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

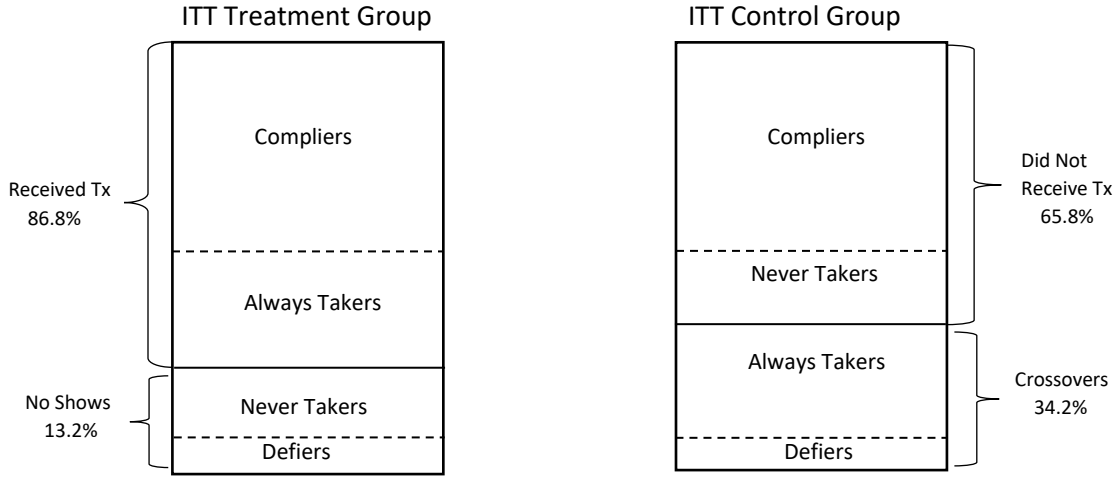
The ITT effect estimates compared outcomes for children assigned to the VPK treatment condition with those assigned to the control condition irrespective of actual participation. In addition, we want TOT estimates of the effect of VPK on the children who actually participated.

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

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

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

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

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

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).

<u>The Complier Effect Estimate as a TOT Effect Estimate</u>

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.

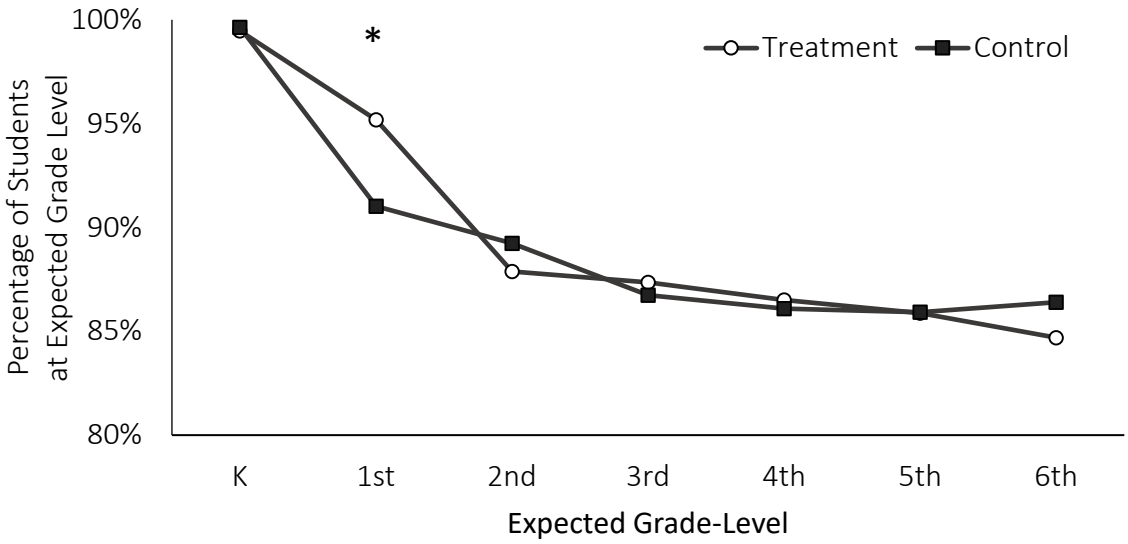
[b] Pooled treatment and control group standard deviations. There are minor variations between the pooled SDs for ITT and TOT; the mean is presented here but effect sizes are computed on the exact values.

