HISD | Research and Accountability
ANALYZING DATA, MEASURING PERFORMANCE.

EVALUATION REPORT
BUREAU OF DISTRICT DATA ANALYSIS
Volume 9, Issue 1, July 2015

Achieve3000® impacts on student reading and STAAR EOC English I, English II, and Biology exams for the 2014–2015 academic year.

By D. Diego Torres, Ph.D.

Based on the Lexile® Framework, a scientific approach to reading and text measurement backed by more than two decades of ongoing research, Achieve3000®, a web-based differentiated reading program used by HISD, is designed to improve student reading ability and comprehension of increasingly complex texts by initially meeting students where they are academically. The present study focuses on ninth and tenth graders and assesses the impact deriving from the use of Achieve3000's reading solutions on (1) the change between students' pre- and post-test Lexile scores as measured by LevelSet™, (2) students' scale score performance on the State of Texas Assessment of Academic Readiness End-Of-Course (STAAR EOC) for the English I, English II, and Biology exams, and (3) students' probability of meeting both the Phase-in 1 and Phase-in 2 passing standard on the STAAR EOC English I and English II. Results suggest that the more students use Achieve3000's web-based software, the greater their gains in post-test Lexile scores, the greater their achievement on the STAAR EOC exams, and the higher their probability of meeting the Phase-in 1 and Phase-in 2 passing standard on the STAAR EOC exams. Based on these findings, it is recommended that teachers in schools that have Achieve3000 licenses actively encourage the completion of Achieve3000 exercises.

Background

Increased literacy early in life has long been acknowledged to be an important factor to educational attainment and psychosocial and physical health in adulthood (DeWalt et al., 2004; Smith, 2009). Failure to achieve a minimum standard of reading comprehension by adulthood is also consequential to the health of poor readers' children (DeWalt & Hink, 2009), extending disadvantage to another generation. Despite this reality of the importance of improved literacy, the 2013 National Assessment of Educational Progress, colloquially referred to as the nation's report card, revealed that only about a third of fourth, eighth, and 12th grade students performed at or above the proficient level of reading (National Center for Education Statistics, 2013). And among proficient readers, smaller proportions were racial or ethnic minorities relative to whites.

In addition to the cold statistics regarding reading achievement, specifically, education policymakers are also concerned about the general noninterest in reading among students in all grades (Greenberg, Gilbert, & Fredrick, 2006), which scholars have attributed to a static system that does not take into account differences among student needs with respect to reading instruction and time spent on reading (Pitcher et al., 2007). To address the general deficiency in reading comprehension and literacy among school-age children, and to meet students where they are educationally, schools and districts across the United States have begun implementing differentiated reading interventions.

One such intervention being used by the Houston Independent School District (HISD) is Achieve3000®, a web-based differentiated reading program. Now concluding its second year of implementation, Achieve3000 has been made available to students in the district's most vulnerable schools. Rooted in the Lexile® Framework developed by MetaMetrics, Inc., Achieve3000 matches students to informational and fictional texts, only augmenting the level of reading difficulty

as students demonstrate readiness. The theory is that by teaching students at a level at which they are most comfortable they will progress more smoothly to more advanced levels of reading comprehension and literacy than they would under a standard curriculum model that fails to appreciate mixed-ability classrooms.

Research Questions

The purpose of the present study is to assess the efficacy of Achieve3000 at raising student achievement. The specific questions this study will answer are as follows:

1) Does students' use of Achieve3000 lead to higher scores on the State of Texas Assessment of Academic Readiness End-Of-Course (STAAR EOC) English I, English II, and Biology subject exams?
2) Does students' use of Achieve3000 lead to higher probability of meeting the Phase-in 1 and Phase-in 2 passing standards on the STAAR EOC English I, English II, and Biology subject exams?
3) Is a higher mean Lexile score associated with the completion of more Achieve3000 exercises?

Data and Method

Data

For the 2014–2015 academic year, a total of 40 schools, at both the elementary and secondary levels, obtained licenses for the use of Achieve3000, making its unique services available to a total of 13,552 students in HISD. This study focused on a subset of those students, namely those in the ninth and tenth grades for whom there were valid pretest Lexile scores, as well as valid scores for the STAAR EOC English I, English II, and Biology exams. The final analytic sample consisted of 6,759 unique students nested within 27 high schools. Forty-three percent of these students (2,924) were in ninth grade during the 2014–2015 academic year, while the remaining 57 percent (3,835) were in tenth grade.

