instrumental variables analysis with randomization as the instrument. As Complier only estimates, however, the CACE estimates omit VPK effects on ITT treatment group Always Takers and ITT control group Crossovers that would be included in a full TOT effect estimate. Both those subgroups are assumed to experience the same effects and exist in the same proportions in their respective ITT conditions. If their common effects are the same or very similar to those of the CACE estimates, the CACE estimates can also be viewed as TOT estimates. For the central achievement test outcomes, explorations of the possible order of magnitude of the Crossover/Always-Taker effects found notable variation but within a moderately restricted range. In particular, all those estimates fell within the confidence intervals for the CACE estimates, which are quite broad. The Crossover/Always-Taker effects would have to be considerably larger or smaller than the Complier effects for them to fall outside those confidence intervals. On that basis, we take the CACE estimates to be acceptable TOT estimates as well.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-455.

Gennetian, L.A., Morris, P.A., Bos, J.M., & Bloom, H.S. (2005). Constructing instrumental variables from experimental data to explore how treatments produce effects. In H.S. Bloom (Ed.), Learning more from social experiments: Evolving analytic approaches. New York: Russell Sage Foundation.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study. Technical report*. Washington DC: US Department of Health and Human Services, Administration for Children and Families.

Table S1: Comparison of Multilevel Logistic Regression and HLM Coefficients and *p*-Values for Binary Outcomes (RCT Analytic Sample, Observed Data)

	Logistic Regression		HLM	
Binary Outcome	Coefficient[a]	*p*-value	Coefficient[b]	*p*-value
Grade Level in 6th Grade	-.084	.535	-.008	.531
IEP (no gifted or physical) in 6th Grade	.378	.010	.033	.010
School Rule Violations in K through 6th Grade	.342	.005	.047	.004
Major Offense in K through 6th Grade	.290	.040	.028	.043
Any Offenses in K through 6th Grade	.245	.027	.039	.025

Notes: Multilevel models with students nested in R-Lists and R-Lists nested in districts.

[a] Log odds ratio.

[b] Estimated difference between treatment and control means.

Table S2: Intent-to-Treat (ITT) Treatment-Control Comparisons on Baseline Variables (RCT Analytic Sample, Weighted Observed Data)

Variable	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
Age (months)	52.9	52.9	3.52	-.067	-.019	.620
Gender (male)	.50	.48	.50	.015	.030	.435
White	.73	.73	.49	.003	.006	.859
Black	.21	.20	.46	.010	.023	.519
Hispanic	.07	.08	.29	-.017	-.059	.133
Non-native English	.05	.05	.25	-.001	-.004	.928
	N = 1852	*N* = 1138		N=2990		

* $p < .05$ for coefficients.

[a] Estimated marginal means from the multilevel analysis model.

[b] Pooled treatment and control group standard deviations.

[c] Coefficients for the ITT treatment-control differences from multilevel models predicting each baseline variable with children nested in R-Lists, R-Lists nested in districts, ITT as the only predictor.

[d] Effect size: Coefficient for the treatment-control difference divided by the pooled standard deviation.

Table S3: Multilevel Logistic Regressions Coefficients for Binary Baseline Covariates (RCT Analytic Sample)

Binary Covariate	Coefficient[a]	Odds Ratio	Std. Error	*t*	*p*-value
Male	.017	1.017	.0823	.201	.840
White	-.024	.976	.0974	-.251	.802
Black	-.002	.998	.1012	-.020	.984
Hispanic	.040	1.041	.1064	.376	.707
Non-native English	.040	1.041	.1058	.377	.706

Notes: Multilevel logistic regression with students nested in R-Lists and R-Lists nested in district.

[a] Log odds ratio.

Table S4: Intent-to-Treat (ITT) Treatment-Control Comparisons on Baseline Variables for Observed and Weighted Data with Attrition (RCT Analytic Sample)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
TCAP Reading in Third Grade (Observed Values)						
Age (months)	53.45	53.59	3.42	-.142	-.042	.319
Gender (male)	.48	.49	.50	-.010	-.020	.637
White	.66	.66	.50	-.007	-.014	.713
Black	.21	.21	.46	.002	.005	.893
Hispanic	.15	.14	.42	.006	.015	.705
Non-native English	.14	.14	.42	.003	.008	.845
TCAP Reading in Third Grade (Weighted Observed Values)						
Age (months)	53.11	53.29	3.45	-.173	-.050	.249
Gender (male)	.48	.48	.49	-.005	-.009	.832
White	.72	.72	.49	.000	.001	.982
Black	.22	.22	.46	.004	.009	.823
Hispanic	.07	.08	.29	-.006	-.020	.652
Non-native English	.05	.05	.26	-.003	-.011	.809
	N = 1505	*N* = 935		*N* = 2440		
TCAP Math in Third Grade (Observed Values)						
Age (months)	53.45	53.59	3.42	-.149	-.043	.297
Gender (male)	.48	.49	.50	-.010	-.020	.630
White	.66	.66	.50	-.006	-.012	.739
Black	.21	.21	.46	.003	.006	.872
Hispanic	.15	.14	.42	.005	.013	.758
Non-native English	.14	.14	.42	.002	.005	.893
TCAP Math in Third Grade (Weighted Observed Values)						
Age (months)	53.11	53.29	3.45	-.177	-.051	.237
Gender (male)	.48	.49	.49	-.004	-.008	.848
White	.72	.71	.49	.001	.002	.956
Black	.22	.22	.46	.004	.009	.827
Hispanic	.07	.08	.29	-.006	-.022	.624
Non-native English	.05	.05	.26	-.003	-.013	.777
	N = 1506	*N* = 936		*N* = 2442		

