The District-wide Effectiveness of the Achieve3000 Program:

A Quasi-Experimental Study

Geoffrey D. Borman

Measured Decisions, Inc.

University of Wisconsin—Madison

So Jung Park

Sookweon Min

University of Wisconsin—Madison

Completed: January 15, 2015

Introduction

A significant body of research has demonstrated the importance of literacy skills for students' overall academic success (e.g., Cunningham & Stanovich, 1998). In particular, scholars have emphasized improving students' literacy in the early grades as crucial to helping promote long-term academic success (Whitehurst & Lanigan, 2001). Currently, however, a number of students are struggling with literacy in American schools. According to the National Center for Educational Statistics (NCES, 2011), approximately two-thirds of 4th and 8th grade students are not successful in reaching proficient-level reading scores. In order to successfully meet the high demands on students in literacy, there have been numerous programs and interventions for the development of students' literacy using various strategies.

The Achieve3000 programs are designed to improve students' reading and writing skills by differentiated online instruction. Achieve3000 programs are promulgated on the idea that high levels of reading and writing skills play a crucial role in students' academic success, college readiness, and preparation for the job market (Achieve3000, 2014). In order to move students to higher levels of reading and writing proficiency, Achieve3000 programs teach all students at their individual reading levels, while continually challenging them to achieve the next level of success (Achieve3000, 2014). By providing students differentiated instruction, the program intends to meet the educational needs of individual students effectively, and also to prepare them to thrive in the society in the long run (Achieve3000, 2014).

Achieve3000 has specific program options for students in various grade levels: *KidBiz3000* (for grades 2-5); *TeenBiz3000* (for grades 6-8); and *Empower3000* (for

grades 9-12). These programs are designed to help students read at their Lexile®/reading level by providing them differentiated online instruction, which adjusts content according to each student's Lexile®/reading level. All of the programs align with the objectives of the Common Core State Standards to provide students the content area literacy skills they need to succeed on the standards (Achieve3000, 2014). In addition, Achieve3000 products offer programs for diverse student groups, including English language learners (ELLs), struggling readers, and adult learners. To be specific, the program for ELLs, *Achieve Language*, provides additional support to English language learners as well as to their teachers in order to effectively meet ELL students' linguistic needs. The program for struggling students, *Achieve Intervention*, offers intensive-evidence based intervention to students who are reading below their level to close the learning gaps among students. The program for adult learners, *Spark3000*, focuses on improving learners' literacy skills and so helping their success in career (Achieve3000, 2014).

Correlational and pre-post evaluations of Achieve3000 have shown Lexile®/reading growth across all grade levels. Greater gains were realized for participating students who completed at least two reading sessions per week, students who scored 75% or higher on the multiple choice activity, students who scored below grade level (two or more years), and students who were English Language Learners (National Lexile Study, 2014). In this evaluation of Achieve3000, we utilize student-level data from Chula Vista, California to assess the pre-to-post California Standards Test (CST) English-language-arts achievement outcomes for Achieve3000 students relative to a comparison group of students from the district formed using Inverse Probability-of-Treatment Weighting propensity score matching methods.

Method

Data

We employed English-language-arts CST data provided by the Chula Vista, California school district for this evaluation of Achieve3000 programs offered throughout the district. The district implemented the Achieve3000 programs during the 2010-2011 through 2011-2012 academic years. In the 2010-2011 academic year, 16 schools in the Chula Vista district implemented Achieve3000 programs. Table 1 presents the specific information on the program schools implementing Achieve3000 programming in the district, including the number of enrolled students and the number of Achieve3000 students in each school.

Achieve3000 Implementation in Chula Vista

Schools implementing the Achieve3000 programs in the Chula Vista district had Professional Learning Services (PLS) that they purchased with their order. These Professional Learning Services provided 1-3 days of on-site support from an Implementation Manager. At the beginning of the school year, this support provided training for teachers new to the school and offered the opportunity to share new program features with returning teachers. Throughout the year, these services also involved classroom modeling, which was conducted by the Implementation Manager who modeled a lesson for the teachers in order to observe best practices, as suggested by Achieve3000. The PLS also provided one-on-one consulting sessions allowing the Implementation Manager to review student data along with the participating teachers.

Table 1. Information on Achieve3000 Schools in the Chula Vista District

School Name	# Students	# Achieve3000 students	% Achieve 3000 students
Chula Vista Hills Elementary School	557	54	0.10
Chula Vista Learning Communities Charter School	959	259	0.27
Discovery Charter School	853	279	0.33
Eastlake ES	591	179	0.30
Rohr ES	373	54	0.14
Hazel Goes Cook ES	481	147	0.31
Heritage ES	892	338	0.38
Juarez-Lincoln ES	601	43	0.07
Liberty ES	847	271	0.32
Loma Verde ES	493	42	0.09
Olympic View ES	541	107	0.20
Rosebank ES	659	201	0.31
The Daly Academy	24	4	0.17
Thurgood Marshall ES	727	31	0.04
Valley Vista ES	548	112	0.20
Wolf Canyon ES	936	330	0.35

Sample

The treatment group consisted of all participants in Achieve3000 programs in the Chula Vista district. We identified a total of 2,625 students as Achieve3000 participants. Students excluded from this sample, however, included 645 students who lacked a 2011 California Standards Test (CST) English-language arts score or a 2012 English-language arts CST score, and 18 duplicated cases. Additionally, we exclude all 3rd graders (5 students in treatment group) and 9th graders (0 students in treatment group) from the

sample. The 3rd graders were excluded because most of them were retained and did not have 2nd grade test (pretest) scores (there is no 2nd grade standardized test in CA). The 9th grade students were excluded because there were no treatment students at that grade level. As a result, 1,957 4th through 8th graders had complete data in the Achieve3000 treatment group and were deemed eligible for the quasi-experimental study in the Chula Vista district.