Measures

There were several dependent variables in this study. One was students' Lexile post-treatment score on Achieve3000's LevelSet™ test. Also of interest was student achievement on STAAR EOC subject exams, including the English I, the English II, and the biology tests. Student achievement here is defined as the scale score value earned on each of the respective tests and whether students met the passing standard of each test at the Phase-in 1 or Phase-in 2 levels. Disaggregating the sample by grade and the specific outcome being assessed resulted in a sample size of 1,631 students for Lexile scores (ninth and tenth grade), 2,924 students for STAAR EOC English I scores (ninth grade only), 3,835 students for STAAR EOC English II scores (tenth grade only), and 2,572 students for STAAR EOC Biology scores (ninth grade only). None of these samples is exclusive.

All analyses included controls for demographic characteristics such as sex, race/ethnicity, and economic disadvantage. Also controlled for was whether a student was classified as special education, gifted/talented, limited English proficient, or at-risk of dropping out. Recognizing that there may be differences between schools that may have impacts on student achievement above and beyond the use of Achieve3000 by students, this study also accounted for school-level proportion black and proportion of economically disadvantaged students.

Analytic Strategy

This study utilizes treatment effect analyses that account for unobserved differences among students. **Appendix A** provides a brief explanation of the specific method used in this study along with references to works that deliver more comprehensive treatment of the same.

For the purposes of this study, treatment is a four-category variable based on the number of Achieve3000 activities that students completed over the course of the academic year. Because the availability of Achieve3000 was limited to a few schools, and because student participation was voluntary, the valid comparison is between students who did not complete any activities and students who completed one or more exercises. The students completing zero Achieve3000 activities constituted the reference group. The students completing one or more exercises were divided into three groups: those completing between one and five activities, those completing between six and ten activities, and those completing more than ten activities.

The foregoing strategy was only used with respect to the STAAR EOC outcomes of interest. Since the Lexile score outcome necessarily excludes students who do not complete at least one Achieve3000 activity, the analysis of Achieve3000's relationship to the change from pre- to post-treatment Lexile scores was limited to linear regression modeling. Modeling accounted for

between-school differences in Achieve3000 usage and regression line slopes were allowed to be different for each school as a consequence of the number of activities completed. Again, Appendix A includes a more technical explanation of the method used with respect to the Lexile score outcome.

Results

Summary Statistics

Summary statistics of all control factors are presented by treatment group status in **Table 1** and **Table 2**. Table 1 shows means and proportions for each covariate for ninth graders. Table 2 shows the same for tenth graders. For comparison purposes, means and proportions for all covariates are also included for all ninth and tenth-grade students who were enrolled in HISD during the fall of the 2014–2015 academic year.

Comparing all HISD ninth graders with the total number of ninth graders in schools that had access to Achieve3000, gender and special education were the only predictors for which the distributions were about equal. The proportion of gifted/talented ninth graders in Achieve3000 schools was appreciably smaller than the proportion of gifted/talented ninth graders in all HISD high schools (8.5 percent versus about 12.0 percent). Similarly, while white students constituted about nine percent of all ninth-grade students in HISD, they constituted closer to two percent among those in Achieve3000 schools. Interestingly, black ninth graders were also underrepresented in Achieve3000 schools (21.1 percent) relative to their proportion among all ninth graders in the district (26.2 percent). A greater percentage of Hispanic ninth graders was in Achieve3000 schools (74.5 percent) than was in all HISD high schools (60.8 percent). For each of the remaining covariates, often negatively related with student achievement, the proportion of ninth-grade students in Achieve3000 was larger than the proportion of ninth-grade students in all HISD high schools.

The foregoing trends were mirrored among tenth-grade students, as shown in Table 2.

Turning to the means by treatment group status for each of the specific outcomes used in this study, it is apparent that Achieve3000 ninth-grade students who completed more rather than less activities performed better on the STAAR EOC English I exam (see **Figure 1**). The mean scale score among Achieve3000 ninth graders who completed more than ten activities was 107 and 87 points higher than the mean among ninth graders who completed, respectively, one to five and six to ten activities. These differences in means were statistically significant at the $p < .01$ level. Achieve3000 students who completed more than ten activities did not, however, perform significantly better than their peers who completed zero activities

Table 1. Summary statistics for all HISD 9th graders and for those at Achieve3000 schools by treatment status, 2014-2015.