Table S4 (continued)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
TCAP Science in Third Grade (Observed Values)						
Age (months)	53.45	53.59	3.42	-.145	-.042	.308
Gender (male)	.48	.49	.50	-.011	-.022	.590
White	.66	.66	.50	-.006	-.012	.737
Black	.21	.21	.46	.003	.006	.885
Hispanic	.15	.14	.42	.006	.013	.744
Non-native English	.14	.14	.42	.003	.006	.876
TCAP Science in Third Grade (Weighted Observed Values)						
Age (months)	53.11	53.29	3.45	-.175	-.051	.243
Gender (male)	.48	.49	.49	-.005	-.010	.817
White	.72	.71	.49	.001	.002	.956
Black	.22	.22	.46	.004	.008	.834
Hispanic	.07	.08	.29	-.006	-.021	.631
Non-native English	.05	.05	.26	-.003	-.012	.789
	N = 1506	*N* = 935		*N* = 2441		
TNReady ELA in Sixth Grade (Observed Values)						
Age (months)	53.23	53.31	3.47	-.085	-.025	.549
Gender (male)	.49	.50	.50	-.007	-.013	.742
White	.67	.68	.50	-.010	-.019	.592
Black	.21	.20	.46	.006	.013	.737
Hispanic	.14	.14	.41	.159	.383	.698
Non-native English	.14	.13	.42	.006	.015	.707
TNReady ELA in Sixth Grade (Weighted Observed Values)						
Age (months)	52.87	52.94	3.51	-.067	-.019	.648
Gender (male)	.49	.49	.50	.001	.002	.958
White	.73	.72	.49	.007	.014	.705
Black	.22	.21	.46	.010	.021	.576
Hispanic	.07	.09	.29	-.017	-.058	.170
Non-native English	.05	.05	.25	-.002	-.008	.845
	N = 1624	*N* = 988		*N = 2612*		
TNReady Math in Sixth Grade (Observed Values)						
Age (months)	53.24	53.29	3.47	-.045	-.013	.749
Gender (male)	.49	.49	.50	-.003	-.007	.866
White	.67	.68	.50	-.009	-.018	.616
Black	.20	.20	.46	.003	.006	.886
Hispanic	.14	.14	.41	.009	.022	.589
Non-native English	.14	.13	.42	.009	.022	.574

Table S4 (continued)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
TNReady Math in Sixth Grade (Weighted Observed Values)						
Age (months)	52.89	52.90	3.50	-.013	-.004	.928
Gender (male)	.50	.49	.50	.006	.012	.779
White	.73	.72	.49	.007	.015	.695
Black	.22	.21	.46	.007	.016	.677
Hispanic	.07	.08	.29	-.015	-.052	.219
Non-native English	.05	.05	.25	-.001	-.003	.942
	N = 1630	*N* = 996		*N* = 2626		
TNReady Science in Sixth Grade (Observed Values)						
Age (months)	53.25	53.28	3.47	-.032	-.009	.823
Gender (male)	.50	.49	.50	.006	.012	.779
White	.73	.72	.50	.007	.014	.695
Black	.22	.21	.45	.007	.016	.677
Hispanic	.07	.08	.41	-.015	-.037	.219
Non-native English	.05	.05	.41	-.001	-.002	.942
TNReady Science in Sixth Grade (Weighted Observed Values)						
Age (months)	52.88	52.90	3.51	-.016	-.004	.916
Gender (male)	.50	.49	.50	.007	.013	.756
White	.73	.72	.49	.003	.007	.858
Black	.22	.20	.46	.011	.023	.539
Hispanic	.07	.08	.29	-.013	-.046	.275
Non-native English	.05	.05	.25	-.002	-.007	.866
	N = 1615	*N* = 976		*N* = 2591		
Attendance in Sixth Grade (Observed Values)						
Age (months)	53.25	53.31	3.49	-.061	-.017	.661
Gender (male)	.49	.50	.50	-.005	-.011	.784
White	.67	.68	.50	-.011	-.021	.558
Black	.20	.20	.46	.002	.003	.928
Hispanic	.14	.13	.41	.011	.026	.509
Non-native English	.14	.13	.41	.011	.027	.500
Attendance in Sixth Grade (Weighted Observed Values)						
Age (months)	52.89	52.94	3.54	-.047	-.013	.743
Gender (male)	.49	.49	.50	.002	.004	.921
White	.73	.72	.49	.005	.011	.774
Black	.22	.21	.46	.008	.016	.664
Hispanic	.07	.08	.29	-.014	-.048	.253
Non-native English	.05	.05	.25	.000	.001	.989
	N = 1675	*N* = 1021		*N* = 2696		