A total of 7,675 students, who enrolled at other demographically similar Chula Vista district schools that did not have access to the Achieve3000 program and had 2011 and 2012 California Standards Test (CST) English language-arts scores, were identified as a possible comparison group. Of 7,675 control group students, 27 students who had duplicate records, 14 students from 3rd grade, and 36 students from 9th grade were excluded from the sample. As a result, 7,598 4th through 8th graders who had complete data were deemed eligible for the comparison group sample

Table 2. Information on Analytic Sample

	Treatment Group	Control Group
Initial sample	2,625	27,524
No information on pretest score or posttest score	645	19,849
Duplicated records	18	27
9th grade students	0	36
3rd grade students	5	14
Final analytic sample	1,957	7,598

Measures

Dependent Variable. We used students' English-language Arts (ELA) scores on the California Standards Test (CST) in the year 2011-12 as the outcome measures in our models. Students' ELA scores on the CST served as posttest outcomes for students.

Independent Variables. For students' academic background information, we used their ELA scores on the CST in the year 2010-2011. These scores served as pretest measures of students' academic achievement before they joined the Achieve3000 program. In addition to students' pretest information, their gender, race/ethnicity, socioeconomically disadvantaged status (SED), English Language Learner (ELL) status, and grade level information were also included as student demographic background information. Gender was dummy coded (1 = Female, 0 = Male). Race/ethnicity indicators were coded as four dummy variables: Asian (1 = Asian, 0 = non-Asian), Hispanic (1 = Hispanic, 0 = non-Hispanic), Black (1 = Black, 0 = non-Black), and others (1 = other, 0 = non-other; Whites as a reference group). Socioeconomically disadvantaged status (SED) was determined by whether a student qualified for a free or reduced-price lunch (1 = FRL status, 0 = non-FRL status). English Language Learner (ELL) status was dummy coded (1 = ELL, 0 = non-ELL). Finally, we include indicators of students' grade level, 4^{th} through 8^{th} grades. Each grade level indicator, 5 through 8, was dummy coded, with 4^{th} grade as the reference group.

Analytical Approach

To attenuate possible selection bias, we used a propensity score method, Inverse Probability-of-Treatment Weighting (IPTW), to match Achieve3000 students and comparison students. The goal of this method was to compare observed comparison units

having a similar probability of being selected for the treated group. Using a logistic regression model, we estimated each student's conditional probability of selection for the Achieve3000 program as a function of student background characteristics and pretest scores. Specifically, the model included the following individual student characteristics: 2010-2011 CST ELA scores; indicators for race/ethnicity (Asian, Hispanic, African Americans, and others); gender; socioeconomically disadvantaged status (SED); English Language Learner (ELL) status; and grade information (4^{th} - 8^{th} grades). We also included interaction terms between pretest and the grade-level indicators in the model. The fitted values of this model estimate the probabilities, or propensities, that children in the sample will be offered Achieve3000 and provide an index that optimally summarizes the information the covariates contain (Murnane & Willett, 2010).

Specifically, we utilized Inverse Probability-of-Treatment Weighting (IPTW) methods. IPTW methods allowed us to weight each person by the inverse of the estimated propensity score (Rosenbaum, 1987; Sharkey & Sampson, 2010). Both Cohen's d[1] and the variance ratio[2] were computed in order to check that treatment and control group students were balanced on covariates resulting from IPTW (Rubin, 2001). After the balance diagnostic tests, we used a Weighted Least Squares (WLS) regression model with IPTW weights to estimate the average Achieve3000 treatment effect.

[1] $d = (\bar{X}_t - \bar{X}_c)/\sqrt{S_t^2 - S_c^2}$, d should be close to zero ($|d| < 0.1$)
[2] $v = S_t^2/S_c^2$, v should be close to one ($4/5 < v < 5/4$)

Results

Descriptive Statistics

In order to understand the background information of the control group and that of the treatment group, descriptive statistics were examined before the main analyses of this study. Table 3 provides information on the difference in posttest scores between the control group and the treatment group. Results in Table 3 suggest that the majority of students in the study were from grades 4 through 6 and that the average post-test scores for the treatment students range from 379.36 to 402.83, whereas those for the control students range from 361.83 to 387.97.

Table 3. Comparison of Posttest scores between the Control Group and the Treatment Group

	Treatment Group			Control Group		
Grade	Student N	M	SD	Student N	M	SD
4th grade	502	379.36	60.86	2,504	361.83	57.79
5th grade	679	392.46	54.57	2,347	387.97	51.66
6th grade	746	379.86	53.03	2,308	378.72	51.18
7th grade	18	402.83	44.57	240	384.66	45.30
8th grade	12	384.14	55.98	199	376.16	54.41
ALL	1,957	384.34	55.84	7,598	376.23	54.41

In addition to the information on the posttest scores in Table 3, Table 4 summarizes all key demographic characteristics and pretest measures of both the treatment group and the control group. We evaluated the statistical significance of all treatment-comparison differences at baseline using a *t* test for the continuous measures and a χ^2 analysis for the dichotomous outcomes.