	All HISD 9th Graders		Achieve3000 Schools Only									
			Completed Zero Activities		Completed between 1 and 5 Activities		Completed between 6 and 10 Activities		Completed More than 10 Activities		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
Total Enrollment	16,188		630		1,428		349		517		2,924	
Variables												
Gender												
Female	7,710	47.6	276	43.8	656	45.9	162	46.4	234	45.3	1,328	45.4
Male	8,478	52.4	354	56.2	772	54.1	187	53.6	283	54.7	1,596	54.6
Race/Ethnicity												
White	1,424	8.8	14	2.2	33	2.3	16	4.6	8	1.5	71	2.4
Black	4,240	26.2	95	15.1	230	16.1	101	28.9	191	36.9	617	21.1
Hispanic	9,839	60.8	501	79.5	1,144	80.1	224	64.2	310	60.0	2,179	74.5
Other Race	685	4.2	20	3.2	21	1.5	8	2.3	8	1.5	57	1.9
Special Education	1,516	9.4	51	8.1	157	11.0	29	8.3	49	9.5	286	9.8
Gifted/Talented	1,931	11.9	95	15.1	90	6.3	30	8.6	34	6.6	249	8.5
LEP	3,539	21.9	194	30.8	443	31.0	104	29.8	118	22.8	859	29.4
Economic Disadvantage												
Free/Reduced Lunch Eligible	6,293	38.9	281	44.6	664	46.5	144	41.3	210	40.6	1,299	44.4
In Poverty	5,476	33.8	245	38.9	545	38.2	154	44.1	236	45.6	1,180	40.4
At-Risk	11,163	69.0	464	73.7	1,151	80.6	269	77.1	402	77.8	2,286	78.2

Source: PEIMS 2014 Fall Snapshot.

Table 2. Summary statistics for all HISD 10th graders and for those at Achieve3000 schools by treatment status, 2014-2015.

	All HISD 10th Graders		Achieve3000 Schools Only									
			Completed Zero Activities		Completed between 1 and 5 Activities		Completed between 6 and 10 Activities		Completed More than 10 Activities		Total	
	N	%	N	%	N	%	N	%	N	%	N	%
Total Enrollment	12,906		741		1,593		719		782		3,835	
Variables												
Gender												
Female	6,448	50.0	366	49.4	754	47.3	364	50.6	406	51.9	1,890	49.3
Male	6,458	50.0	375	50.6	839	52.7	355	49.4	376	48.1	1,945	50.7
Race/Ethnicity												
White	1,404	10.9	10	1.3	26	1.6	12	1.7	26	3.3	74	1.9
Black	3,261	25.3	161	21.7	438	27.5	155	21.6	205	26.2	959	25.0
Hispanic	7,630	59.1	545	73.5	1,112	69.8	542	75.4	537	68.7	2,736	71.3
Other Race	611	4.7	25	3.4	17	1.1	10	1.4	14	1.8	66	1.7
Special Education	1,049	8.1	53	7.2	162	10.2	40	5.6	42	5.4	297	7.7
Gifted/Talented	2,227	17.3	106	14.3	157	9.9	106	14.7	158	20.2	527	13.7
LEP	2,229	17.3	176	23.8	343	21.5	157	21.8	127	16.2	803	20.9
Economic Disadvantage												
Free/Reduced Lunch Eligible	4,982	38.6	321	43.3	606	38.0	313	43.5	374	47.8	1,614	42.1
In Poverty	3,945	30.6	276	37.2	702	44.1	284	39.5	276	35.3	1,538	40.1
At-Risk	8,735	67.7	605	81.6	1,336	83.9	564	78.4	512	65.5	3,017	78.7

Source: PEIMS 2014 Fall Snapshot.

or all ninth graders in the district on the STAAR EOC English I exam. The English I mean scale score among all ninth-grade students was significantly greater than the English I mean among Achieve3000 ninth grade students who completed one to five activities ($p < .001$).

Figure 1 also shows mean STAAR EOC Biology scale scores by student group and treatment status for all ninth-grade students in HISD and for ninth-grade students in schools that had licenses for Achieve3000. The mean Biology scale score among all HISD ninth graders was not significantly different than the mean scale score among any of the two bottom Achieve3000 treatment groups. The mean biology scale score among Achieve3000 students who completed more than ten activities was significantly higher than that achieved by Achieve3000 students who completed one to five activities, by 70 points ($p < .05$). There were no other statistically significant mean group differences on the STAAR EOC Biology exam.

The mean STAAR EOC English II scale score, as shown in **Figure 2**, was 3757 among all tenth graders in the district. This mean was not significantly different than that among Achieve3000 students who completed zero (3749) or six to ten activities (3691). Achieve3000 students who completed one to five activities had a mean about 66 points lower than that experienced by all tenth graders in the district (statistically significant at the $p < .001$), while those who completed more than ten activities had a mean of 163 points higher (statistically significant at the $p < .001$) than that experienced by all tenth graders in the district. Among Achieve3000 students, those who completed more than ten activities performed significantly better than those who completed six to ten activities, who, in turn performed better than those who completed one to five activities.