Table S4 (continued)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
Expected Grade Level in Sixth Grade (Observed Values)						
Age (months)	53.26	53.31	3.49	-.053	-.015	.700
Gender (male)	.49	.50	.50	-.005	-.010	.796
White	.67	.68	.50	-.011	-.021	.551
Black	.20	.20	.46	.002	.004	.911
Hispanic	.14	.13	.41	.011	.026	.517
Non-native English	.14	.13	.41	.011	.026	.507
Expected Grade Level in Sixth Grade (Weighted Observed Values)						
Age (months)	52.90	52.94	3.54	-.038	-.011	.793
Gender (male)	.49	.49	.50	.002	.004	.928
White	.73	.72	.49	.005	.010	.786
Black	.22	.21	.46	.008	.017	.650
Hispanic	.07	.08	.29	-.014	-.048	.250
Non-native English	.05	.05	.25	.000	.000	.995
	N = 1678	*N* = 1021		*N* = 2699		
IEP (no gifted or physical) in Sixth Grade (Observed Values)						
Age (months)	53.26	53.31	3.49	-.056	-.016	.686
Gender (male)	.49	.50	.50	-.005	-.010	.808
White	.67	.68	.50	-.011	-.021	.556
Black	.20	.20	.46	.002	.004	.914
Hispanic	.14	.13	.41	.010	.025	.519
Non-native English	.14	.13	.41	.011	.026	.509
IEP (no gifted or physical) in Sixth Grade (Weighted Observed Values)						
Age (months)	52.90	52.94	3.54	-.041	-.012	.774
Gender (male)	.50	.49	.50	.002	.005	.908
White	.73	.72	.49	.005	.010	.781
Black	.22	.21	.46	.008	.017	.654
Hispanic	.07	.08	.29	-.014	-.048	.249
Non-native English	.05	.05	.25	.000	.000	.997
	N = 1679	*N* = 1021		*N* = 2700		
School Rule Violations in Kindergarten through Sixth Grade (Observed Values)						
Age (months)	53.26	53.29	3.49	-.034	-.010	.809
Gender (male)	.49	.50	.50	-.008	-.015	.712
White	.67	.68	.51	-.013	-.025	.491
Black	.21	.20	.46	.005	.011	.786
Hispanic	.14	.13	.41	.010	.023	.559
Non-native English	.14	.13	.42	.006	.014	.730

Table S4 (continued)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
School Rule Violations in Kindergarten through Sixth Grade (Weighted Observed Values)						
Age (months)	52.91	52.91	3.53	-.005	-.002	.970
Gender (male)	.49	.49	.50	.002	.004	.919
White	.73	.73	.49	.000	.000	.996
Black	.22	.21	.46	.011	.023	.551
Hispanic	.07	.08	.29	-.011	-.039	.362
Non-native English	.05	.05	.25	-.002	-.008	.848
	N = 1619	*N* = 976		*N* = 2595		
Major Offenses in Kindergarten through Sixth Grade (Observed Values)						
Age (months)	53.26	53.31	3.49	-.049	-.014	.731
Gender (male)	.49	.50	.50	-.008	-.015	.707
White	.67	.68	.50	-.013	-.026	.482
Black	.21	.20	.46	.007	.016	.686
Hispanic	.14	.14	.41	.007	.018	.662
Non-native English	.14	.13	.42	.007	.016	.685
Major Offenses in Kindergarten through Sixth Grade (Weighted Observed Values)						
Age (months)	52.91	52.91	3.53	-.009	-.003	.952
Gender (male)	.50	.50	.50	.001	.002	.968
White	.72	.72	.49	.002	.004	.916
Black	.22	.21	.46	.013	.028	.466
Hispanic	.07	.08	.29	-.016	-.055	.200
Non-native English	.05	.05	.25	-.002	-.006	.888
	N = 1618	*N* = 974		*N* = 2592		
Any Offenses in Kindergarten through Sixth Grade (Observed Values)						
Age (months)	53.26	53.30	3.46	-.037	-.011	.792
Gender (male)	.49	.50	.50	-.008	-.016	.696
White	.67	.68	.50	-.012	-.025	.495
Black	.21	.20	.46	.006	.013	.728
Hispanic	.14	.13	.42	.008	.019	.632
Non-native English	.14	.13	.42	.005	.013	.748

Table S4 (continued)

	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value
Any Offenses in Kindergarten through Sixth Grade (Weighted Observed Values)						
Age (months)	52.91	52.90	3.53	.000	.000	.998
Gender (male)	.50	.50	.50	.000	.001	.982
White	.72	.72	.49	.002	.003	.927
Black	.22	.21	.46	.013	.028	.460
Hispanic	.07	.08	.29	-.016	-.054	.208
Non-native English	.05	.05	.25	-.002	-.007	.868
	N = 1626	*N* = 980		*N* = 2606		

$p < .05$ for coefficients.

[a] Estimated marginal means from the multilevel analysis model

[b] Pooled treatment and control group standard deviations

[c] Coefficients for the ITT treatment-control differences from a multilevel model with children nested in R-Lists, R-Lists nested in districts, with ITT condition as the only predictor.

[d] Effect size: Coefficient for the treatment-control difference divided by the pooled standard deviation.

Table S5: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Sixth Grade State Achievement Tests (ISS)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p* - value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
Sixth Grade TNReady (Observed Values)								
ELA	322.97	325.67	29.54	-2.70	-.091	.192	-5.47	-.185
Math	319.28	323.47	37.12	-4.19	-.113	.110	-8.43	-.227
Science	755.07	758.00	39.03	-2.93	-.075	.299	-6.08	-.156
	N = 594-607	*N* = 320-335		*N* = 914-942				

$*p < .05$, †p < .10 for coefficients

[a] Covariate-adjusted means generated by multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts. Student level covariates are age, male, White, Black, Hispanic, non-English primary language. Program level covariates are region of the state (west, central east, and east); program operator (school vs. partner community agency); original pilot program or not; program hosted by a high priority school; and urban vs. nonurban location. The multipliers for the ITT coefficients that estimate the TOT coefficients are between 2.0141-2.0743 for sixth grade.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation. Negative signs indicate a less favorable outcome for the treatment group.