Table 4. Comparison of Baseline Characteristics for Achieve3000 Program Students before PS adjustment (n = 1,957)

Variable	Condition	N	M	*SD*	Mean difference (imbalance)	t	χ^2
Pre-test Score	Control	7,612	376.55	59.28	4.32	-2.87**	
	Treatment	1,962	380.87	59.42			
Hispanic	Control	7,612	0.71	0.45	-0.10		66.14***
	Treatment	1,962	0.61	0.49			
Asian	Control	7,612	0.13	0.34	0.04		21.93***
	Treatment	1,962	0.17	0.38			
White	Control	7,612	0.11	0.32	0.03		14.85***
	Treatment	1,962	0.15	0.35			
Black	Control	7,612	0.03	0.17	0.01		7.22**
	Treatment	1,962	0.04	0.20			
Other	Control	7,612	0.02	0.14	0.01		8.84**
	Treatment	1,962	0.03	0.17			
Female	Control	7,612	0.51	0.50	0.00		0.71
	Treatment	1,962	0.51	0.50			
SED	Control	7,612	0.52	0.50	-0.14		130.90***
	Treatment	1,962	0.37	0.48			

ELL	Control	7,612	0.32	0.47	-0.10	80.48***
	Treatment	1,962	0.22	0.41		
4th grade	Control	7,612	0.33	0.47	-0.07	38.51***
	Treatment	1,962	0.26	0.44		
5th grade	Control	7,612	0.31	0.46	0.04	10.42**
	Treatment	1,962	0.35	0.48		
6th grade	Control	7,612	0.30	0.46	0.08	42.90***
	Treatment	1,962	0.38	0.49		
7th grade	Control	7,612	0.03	0.17	-0.02	29.69***
	Treatment	1,962	0.01	0.10		
8th grade	Control	7,612	0.03	0.16	-0.02	29.00***
	Treatment	1,962	0.01	0.08		

Note * $p < .05$. ** $p < .01$. *** $p < .001$

Balance Checks

In order to check whether the PS adjustment approach removed the observed baseline differences in covariates, we compared the balance result from before conducting the PS adjustment and the result from after conducting the adjustment. We first checked the overlap of the treatment and control group's distribution for PS-logit (see Appendix 1). Then, we deleted 28 non-overlapping cases, which have non-corresponding treatment students, because it is usually not possible to achieve balance with groups that show regions of non-overlap on the PS-logit (Steiner & Cook, 2013).

Table 5 shows Cohen's *d* and variance ratio before and after PS-adjustment. Before the PS adjustment, there was an imbalance between the treatment group and the comparison group, indicated by Cohen's *d* values of 0.10 or greater as well as a variance ratio of less than 0.8 or greater than 1.25. After the PS adjustment, most Cohen's *d* values were close to 0, and the variance ratios for the variables were close to 1. Due to the small number of non-overlapping cases, the results of the balance diagnostic tests (both Cohen's *d* and the variance ratio) are nearly the same for both the whole and overlapping samples. Appendices 2-4 present these results visually. For the following analyses, we report the results from the overlapping sample.

Analyses of Achieve3000 Outcomes

The main analyses compare the achievements of the Achieve3000 students to those of the comparison students. Table 6 presents the results of Weighted Least Squares (WLS) regression models estimating the differences between the Achieve3000 students and their counterparts on post-test scores after we excluded 28 non-overlapping cases.

Table 5. Comparison of Balance Checks

	Before PS Adjustment		After PS Adjustment (Whole sample)		After PS Adjustment (Overlapping Sample)	
	Cohen's *d*	Variance Ratio	Cohen's *d*	Variance Ratio	Cohen's *d*	Variance Ratio
Pre-test score	0.07	1.00	0.002	1.02	0.003	1.02
Hispanic	-0.20	1.15	-0.03	1.02	-0.03	1.02
Asian	0.12	1.25	0.02	1.04	0.02	1.03
Black	0.07	1.40	0.01	1.05	0.01	1.04
Other	0.07	1.56	0.01	1.04	0.01	1.05
Female	-0.01	1.00	-0.01	1.00	-0.01	1.00
SED	-0.30	0.94	-0.04	1.00	-0.04	1.00
ELL	-0.24	0.78	-0.02	1.00	-0.02	1.00
4^{th} grade	-0.16	0.86	0.00	1.00	0.00	1.00
5^{th} grade	0.08	1.06	0.00	1.00	0.00	1.00
6^{th} grade	0.16	1.12	0.01	1.01	0.01	1.01
7^{th} grade	-0.16	0.30	-0.03	0.81	-0.03	0.81
8^{th} grade	-0.16	0.24	-0.01	0.94	0.01	1.08

The research sample was evaluated with two models; the first model included only the treatment indicator and the second model examined the outcomes for treatment and comparison after controlling for students' pretest scores as well as their demographic information.