Similar to the trends just highlighted with respect to the STAAR EOC exams, it was evident that more activities completed was associated with a greater likelihood of having met the passing standard on the STAAR EOC English I and English II exams, not accounting for differences across covariates. **Figure 3** shows the did-not-meet/did-meet standard rate on both the English I and English II exams at the Phase-in 1 standard. **Figure 4** shows the same at the Phase-in 2 standard. Relative to their Achieve3000 peers who completed ten or fewer activities, a greater proportion of ninth and tenth graders across the district met the Phase-in 1 and Phase-in 2 passing standard on the English I and English II exams. Contrastingly, the Achieve3000 students who completed greater than ten activities met the Phase-in 1 and Phase-in 2 passing standard at about the same rate or higher as all ninth and tenth graders in the district.

Finally, unadjusted post-treatment Lexile distributions by treatment status are shown in **Figure 5** for all Achieve3000 students who completed at least one activity and who also completed both the pre-treatment and a post-treatment Lexile test. Matching vertical lines are drawn at the mean of each distribution to highlight the fact that the completion of more activities is related to higher post-treatment Lexile scores.

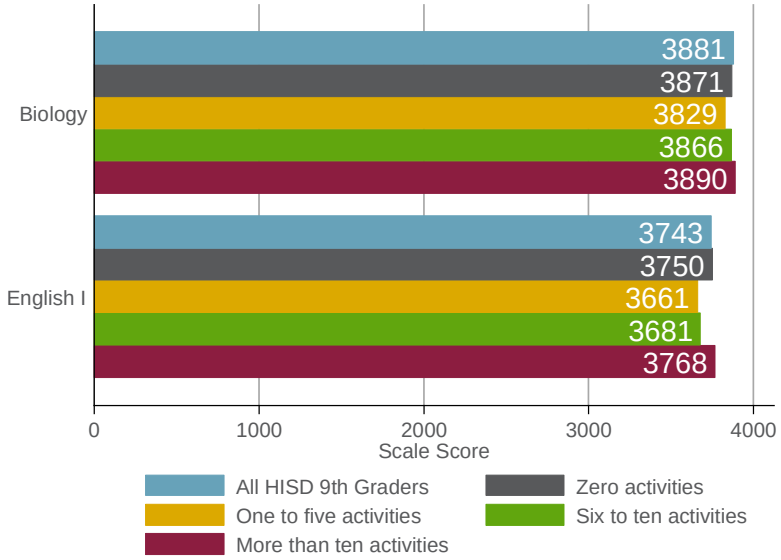

Figure 1. Mean 9th Grade STAAR EOC Scale Score by Treatment Status, Biology and English I.

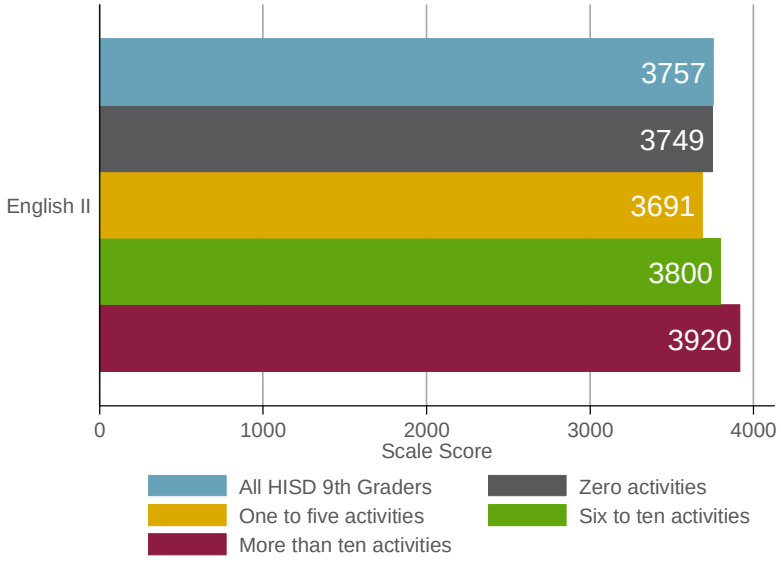

Figure 2. Mean 10th Grade STAAR EOC Scale Score by Treatment Status, English II.