[e] The 2SLS analysis model yields p-values for statistical significance that are the same for the ITT and TOT coefficients.

Table S6: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Third through Sixth Grade Achievement Tests not Restricted by Grade Level (RCT Analytic Sample)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
Third Grade TCAP (Observed Values)								
Reading	746.1	748.2	34.33	-2.13	-.062	.146	-4.05	-.118
Mathematics	755.9	760.2	35.57	-4.22*	-.119	.006	-8.02*	-.225
Science	748.6	752.2	35.32	-3.58*	-.101	.016	-6.80*	-.192
Third Grade TCAP (Weighted Observed Values)								
Reading	746.9	750.1	33.60	-3.26*	-.097	.027	-6.19*	-.184
Mathematics	755.6	761.0	34.87	-5.40*	-.155	.000	-10.24*	-.293
Science	750.0	754.1	35.49	-4.03*	-.114	.008	-7.64*	-.215
	N = 1505- 1506	*N* = 935-936		*N* = 2440-2442				
Fourth Grade TCAP Cohort 1 (Observed Values)								
Reading	745.2	749.4	35.50	-4.28*	-.120	.029	-8.89*	-.251
Mathematics	756.9	763.7	39.96	-6.75*	-.169	.002	-14.04*	-.351
Science	748.3	754.3	35.24	-5.97*	-.169	.002	-12.44*	-.353
Fourth Grade TCAP Cohort 1 (Weighted Observed Values)								
Reading	746.0	750.8	34.43	-4.76*	-.138	.015	-9.90*	-.288
Mathematics	757.9	764.7	38.37	-6.83*	-.178	.002	-14.20*	-.370
Science	749.5	756.2	34.97	-6.69*	-.191	.001	-13.95*	-.399
	N = 1081-1083	*N* = 510-511		*N* = 1591-1594				
Fifth Grade TNReady Cohort 2 (Observed Values)								
ELA	309.7	313.6	30.66	-3.87†	-.126	.050	-6.86†	-.224
Mathematics	310.2	315.4	40.30	-5.21*	-.129	.045	-9.21*	-.229
Science	748.0	750.6	38.64	-2.63	-.068	.276	-4.63	-.120
Fifth Grade TNReady Cohort 2 (Weighted Observed Values)								
ELA	308.8	314.0	31.10	-5.15*	-.166	.010	-9.14*	-.294
Mathematics	309.7	315.4	39.88	-5.73*	-.144	.025	-10.12*	-.253
Science	746.8	751.0	37.78	-4.20†	-.111	.073	-7.41†	-.196
	N = 593-599	*N* = 499-502		*N* = 1092-1101				

Table S6 (continued)

Sixth Grade TNReady (Observed Values)								
ELA	321.2	325.0	29.88	-3.83*	-.128	.002	-7.18*	-.240
Mathematics	317.1	323.6	36.33	-6.46*	-.178	.000	-12.12*	-.333
Science	750.4	755.6	39.38	-5.18*	-.132	.002	-9.83*	-.249
Sixth Grade TNReady (Weighted Observed Values)								
ELA	320.5	325.1	30.30	-4.56*	-.151	.000	-8.56*	-.282
Mathematics	316.8	324.5	36.19	-7.70*	-.213	.000	-14.44*	-.399
Science	750.0	756.4	39.12	-6.35*	-.163	.000	-12.06*	-.308
	N = 1615-1630	*N* =976-996		*N* = 2591-2626				

∗p < .05, †p < .10 for coefficients.

[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients is between 1.8965-1.8990 with third grade, 2.0799-2.0842 for fourth grade, 1.7643-1.7740 for fifth grade, and 1.8751-1.8972 for sixth grade.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients.