The findings presented in Table 6 show that the overall test score for the treatment group is higher than that of the comparison group and the differences between the two groups are statistically significant. These posttest differences held even after controlling for students' pretest scores and demographic information. The overall average effect sizes of Achieve3000 programs, across all grade level, are 0.05 before adjusting covariates and 0.04 after adjusting covariates (see Table 12).

However, the average results may mask key differences across grade levels in terms of the Achieve3000 programs effectiveness. As the results in Tables 7-11 show, the outcomes of Achieve3000 did, indeed, vary across grade levels. To be specific, the outcomes favored the Achieve3000 students in grades 4, 7, and 8 (see Table 7, 10, and 11, respectively), whereas there were no posttest differences between treatment and comparison students in grades 5 and 6 (see Table 8 and 9, respectively). Table 12 summarizes the effect sizes of the Achieve3000 program as revealed by the subgroup analysis by grade level. The effect sizes (*ES)* were calculated as WLS regression coefficients divided by the standard deviations of the respective outcomes.

Table 6. WLS Regression Model Predicting Overall Achieve3000 Program Effect

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	377.36***	0.80	104.84***	2.92
Treatment	2.61*	1.13	2.35***	0.67
pre-test score			0.69***	0.01
Hispanic			-2.35*	1.10
Asian			1.52	1.30
Black			-6.87**	2.11
Other			-1.16	2.47
Female			2.08**	0.67
SED			-5.74***	0.74
ELL			-8.22***	0.86
5^{th} grade			38.54***	0.85
6^{th} grade			9.13***	0.84
7^{th} grade			23.84***	2.22
8^{th} grade			27.16***	2.45

Note *p < .05. **p < .01. ***p < .001

Table 7. WLS Regression Model Predicting Achieve3000 Program Effect for 4th Graders

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	363.49***	1.52	115.57***	5.27
Treatment	6.95**	2.16	5.96***	1.35
Pre-test score			0.67***	0.01
Hispanic			-1.42	2.24
Asian			4.39	2.62
Black			-6.67	4.24
Other			-4.71	4.55
Female			-2.44	1.36
SED			-8.15***	1.48
ELL			-10.92***	1.62

Note *$p < .05$. **$p < .01$. ***$p < .001$

Table 8. WLS Regression Model Predicting Achieve3000 Program Effect for 5th Graders

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	388.82***	1.36	137.67***	4.78
Treatment	1.00	1.93	0.68	1.13
Pre-test score			0.70***	0.01
Hispanic			-1.79	1.85
Asian			2.73	2.21
Black			-5.61	3.65
Other			1.33	4.35
Female			6.04***	1.14
SED			-1.75	1.24
ELL			-7.33***	1.42

Note *$p < .05$. **$p < .01$. ***$p < .001$

Table 9. WLS Regression Model Predicting Achieve3000 Program Effect for 6th Graders

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	379.21***	1.35	105.29***	4.94
Treatment	-1.76	1.90	-1.02	1.09
Pre-test score			0.72***	0.01
Hispanic			-2.94	1.80
Asian			-1.07	2.11
Black			-5.66	3.36
Other			1.60	4.12
Female			2.81*	1.10
SED			-6.56***	1.21
ELL			-5.83***	1.51

Note *$p < .05$. **$p < .01$. ***$p < .001$

Table 10. WLS Regression Model Predicting Achieve3000 Program Effect for 7th Graders

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	384.88***	3.69	119.32***	17.25
Treatment	19.04***	5.53	17.95***	4.16
Pre-test score			0.70***	0.04
Hispanic			-4.16	5.04
Asian			-5.36	6.66
Black			-29.42***	8.76
Other			19.17	17.22
Female			6.79	3.49
SED			-5.03	4.50
ELL			1.12	5.79

Note *p < .05. **p < .01. ***p < .001

Table 11. WLS Regression Model Predicting Achieve3000 Program Effect for 8th Graders

	Treatment Effect		Treatment Effect after Adjusting Other Covariates	
	Coefficients	*SE*	Coefficients	*SE*
(Intercept)	374.35***	4.48	104.33***	24.62
Treatment	13.88*	6.22	8.14*	3.59
Pre-test score			0.79***	0.06
Hispanic			-9.61	5.54
Asian			12.05	16.55
Black			-28.23	14.83
Other			-20.22	22.66
Female			-6.00	3.77
SED			-3.76	4.02
ELL			-5.97	5.61

Note *$p < .05$. **$p < .01$. ***$p < .001$

Table 12. Effect Size of the Achieve3000 program by Grade Level[3][4]

	Achieve3000 Effect Size	Achieve3000 Effect Size After Adjusting Other Covariates
4th grade	0.12	0.10
5th grade	0.02	0.01
6th grade	-0.03	-0.02
7th grade	0.42	0.40

[3] In order to make sure that different effect sizes by grade levels in Table 12 are not due to poor matching, we checked the balance on the pretest measure for each grade. The results show that pretest differences after PS-adjustment are not statistically significant for each grade. For detailed information, see Appendices 5 and 6, respectively.

[4] The overlapping sample was used to obtain the results shown in Table 6~12. Although discarding non-overlapping cases on the observed PS-logit is often recommended, doing so can result in reduced generalizability of results (Steiner & Cook, 2013). We conducted a sensitivity analysis using the whole sample including non-overlapping cases. The results show that the effect size for the sample group is the same as the result for the sample after excluding the 28 non-overlapping cases, except for the 8th graders: the effect size for 8th graders is 0.18 before adjusting other covariates and also 0.16 after adjusting other covariates.