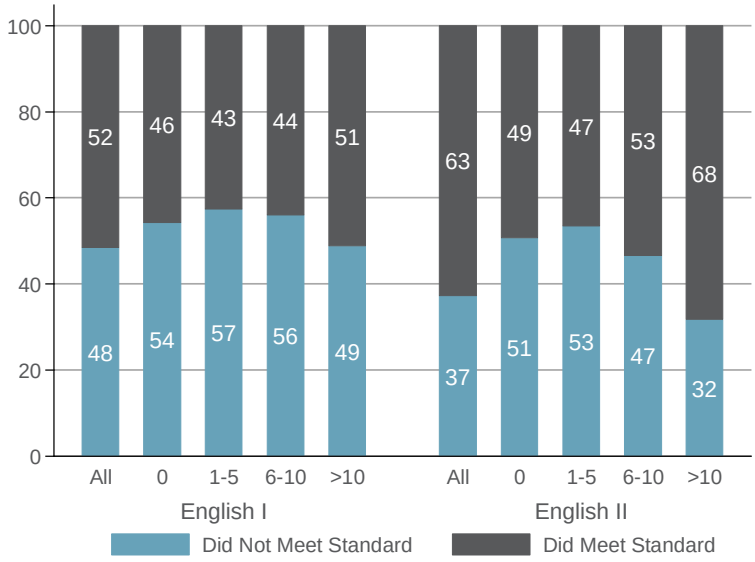

Figure 3. STAAR EOC Phase-in 1 Did-Not-Meet/Did-Meet Standard Rates by Treatment Status by Subject.

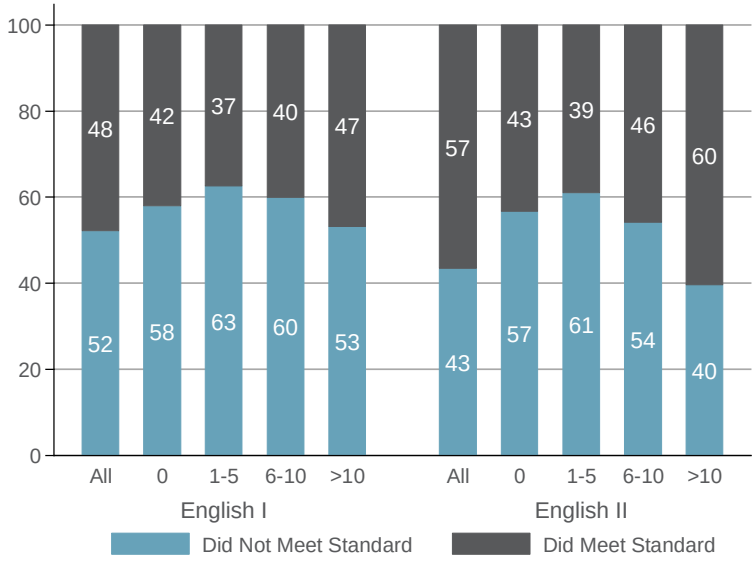

Figure 4. STAAR EOC Phase-in 2 Did-Not-Meet/Did-Meet Standard Rates by Treatment Status by Subject.

Treatment Effects of Achieve3000

What was the impact of Achieve3000 on student achievement after addressing selection issues and adjusting for student- and school-level controls? **Table B1** (page 11) presents, for each level of Achieve3000 usage, the average treatment effect on the scale scores for the STAAR EOC English I, English II, and biology exams. The bottom panel shows what the mean score would be on the respective exam had all students not used Achieve3000 web-based solutions. The second column reveals that ninth graders in Achieve3000 schools would have a mean of 3662 on the STAAR EOC English I exam had no students utilized Achieve3000. The differences in mean outcomes between non-usage of Achieve3000

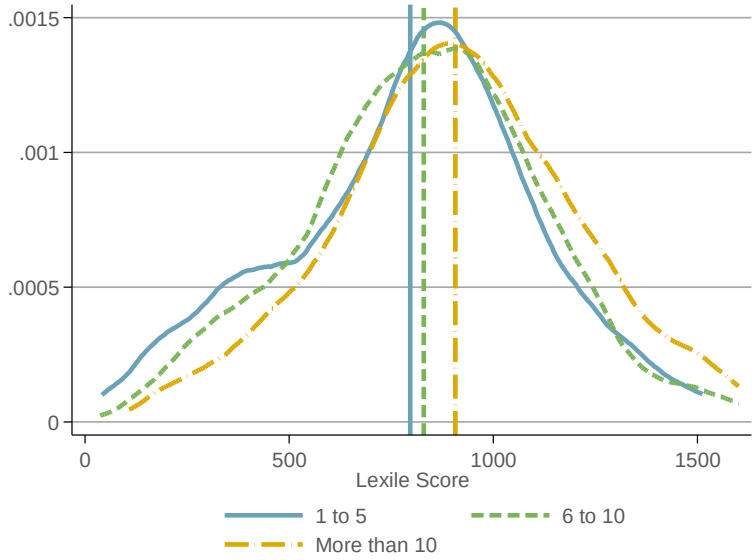

Figure 5. Post-treatment Lexile Score Distributions by Number of Activities Completed with Vertical Lines Drawn at the Distribution Means.

and both one to five activities completed and six to ten activities completed were not statistically different from zero. If all students had completed eleven or more Achieve3000 activities, the mean scale score on the English I exam would be 93 points higher than had they completed zero activities. This was statistically significant at the $p < .001$ level.