Table S7: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Third through Sixth Grade State Achievement Tests for Students at Expected Grade Level (RCT Analytic Sample)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value	Coefficient for T-C Difference[c]	Effect Size[d]
Third Grade TCAP (Observed Values)								
Reading	746.1	748.2	34.33	-2.13	-.062	.146	-4.05	-.118
Mathematics	755.9	760.2	35.57	-4.22*	-.119	.006	-8.02*	-.225
Science	748.6	752.2	35.32	-3.58*	-.101	.016	-6.80*	-.192
Third Grade TCAP (Weighted Observed Values)								
Reading	746.9	750.1	33.60	-3.26*	-.097	.027	-6.19*	-.184
Mathematics	755.6	761.0	34.87	-5.40*	-.155	.000	-10.24*	-.293
Science	750.0	754.1	35.49	-4.03*	-.114	.008	-7.64*	-.215
	N = 1505-1506	*N* = 935-936		*N* = 2440-2442				
Fourth Grade TCAP Cohort 1 (Observed Values)								
Reading	747.2	751.6	34.85	-4.41*	-.126	.029	-9.16*	-.263
Mathematics	759.2	765.0	39.48	-5.81*	-.147	.011	-12.05*	-.305
Science	749.6	754.8	34.45	-5.18*	-.150	.009	-10.76*	-.312
Fourth Grade TCAP Cohort 1 (Weighted Observed Values)								
Reading	748.9	753.0	33.02	-4.10*	-.124	.040	-8.52*	-.258
Mathematics	760.8	765.9	37.05	-5.02*	-.136	.024	-10.42*	-.281
Science	751.4	756.2	33.62	-4.88*	-.145	.014	-10.14*	-.302
	N = 947-948	*N* = 460		*N* = 1407-1408				
Fifth Grade TNReady Cohort 2 (Observed Values)								
ELA	310.7	314.0	30.26	-3.31	-.109	.112	-5.87	-.194
Mathematics	312.9	317.2	40.49	-4.29	-.106	.124	-7.57	-.187
Science	750.9	753.1	38.45	-2.18	-.057	.392	-3.84	-.100
Fifth Grade TNReady Cohort 2 (Weighted Observed Values)								
ELA	309.9	314.3	30.31	-4.38*	-.145	.036	-7.78*	-.256
Mathematics	312.9	316.9	39.45	-3.99	-.101	.141	-7.04	-.178
Science	749.7	753.6	37.15	-3.93	-.106	.109	-6.93	-.187
	N = 517-522	*N* = 445-448		*N* = 962-970				
Sixth Grade TNReady (Observed Values)								
ELA	325.4	328.2	27.89	-2.81*	-.101	.022	-5.24*	-.188
Mathematics	321.4	326.2	33.80	-4.84*	-.143	.001	-9.04*	-.267
Science	753.7	757.7	38.67	-4.02*	-.104	.019	-7.59*	-.196

Table S7 (continued)

Sixth Grade TNReady (Weighted Observed Values)								
ELA	325.5	328.7	27.50	-3.25*	-.118	.009	-6.07*	-.221
Mathematics	321.8	327.1	32.88	-5.27*	-.160	.000	-9.84*	-.299
Science	753.8	758.6	37.82	-4.74*	-.125	.006	-8.94*	-.236
	N = 1399-1413	*N* = 871-891		*N* = 2270-2304				

∗p < .05 for coefficients.

Notes. Only students at or above expected grade levels are included.

[a] Covariate-adjusted means generated by the multilevel analysis models with covariates set at the grand means for the sample.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients is 1.8965-1.8990 with third grade, 2.0747-2.0794 for fourth grade, 1.7634-1.7737 for fifth grade, and 1.8657-1.8886 for sixth grade.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients

Table S8: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Grade Level and Special Education Status at the End of Sixth Grade (ISS)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-Value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
				Observed Values				
On grade	.884	.863	.328	.021	.063	.364	.041	.125
IEP	.111	.071	.298	.040†	-.135	.058	.080	-.270
	N = 624-625	*N* = 340		*N* = 964-965				

*$p < .05$, †$p < .10$ for coefficients

Note. On grade is a binary variable: 1=at or above expected grade level, 0 = below expected grade level. IEP = Individualized Educational Program as the formal special education designation.

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for the treatment-control differences from multilevel multiple models with children nested in R-Lists and R-Lists nested in districts. Covariates are the same as in previous models. The multiplier for ITT coefficients that estimates TOT coefficients is 1.9944 for expected grade level and 1.9936 for IEP.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation. Negative signs indicate a less favorable outcome for the treatment group.

[e] The 2SLS analysis model yields p-values for statistical significance that are the same for the ITT and TOT coefficients.

Table S9: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for On Grade Level from Kindergarten through Sixth Grade (RCT Analytic Sample)

	ITT					TOT		
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
On Grade Level (Observed Values)								
Kindergarten	.997	.997	.037	-.001	-.019	.590	-.001	-.035
First grade	.952	.935	.224	.017*	.077	.049	.033*	.146
Second grade	.901	.907	.297	-.006	-.021	.590	-.012	-.040
Third grade	.891	.889	.313	.003	.009	.814	.005	.017
Fourth grade	.884	.882	.322	.002	.006	.882	.004	.011
Fifth grade	.880	.881	.324	.000	-.001	.974	-.001	-.002
Sixth grade	.872	.881	.329	-.008	-.025	.531	-.016	-.047
On Grade Level (Weighted Observed Values)								
Kindergarten	.995	.996	.043	-.001	-.020	.546	-.002	-.038
First grade	.942	.920	.245	.022*	.089	.025	.041*	.169
Second grade	.882	.889	.322	-.007	-.023	.568	-.014	-.043
Third grade	.872	.869	.338	.003	.010	.796	.007	.020
Fourth grade	.864	.862	.347	.002	.007	.862	.005	.013
Fifth grade	.859	.859	.350	.000	-.001	.976	-.001	-.002
Sixth grade	.851	.860	.354	-.009	-.026	.528	-.017	-.049
	N=1678-1852	*N*=1021-1138		*N*=2699-2990				

$*p < .05$, $\dagger p < .10$ for coefficients.

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients range from 1.8907 to 1.9088.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients.