8th grade	0.29	0.17
ALL	0.05	0.04

Relationships Between Treatment Participation and Achievement

In order to assess the relationship between program participation and Achive3000 outcomes, we examined several measures of students' program participation and the associations between these measures and student posttest outcomes. Table 13 provides the descriptive information on five different groups for the program-participation fidelity analyses: students who completed on average, one activity per week (Group 1); students who completed on average, two activities per week (Group 2); students who scored 75% or higher on the multiple choice activity (Group 3); students who completed on average, one activity per week and scored 75% or higher on the multiple choice activity (Group 4); and students who completed on average, two activities per week and scored 75% or higher on the multiple choice activity (Group 5).

Table 13. Groups Defined for the Achieve3000 Participation Analysis

Group	Program participation	Student N
Group 1	Students who completed on average, one activity per week	1,184
Group 2	Students who completed on average, two activities per week	631
Group 3	Students who scored 75% or higher on the multiple choice activity	565
Group 4	Students who completed on average, one activity per week and scored 75% or higher on the multiple choice activity	438
Group 5	Students who completed on average, two activities per week and scored 75% or higher on the multiple choice activity	274

Table 14 demonstrates the results of the program participation analyses before controlling for students' background information. The results shown in Table 14 indicate

that students who were involved in activities more frequently and those who scored higher on activities tended to achieve significantly higher on their posttest scores.

In addition, Table 15 displays the relationship between Achieve3000 participation after controlling for students' demographic information and their academic backgrounds. Similar to models shown in Table 14, the results of Table 15 also indicate that more frequent and more successful involvement in Achieve3000 activities were statistically significant predictors of Achieve3000 students' posttest scores, even after taking student background into account. These results suggest that the student participation measures, measuring fidelity of Achieve3000 participation, are important predictors of differences in students' Achieve3000 program achievement outcomes.

Table 14. Relationships Between Program Participation and Posttest Outcomes

	Posttest Outcomes by Group				
	Group 1 Coefficient (*SE*)	Group 2 Coefficient (*SE*)	Group 3 Coefficient (*SE*)	Group 4 Coefficient (*SE*)	Group 5 Coefficient (*SE*)
Intercept	370.10***	377.17***	371.76***	374.23***	378.74***
	(1.97)	(1.51)	(1.42)	(1.36)	(1.33)
Program Participation	23.54***	22.23***	45.51***	48.01***	45.53***
	(2.53)	(2.65)	(2.60)	(2.83)	(3.49)

Note *$p < .05$. **$p < .01$. ***$p < .001$

Conclusion

Given the strong link between literacy and overall academic performance, the significance of students' reading and writing skills has become increasingly emphasized in American education (Whitehurst & Lanigan, 2001). While there have been a number of interventions aimed at improving reading and writing skills, the effectiveness of those programs has been mixed (D'Agostino & Murphy, 2004). By examining the effectiveness

of Achieve3000, one of the recent efforts to help students develop their skills in reading and writing by using differentiated online instruction, we expect to improve our understanding on how the program currently works for enhancing students' reading achievement and to provide suggestions to improve the program's quality in the long run.

The findings of this study provide evidence that students participating in the Achieve3000 programs had statistically higher California Standards Test English-language-arts outcomes, relative to a well-matched comparison group of similar students, after one-year of implementation in the Chula Vista, CA school district. This result is particularly encouraging since the statistically significant difference favoring the program group holds even after controlling for students' various backgrounds including ethnicity, gender, students' socioeconomic status, and their language status. Because our balance

Table 15. Relationships Between Program Participation and Posttest Outcomes after Adjusting Other Covariates

	Posttest Outcomes by Group				
	Group 1	Group 2	Group 3	Group 4	Group 5
	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)	Coefficient (*SE*)
Intercept	101.11***	102.86***	113.06***	112.66***	109.75***
	(6.51)	(6.51)	(6.72)	(6.72)	(6.71)
Pre-test score	0.70***	0.70***	0.68***	0.68***	0.69***
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Hispanic	-0.54	-0.79	0.20	-0.06	-0.64
	(2.30)	(2.29)	(2.32)	(2.32)	(2.33)
Asian	4.65	4.39	5.60*	5.02	4.73
	(2.69)	(2.70)	(2.70)	(2.70)	(2.71)
Black	-4.30	-4.10	-1.91	-2.68	-3.10
	(4.22)	(4.23)	(4.24)	(4.24)	(4.25)
Other	-0.37	-0.54	-0.56	-1.08	-1.42
	(4.78)	(4.78)	(4.83)	(4.83)	(4.85)
Female	1.66	1.72	1.38	1.16	1.54
	(1.51)	(1.51)	(1.52)	(1.53)	(1.53)
SED	-5.39**	-5.52**	-5.22**	-5.13**	-5.23**
	(1.68)	(1.68)	(1.69)	(1.70)	(1.70)