Among tenth graders taking the English II exam (column 3 of Table B1), the mean scale score would be 3758 had none of them completed any Achieve3000 activities. The mean would be similar had all students completed only between one and five or six and ten activities. However, the average treatment effect for completing greater than ten activities is an additional 89 scale score points. Since tenth graders taking the English II exam in 2014–2015 also took the English I exam in 2013–2014, the latter exam serves as another student-level control. Adjusting for previous years' scores, the average treatment effect deriving from the usage of Achieve3000 web-based solutions is somewhat attenuated, but the result is basically the same (column 4 of Table B1). Had all students with access to Achieve3000 completed more than ten activities, the average treatment effect would be an additional 44 scale score points over the potential outcome mean on the STAAR EOC English II exam for non-usage of Achieve3000 of 3759.

The final column of Table B1 shows the average treatment effect deriving from the usage of Achieve3000's web-based differentiated solutions on ninth graders' STAAR EOC biology exam. The bottom portion of the panel shows what the mean scale score would be (3801) had all students failed to complete at least one Achieve3000 activities. Had all students completed between one and five activities, the mean would be the same had none of them completed any activities. Had all students completed between six and ten, or more than ten, activities, the average treatment effect would be an additional 68 ($p < .01$) or 96 ($p < .001$) scale score points, respectively.

Table B2 and **Table B3** (pages 11–12) show, for each level of Achieve3000 usage, the probability of meeting the passing standard on the STAAR EOC English I, English II, and biology exams at the Phase-in 1 and Phase-in 2 level, respectively. Since the potential outcome means and average treatment effects are similar across the two Phase-in levels, the results of both tables are discussed in tandem.

First, had all ninth-grade students in Achieve3000 schools failed to complete any Achieve3000 activities, the mean probability of meeting the passing standard on the STAAR EOC English I exam (the second column of Tables B2 and B3) would be about 41 percent at the Phase-in 1 level and 36 percent at the Phase-in 2 level. This mean probability of meeting these passing standards would be the same had all students completed between one and five or between six and ten Achieve3000 activities. Had all students with access to Achieve3000 completed greater than ten activities, the mean probability of meeting the passing standard would be higher by about 10 percentage points. This average treatment effect was statistically significant at the $p < .001$.

Turning to the STAAR EOC English II exam (the third column in Tables B2 and B3), the potential mean probability of meeting the passing standard among tenth graders with access to Achieve3000 would be about 50 percent at the Phase-in 1 level and 44 percent at the Phase-in 2 level. Similar to the average treatment effect shown for the English I exam, had all students completed only between one and five or between six and ten Achieve3000 activities, the mean probability of meeting the passing standard would be the same, regardless of Phase-in level. Had all tenth graders with access to Achieve3000 completed more than ten activities, the mean probability of meeting the passing standard would be more than 13 percentage points higher at the Phase-in 1 level and more than 10 percentage points higher at the Phase-in 2 levels. Though somewhat attenuated, the same trend of treatment effects are achieved when tenth-grade students' English I scores from ninth grade are controlled (the fourth column of Tables B2 and B3). Had all tenth graders with access to Achieve3000 completed more than ten activities, the mean probability of meeting the passing standard would be more than 9 percentage points higher at both the Phase-in 1 and by more than almost 7 percentage points at the Phase-in 2 level.

Had all students failed to complete at least one Achieve3000 activity, their mean probability of meeting the passing standard for the STAAR EOC biology exam would be about 71 percent at the Phase-in 1 level and about 59 percent at the Phase-in 2 level (see the final columns of Tables B2 and B3). The average treatment effect increases significantly with the completion of more exercises. If all students with access to Achieve3000 had completed between one and five of the web-based activities, the mean probability of meeting the passing standard on the STAAR EOC biology exam would be 7 percentage points higher (significant at the $p < .01$ level) at both Phase-in levels than the mean probability had none of them completed a single activity. Had all students completed between six and ten activities, the mean probability of meeting the passing standard would be about 12 percentage points higher (significant at the $p < .001$ level) at both Phase-in levels than the mean probability had none of them completed a single activity. Had all students completed more than ten activities, the mean probability of meeting the passing standard would be about 17 percentage points higher ($p < .001$ level) at the Phase-in 1 level and about 21 percentage points higher ($p < .001$ level) at the Phase-in 2 level than the mean probability had none of them completed a single activity.