[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.

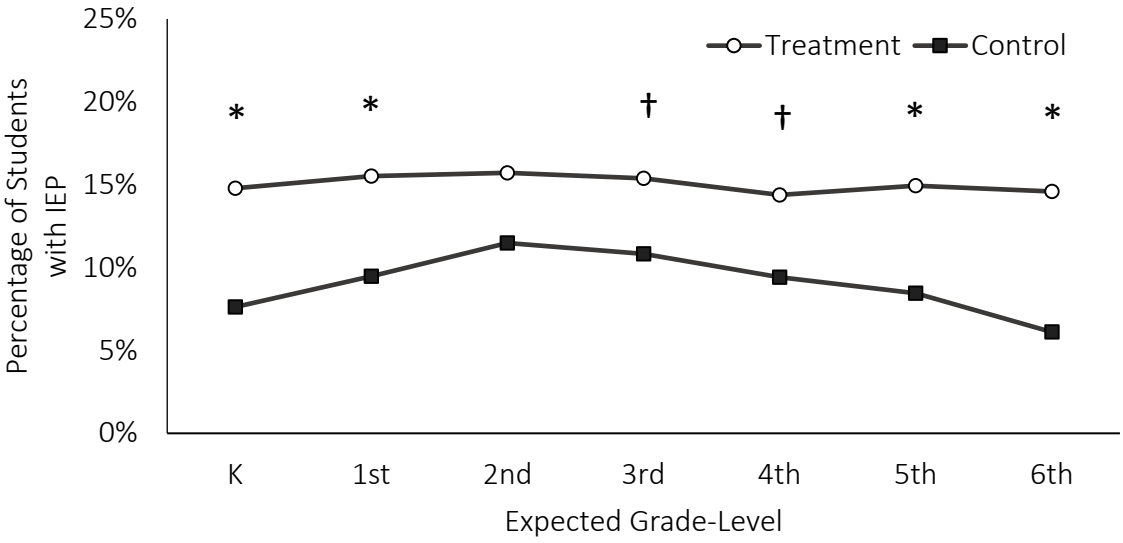
[e] The p-values for statistical significance that are the same for the ITT and TOT coefficients

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



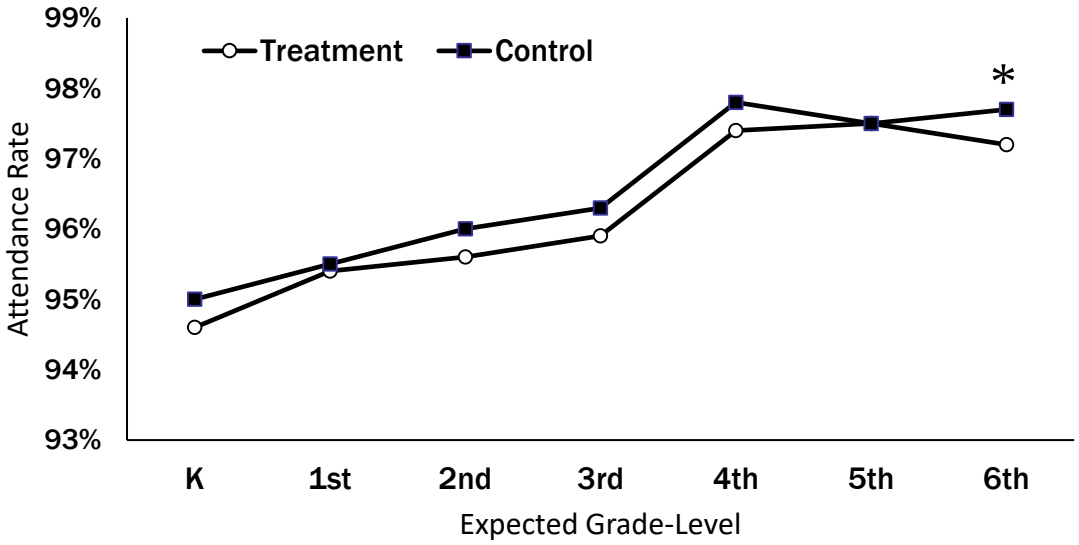
Note. Asterisks indicate $p < .05$ and obelisks indicate $p < .10$. Detailed results for kindergarten through sixth grade are located in Supplemental Table S5.

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.