ELL	-9.44***	-9.49***	-9.89***	-9.84***	-9.93***
	(2.08)	(2.08)	(2.10)	(2.11)	(2.11)
5th grade	35.46***	35.23***	33.32***	33.66***	34.02***
	(2.00)	(2.00)	(2.02)	(2.02)	(2.02)
6th grade	5.10**	5.04 **	2.89	3.31	3.58
	(1.92)	(1.92)	(1.95)	(1.94)	(1.95)
7th grade	24.93**	25.17**	24.14**	24.26**	23.95**
	(7.96)	(7.97)	(7.91)	(7.92)	(7.95)
8th grade	29.05**	29.19**	27.02**	26.44**	26.89**
	(9.69)	(9.71)	(9.63)	(9.63)	(9.67)
Program Participation	6.74***	6.09***	12.59***	13.23***	12.65***
	(1.59)	(1.66)	(1.78)	(1.92)	(2.25)

Note *$p < .05$. ** $p < .01$. *** $p < .001$

checks revealed an overall baseline effect size difference between our Achieve3000 student group (n = 1,957) and the comparison group (n = 7,598) of d = 0.003, this suggests that the baseline equivalence of our quasi-experimental matched comparison group meets widely recognized standards for assessing pre-intervention differences between groups (Ho, Imai, King, & Stuart, 2007; Rubin, 2001; What Works Clearinghouse, 2014). Namely, because the pretest difference separating the Achieve3000 and comparison group was less than d = .25, and because we used OLS regression adjustment for the small pretest difference of d = 0.003, we can be more confident that the posttest differences favoring Achieve3000 students were not a result of any pre-existing achievement differences.

In addition, we found that the effectiveness of the program varied considerably across grades, though positive and statistically significant posttest differences were found across several grade levels. This evidence, though exploratory, may help in formative ways to identify which grade levels seem to benefit more or less from the Achieve3000 program. In particular, the largest effect sizes were found for 7th grade, d = .40, and 8th

grade, d = .17. This result may suggest that the Achieve3000 program for 7th and 8th grade students is particularly effective. However, these larger effect sizes are ultimately based on relatively small sample sizes and are, therefore, estimated with less precision.

Due to data limitations, this study does not account for the multi-level structure of the data (the clustering of students within schools). Further research may address this limitation by examining how the program's impact varies across schools. Additionally, future research may examine whether the program has a longer-term impact on student learning.

References

Achieve3000 (2014) Achieve3000 official website. Retrieved from http://www.achieve3000.com/

Cunningham, A E, & Stanovich, K E (1998). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945.

D'Agostino, J. V., & Murphy, J. A. (2004). A meta-analysis of Reading Recovery in United States schools. *Educational Evaluation and Policy Analysis, 26*(1), 23-28.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236.

Murnane, R. J., & Willett, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* Oxford University Press.

National Center for Educational Statistics. (2011). *National assessment of educational*

progress: 2011 reading assessments. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387-394.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3-4), 169-188.

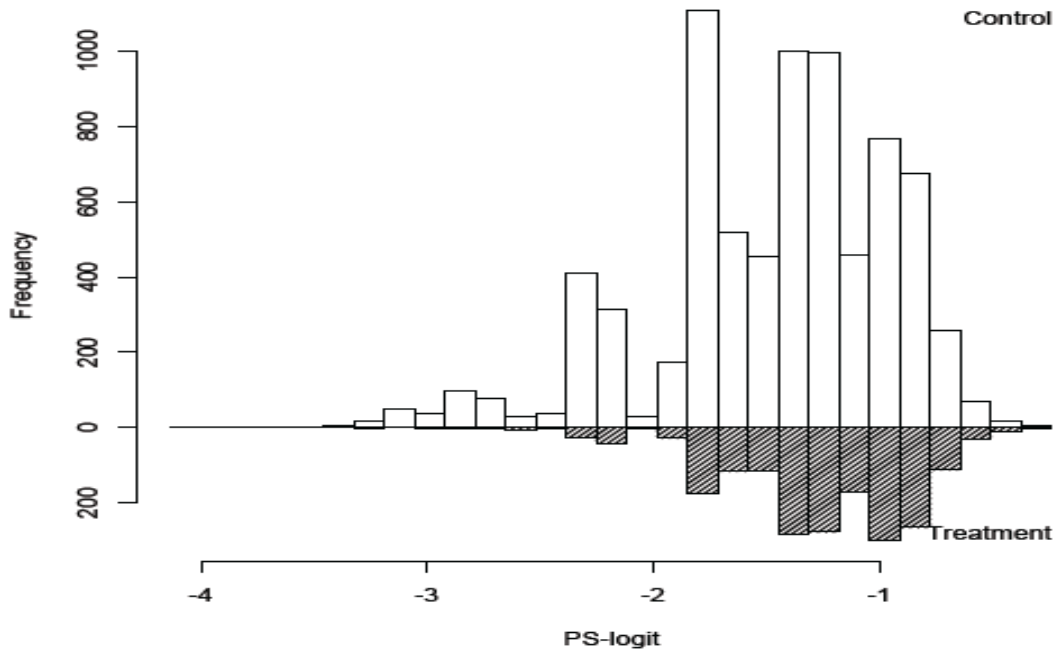
Sharkey, P., & Sampson, R. J. (2010). Destination effects: Residential mobility and trajectories of adolescent violence in a stratified metropolis. *Criminology*, *48*(3), 639-681.