Finally, for those students who had pre-treatment Lexile scores on LevelSet and completed at least one Achieve3000 activity, how large were their collective gains on the post-treatment Lexile score? The estimates in **Table B4** (page 12) answer this question. The second column presents estimates of a restricted model for comparison purposes. The main focus here will deal with the estimates obtained in the full model. Controlling for all student- and school-level factors mentioned in the previous section, and accounting for the initial Lexile score, the school mean Lexile score was just under 900 points (with a standard deviation of about 20 points). The average gain per 10 Achieve3000 activities completed above the mean number of activities completed was about 16 points (with a standard deviation of 7 points). The estimated correlation of .77 between the school mean score and the number of activities completed (lower panel and final column of Table B4) can be interpreted to say that, net of covariates, schools with larger mean Lexile scores for students with an average number of activities completed than other schools tend to have larger gains in Lexile scores than those other schools as well.

Although it does not address potential issues of self-selection into greater rather than less usage of Achieve3000, this association is quite large and can be argued to show, particularly in light of the other findings shown here, that Achieve3000 does indeed lead to improved reading ability and comprehension for students in need of such improvement.

Discussion and Recommendations

The aim of this study was to assess the effectiveness of Achieve3000, a differentiated web-based learning tool, to increase student reading ability and comprehension. Using quasi-experimental analytic techniques to address selection bias on the use of Achieve3000 by those students for whom it was available, the findings here suggest that the completion of increasing number of Achieve3000 activities yields higher levels of achievement on the STAAR EOC English I, English II, and Biology exams. Not only does the use of Achieve3000 lead to greater average treatment effects for scale scores on these exams, but it also leads to higher mean probabilities of meeting the passing standards on the same exams, both at the Phase-in 1 and Phase-in 2 levels. Though it is difficult to make the same causal claims about the impact of Achieve3000 on post-treatment Lexile scores, accounting for pre-treatment Lexile score and other controls, the improvement in performance is appreciable enough to argue, particularly in light

of the findings for the other outcomes examined here, that Achieve3000 has some positive impact.

Based on these findings, it is recommended that teachers in schools that have Achieve3000 licenses actively encourage their students to complete as many of the exercises as possible during the academic school year. Fidelity to such a recommendation under such a highly decentralized system will no doubt be difficult to achieve. It may therefore be important to allow for the incorporation of Achieve3000 reading solutions into normal classroom time or offer incentives to students who complete a greater number of exercises.

References

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* New York: Cambridge University Press.

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., & Pignone, M. P. (2004). Literacy and health outcomes. *Journal of General Internal Medicine*, *19*, 1228-1239.

DeWalt, D. A., & Hink, A. (2009). Health literacy and child health outcomes: A systematic review of the literature. *Pediatrics*, *124*, S265-S274.

Greenberg, D., Gilbert, A. & Fredrick, L. (2006). Reading interest and behavior in middle school students in inner-city and rural settings. *Reading Horizons*, *47*, 159-173.

National Center for Education Statistics. (2013). *A first look: 2013 mathematics and reading*. Retrieved from http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014451.pdf.

Pitcher, S. M., Albright, L. K., DeLaney, C. J., Walker, N. T., Seunarinesingh, K., Mogge, S., …Dunston, P. J. (2007). Assessing adolescents' motivation to read. *Journal of Adolescent & Adult Literacy*, *50*, 378-396.

Smith, M. C. (2009). Literacy in adulthood. In M. C. Smith (Ed.), *Handbook of research on adult learning and development* (pp. 601-635). New York: Routledge.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*, 1281-1301.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, MA: MIT Press.

For additional information contact the HISD Department of Research and Accountability at 713-556-6700 or e-mail Research@Houstonisd.org.

Appendix A

This appendix provides information on the analytic strategy used in this study.

Analytic Strategy

Treatment of STAAR EOC Scale Score Outcomes

Because neither the implementation of Achieve3000 at a specific school nor its use by particular students within those schools are random processes, statistical analyses examining its relationship to specific outcomes must address the non-representative nature deriving from these non-random processes. While simple regression techniques may reveal statistically significant associations between the use of Achieve3000 and students' achievement, such associations may not be viewed as causal since there may be unobserved differences among students that drives them to both do better academically, generally, and to also complete more Achieve3000 activities. To be able to say that Achieve3000 produced specific results and was not merely associated with them, the counterfactual model of causal inference requires the use of statistical methods that remove bias. Failure to account for potential omitted variables or the bias in selecting greater use of Achieve3000 among students can lead to erroneous conclusions about the causal link between students' use of Achieve3000 and their later academic performance. Specifically, given that Achieve3000 is entirely voluntary, the lesser or greater usage of its modules by students, to the extent that such usage may be attributed to students' family background or other observables, must be addressed if unbiased estimates of its effects are to be accurately achieved.