Table S10: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for IEPs from Kindergarten through Sixth Grade (RCT Analytic Sample)

	ITT					TOT		
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
IEPs (Observed Values)								
Kindergarten	.119	.088	.304	.031*	.102	.008	.059*	.194
First grade	.128	.100	.320	.028*	.087	.027	.053*	.165
Second grade	.137	.117	.329	.020	.061	.125	.038	.116
Third grade	.136	.112	.328	.024†	.072	.072	.045†	.137
Fourth grade	.127	.103	.318	.024†	.075	.065	.046†	.143
Fifth grade	.126	.098	.316	.028*	.090	.029	.054*	.171
Sixth grade	.117	.084	.304	.033*	.107	.010	.062*	.203
IEPs (Weighted Observed Values)								
Kindergarten	.131	.093	.321	.038*	.117	.003	.071*	.223
First grade	.141	.109	.338	.032*	.093	.020	.060*	.178
Second grade	.147	.125	.346	.023	.065	.106	.043	.124
Third grade	.143	.119	.340	.024†	.071	.084	.046†	.134
Fourth grade	.132	.106	.330	.026†	.079	.056	.050†	.151
Fifth grade	.134	.100	.326	.034*	.104	.013	.065*	.199
Sixth grade	.126	.081	.310	.045*	.144	.001	.085*	.272
	N=1679-1846	N=1021-1132		*N*=2700-2978				

$*p < .05$, $†p < .10$ for coefficients.

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts with the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients range from 1.8904 to 1.9091.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients.

Table S11: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Attendance from Kindergarten through Sixth Grade (RCT Analytic Sample)

	ITT					TOT		
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
Attendance (Observed Values)								
Kindergarten	.943	.947	.043	-.004*	-.063	.023	-.007*	-.171
First Grade	.952	.954	.039	-.002	-.045	.262	-.003	-.085
Second Grade	.955	.958	.036	-.003†	-.075	.064	-.005†	-.142
Third Grade	.958	.960	.043	-.002	-.051	.215	-.004	-.097
Fourth Grade	.973	.975	.038	-.002	-.050	.230	-.004	-.096
Fifth Grade	.973	.974	.028	-.001	-.035	.406	-.002	-.066
Sixth Grade	.971	.975	.028	-.003*	-.110	.013	-.006*	-.207
Attendance (Weighted Observed Values)								
Kindergarten	.947	.949	.042	-.003	-.065	.100	-.005	-.122
First Grade	.954	.955	.038	-.001	-.024	.547	-.002	-.045
Second Grade	.957	.959	.035	-.002	-.055	.168	-.004	-.104
Third Grade	.960	.962	.041	-.002	-.041	.310	-.003	-.078
Fourth Grade	.975	.977	.036	-.001	-.040	.334	-.003	-.076
Fifth Grade	.975	.975	.027	.000	-.008	.846	.000	-.015
Sixth Grade	.973	.976	.027	-.003*	-.103	.013	-.005*	-.194
	N = 1675-1825	*N* = 1021-1120		*N* = 2696-2945				

$*p < .05$, $†p < .10$ for coefficients

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). Multipliers for the ITT coefficients that estimate the TOT coefficients range from 1.811 to 1.9124. [d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

[e] The p-values that are the same for the ITT and TOT coefficients.

Table S12: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Cumulative Disciplinary Actions through Sixth Grade (ISS)

	ITT						TOT	
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p* - value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
				Observed Values				
School Rules	.225	.163	.392	.062*	-.158	.023	.124	-.316
Major Offenses	.126	.103	.314	.023	-.073	.305	.046	-.146
All Offenses	.253	.195	.416	.058*	-.140	.045	.116	-.278
	N = 604-607	*N* = 329-330		*N* = 933-937				

*$p < .05$, †$p < .10$ for coefficients

Note. School rules: violations of school rules or other administrative issues; major offenses: fighting, bullying, weapon in school, and the like; all offenses: total across school rule and major offenses categories. These are coded for whether there is any infraction recorded in school records cumulatively from K through the sixth grade year (1 = yes, 0 = no).

[a] Covariate-adjusted means generated by the multilevel analysis models.

[b] Pooled treatment and control group standard deviations. There were minor variations between the pooled SDs for the ITT and TOT; the mean is presented here but effect sizes were computed on the exact values.

[c] Coefficients for the treatment-control differences from multilevel models with children nested in R-Lists and R-Lists nested in districts. Covariates are the same as in previous models. The multiplier for ITT coefficients that estimates TOT coefficients is 2.0016for school rule violations, 2.0044 for major offenses, and 1.9976 for all offenses.

Table S13: Intent-to-Treat (ITT) and Treatment-on-Treated (TOT) Effect Estimates for Cumulative Disciplinary Offenses from Kindergarten through Sixth Grade (RCT Analytic Sample)