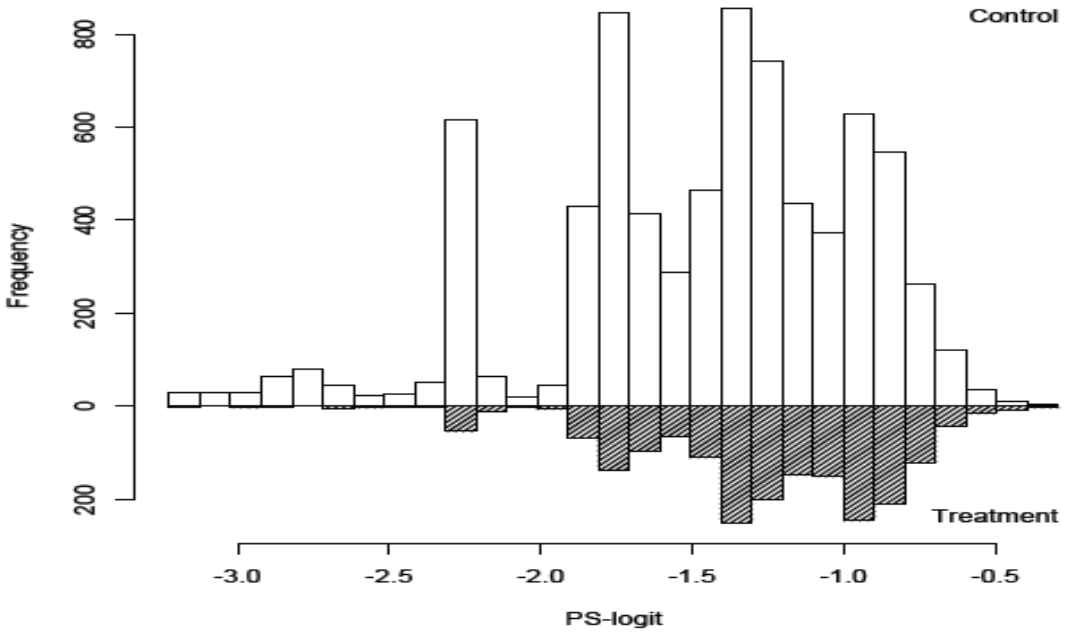
Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. *The Oxford Handbook of Quantitative Methods in Psychology*, *1*, 237.

What Works Clearinghouse (2014). *What Works Clearinghouse procedures and standards handbook* (version 3.0). Washington, DC: Institute of Educations Sciences.

Whitehurst, G l, & Lanigan, C l (2001). Emergent literacy: Development from prereaders to readers In S .. B. Neuman & D. K Dickinson (Eds . .), *Handbook of early literacy research* (pp. 11-29). New York: Guilford.

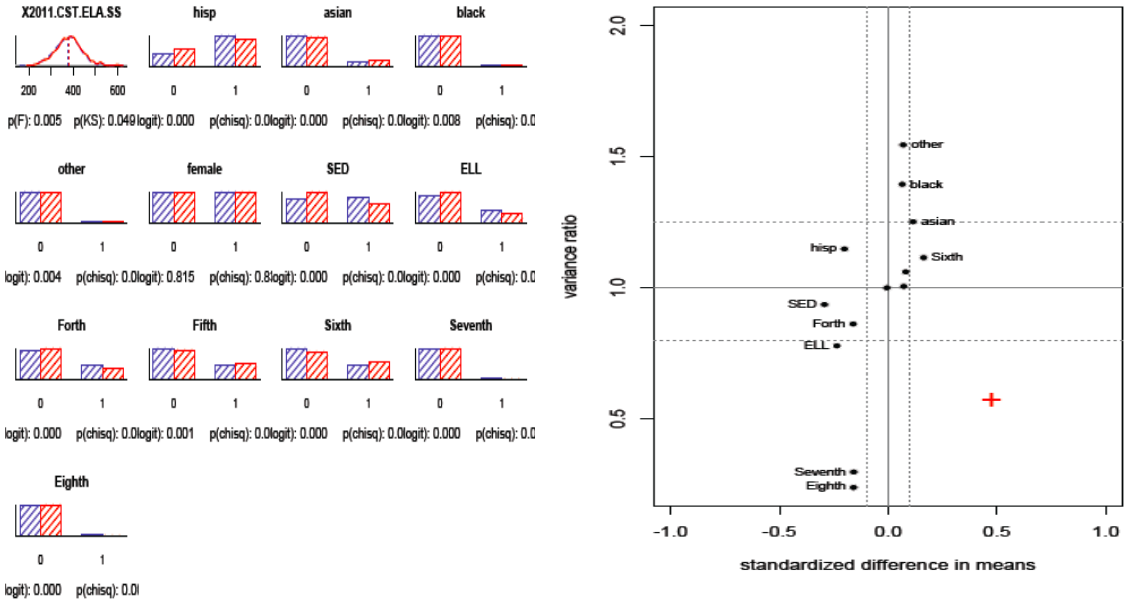
Appendix 1. Overlap of the treatment and control group on the estimated PS-logit