	ITT					TOT		
	Treatment Group Mean[a]	Control Group Mean[a]	Pooled SD[b]	Coefficient for T-C Difference[c]	Effect Size[d]	*p*-value[e]	Coefficient for T-C Difference[c]	Effect Size[d]
School Rule Violations (Observed Values)								
Kindergarten	.011	.008	.095	.003	.032	.422	.006	.059
K-First	.042	.037	.161	.005	.031	.439	.009	.058
K-Second	.049	.041	.193	.007	.038	.343	.014	.072
K-Third	.069	.053	.231	.016†	.069	.092	.030†	.128
K-Fourth	.098	.074	.270	.024*	.089	.028	.045*	.168
K-Fifth	.141	.108	.328	.033*	.101	.013	.062*	.190
K-Sixth	.231	.185	.396	.047*	.119	.004	.088*	.222
School Rule Violations (Weighted Observed Values)								
Kindergarten	.012	.009	.103	.003	.029	.466	.006	.055
K-First	.044	.039	.173	.005	.029	.468	.009	.053
K-Second	.054	.045	.206	.009	.043	.295	.017	.081
K-Third	.079	.060	.249	.019†	.077	.065	.036†	.144
K-Fourth	.110	.083	.287	.028*	.096	.021	.052*	.181
K-Fifth	.154	.119	.341	.035*	.102	.014	.066*	.194
K-Sixth	.249	.194	.409	.055*	.135	.001	.103*	.253
	N = 1619-1825	*N* = 976-1120		*N* = 2595-2945				
Major Disciplinary Offenses (Observed Values)								
Kindergarten	.006	.005	.076	.002	.025	.532	.004	.046
K-First	.015	.010	.116	.004	.039	.334	.008	.073
K-Second	.024	.019	.153	.005	.034	.397	.010	.064
K-Third	.036	.036	.189	.000	-.001	.983	.000	-.002
K-Fourth	.056	.045	.225	.011	.047	.249	.020	.090
K-Fifth	.091	.067	.276	.024*	.086	.037	.045*	.162
K-Sixth	.137	.109	.331	.028*	.083	.043	.052*	.157
Major Disciplinary Offenses (Weighted Observed Values)								
Kindergarten	.007	.005	.080	.002	.025	.538	.004	.046
K-First	.017	.013	.125	.005	.036	.372	.008	.066
K-Second	.028	.023	.163	.005	.028	.489	.009	.053
K-Third	.040	.041	.200	-.002	-.009	.838	-.003	-.016
K-Fourth	.061	.051	.236	.010	.041	.333	.018	.077
K-Fifth	.096	.074	.286	.021†	.075	.074	.041†	.142
K-Sixth	.139	.117	.339	.022	.066	.121	.042	.123
	N = 1618-1825	*N* = 974-1120		*N* = 2592-2945				

Table S13 (continued)

Any Disciplinary Offenses (Observed Values)								
Kindergarten	.015	.011	.114	.004	.033	.401	.007	.062
K-First	.048	.041	.184	.008	.041	.292	.014	.078
K-Second	.060	.053	.226	.007	.029	.472	.012	.054
K-Third	.087	.078	.269	.009	.035	.385	.018	.066
K-Fourth	.124	.103	.309	.021†	.067	.097	.039†	.126
K-Fifth	.182	.150	.368	.032*	.087	.030	.061*	.165
K-Sixth	.273	.234	.429	.039*	.090	.025	.073*	.170
Any Disciplinary Offenses (Weighted Observed Values)								
Kindergarten	.017	.013	.122	.004	.030	.447	.007	.057
K-First	.052	.043	.197	.009	.044	.271	.016	.081
K-Second	.068	.060	.240	.008	.034	.402	.016	.065
K-Third	.097	.086	.287	.012	.040	.327	.022	.076
K-Fourth	.137	.114	.326	.023†	.072	.083	.044†	.135
K-Fifth	.193	.162	.379	.031*	.083	.044	.059*	.156
K-Sixth	.287	.250	.440	.037*	.084	.041	.070*	.159
	N = 1626-1825	*N* = 980-1120		*N* = 2606-2945				

∗$p < .05$, †$p < .10$ for coefficients.

[a] Covariate-adjusted means generated by the multilevel analysis models.

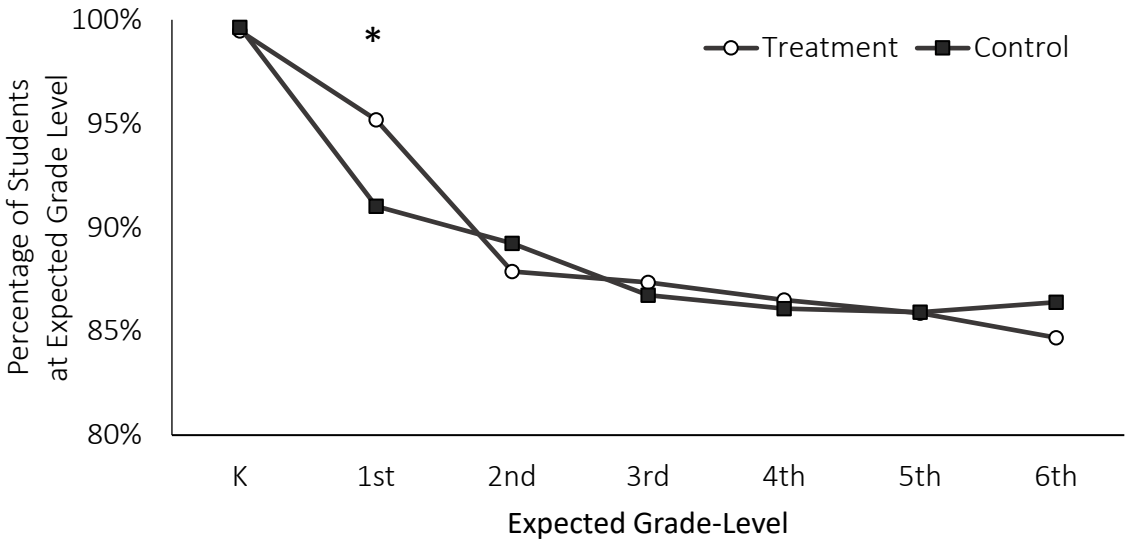
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[c] Coefficients for treatment-control differences from OLS multilevel models with children nested in R-Lists and R-Lists nested in districts and the standard set of covariates (see text). The multipliers for the ITT coefficients that estimate the TOT coefficients range from 1.8772 to 1.8936 for school rule violations, 1.8811 to 1.8907 for major disciplinary offenses, and 1.8765 to 1.8939 for all disciplinary offenses.