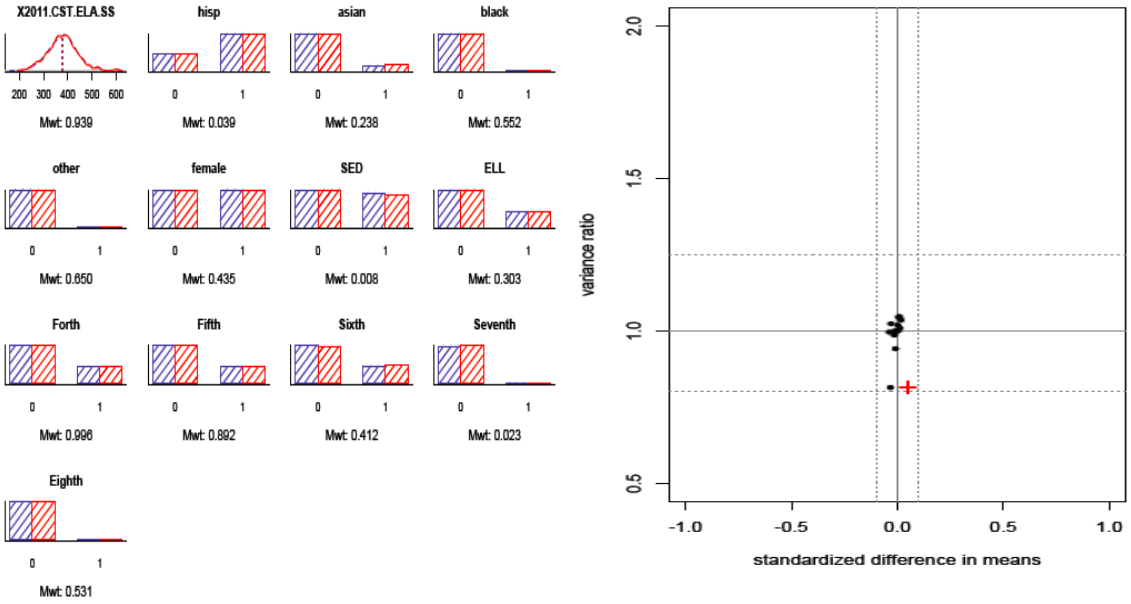


# of non-overlapping cases: 24, breaks: -3.225964 -0.2661278

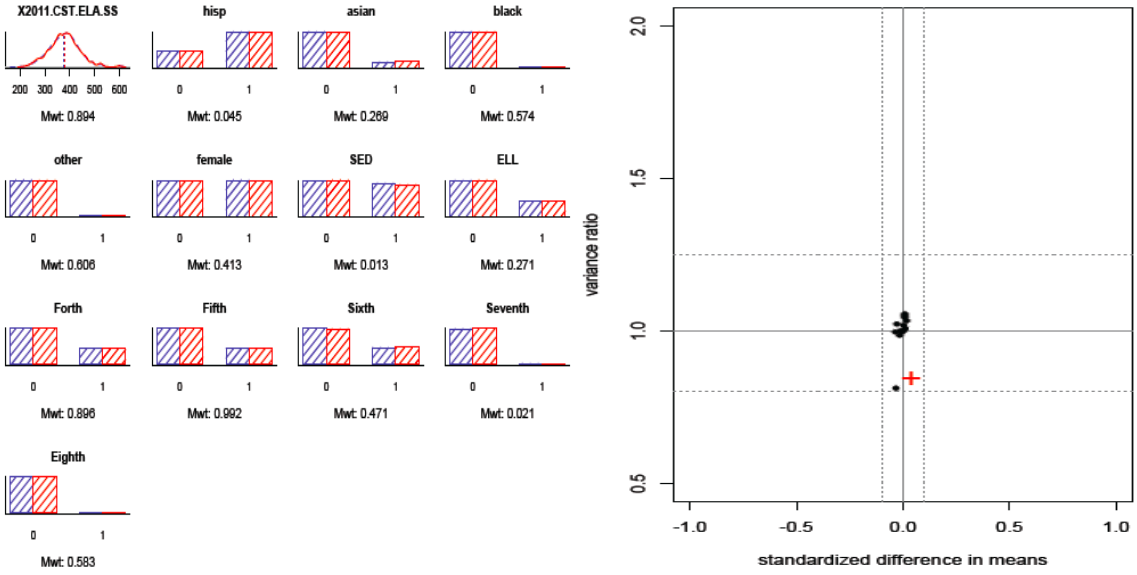
Appendix 2. Balancing plots: Initial imbalance before PS adjustment



Appendix 3. Balance after PS adjustment (with all cases)



Appendix 4. Balance after PS adjustment (with overlapping cases only)



Appendix 5. Pretest differences by grade levels before adjusting IPTW

Grade Level	Control	Treatment	Diff	t-value
4th grade	381.95	393.62	11.67***	-3.78
5th grade	360.34	363.89	3.55	-1.42
6th grade	386.96	387.92	0.96	-0.41
7th grade	384.34	386.61	2.27	-0.18
8th grade	369.69	361.75	-7.94	0.63
All students	376.55	380.87	4.32**	-2.87

Note * p < .05. ** p < .01. *** p < .001

Appendix 6. Pretest differences by grade levels after adjusting IPTW

Grade Level	Control	Treatment	Diff	t-value
4th grade	383.90	384.83	0.93	0.41
5th grade	361.14	361.28	0.13	0.06
6th grade	387.20	386.42	-0.78	-0.39
7th grade	384.51	385.42	1.33	0.22
8th grade	363.80	367.09	3.30	0.56
All students	377.18	377.35	0.16	0.13

Note * *p* < .05. ** *p* < .01. *** *p* < .001