[d] Effect size: coefficient for the treatment-control difference divided by the pooled standard deviation.

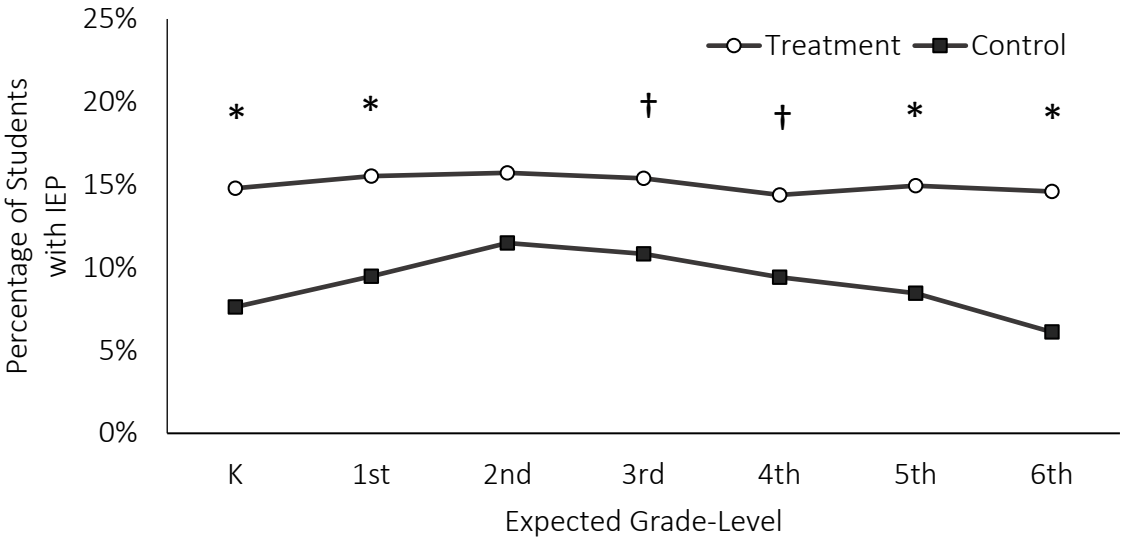
[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients

Figure S1: Grade Level TOT Weighted Means in Sixth Grade (RCT Analytic Sample)



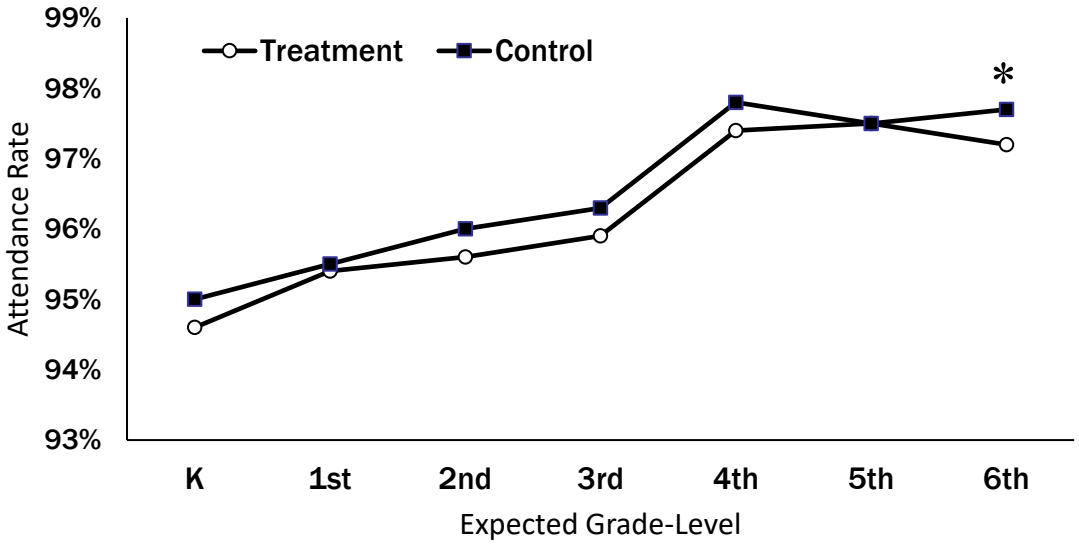
Note. Asterisks indicate $p < .05$ and obelisks indicate $p < .10$. Detailed results for kindergarten through sixth grade are located in Supplemental Table S5.

Figure S2: Special Education Status TOT Weighted Means in Sixth Grade (RCT Analytic Sample)



Note. Asterisks indicate $p < .05$ and obelisks indicate $p < .10$. Detailed results for kindergarten through sixth grade are located in Supplemental Table S6.

Figure S3: Attendance Rates in Kindergarten through Sixth Grade for Weighted TOT Conditions (RCT Analytic Sample)



Note. Asterisks indicate $p < .05$. Results for kindergarten through sixth grade attendance are located in Supplemental Table S7.