To approximate the results that might be obtained if the district had implemented Achieve3000 via randomization, and therefore to make causal claims, this study utilized a form of regression adjustment that is weighted by the inverse of the probability of receiving the treatment received (see Cameron & Trivedi [2005] and Wooldridge [2007, 2010] for a comprehensive overview of the inverse-probability-weighted regression adjustment estimator). Inverse-probability-weighted regression adjustment (IPWRA) estimators first use a model to predict treatment status for each subject in an observational study. This value, the propensity score, is then used to create subject-specific weights equal to the inverse of the probability of receiving the treatment actually received. Treatment-specific predicted outcomes are then predicted by fitting weighted regression models. The difference of the means between treatment groups provides estimates of the average treatment effect (ATE).[1] Robust standard errors were estimated to account for between-school variability.

Treatment of Lexile Score Outcome

The analysis of Achieve3000's relationship to the change from pre- to post-treatment Lexile scores was limited to random coefficient modeling, a form of hierarchical linear modeling that allows the effects of covariates to vary between clusters, or, in this case, schools. The main predictor for this analysis, the total number of Achieve3000 activities completed, was centered at the grand mean and then divided by ten so that the estimate of its relationship to the outcome can be interpreted as the change in Lexile score per ten additional activities completed above the average number of activities completed by all Achieve3000 students. School-specific regression lines were not assumed to be parallel, so this study introduced random slopes for the main predictor. The same controls used in the treatment effects analyses were also used in the random coefficient models.

[1] One could also restrict the computations of the means of the treatment-specific predicted outcomes to focus only on treated subjects, deriving the average treatment effect on the treated (ATET).

Appendix B

Table B1. Treatment level predicted scale score outcomes deriving from the inverse-probability-weighted regression-adjusted estimator.

	English I	English II		Biology
			Controlling for English I Scores	
ATE				
Treatment				
(1 to 5 vs. 0)	5.69	-25.66	-3.54	33.06
	(19.54)	(15.08)	(11.89)	(20.49)
(6 to 10 vs. 0)	11.17	24.96	20.48	68.43**
	(24.19)	(17.61)	(13.59)	(24.64)
(More than 10 vs. 0)	93.35***	89.36***	44.40**	95.81***
	(25.33)	(18.62)	(13.69)	(24.79)
POMean				
Treatment				
0	3661.64***	3758.33***	3758.98***	3801.08***
	(18.03)	(13.82)	(11.51)	(17.48)

Note: All model coefficients are net of all controls listed in the Data and Method section of this research brief. Robust standard errors are in parenthesis.
*p < .05, **p < .01, ***p < .001; two-tailed tests.

Table B2. Treatment level predicted probabilities of meeting the Phase-in 1 passing standard deriving from the inverse-probability-weighted regression-adjusted estimator.

	English I	English II		Biology
			Controlling for English I Met Standard	
ATE				
Treatment				
(1 to 5 vs. 0)	.020	.012	.008	.068**
	(.021)	(.019)	(.017)	(.022)
(6 to 10 vs. 0)	.030	.018	.005	.117***
	(.030)	(.022)	(.020)	(.029)
(More than 10 vs. 0)	.109***	.126***	.093***	.167***
	(.032)	(.023)	(.021)	(.026)
POMean				
Treatment				
0	.405***	.495***	.507***	.708***
	(.019)	(.016)	(.017)	(.020)

Note: All model coefficients are net of all controls listed in the Data and Method section of this research brief. Robust standard errors are in parenthesis.
*p < .05, **p < .01, ***p < .001; two-tailed tests.

Table B3. Treatment level predicted probabilities of meeting the Phase-in 2 passing standard deriving from the inverse-probability-weighted regression-adjusted estimator.

	English I	English II		Biology
			Controlling for English I Met Standard	
ATE				
Treatment				
(1 to 5 vs. 0)	.013	-.005	.005	.072**
	(.021)	(.018)	(.017)	(.023)
(6 to 10 vs. 0)	.031	.001	.008	.118***
	(.029)	(.021)	(.020)	(.032)
(More than 10 vs. 0)	.093***	.098***	.068**	.208***
	(.028)	(.023)	(.022)	(.029)
POMean				
Treatment				
0	.364***	.438***	.440***	.589***
	(.019)	(.015)	(.016)	.020

Note: All model coefficients are net of all controls listed in the Data and Method section of this research brief. Robust standard errors are in parenthesis.
*p < .05, **p < .01, ***p < .001; two-tailed tests.

Table B4. Random coefficient models of Lexile score change due to Achieve3000 usage.

Variables	Restricted Model	Full Model
Fixed Effects		
Intercept	842.63***	892.48***
	(33.06)	(34.13)
# Activities Completed	19.16	15.78***
	(16.95)	(3.08)
Random Effects		
SD of the Intercept	161.58***	19.59***
	(24.26)	(3.79)
SD of # Activities Completed	52.21***	7.05**
	(21.92)	(3.48)
Correlation between Intercept and # Activities Completed	.46	.77
SD of the Residuals	240.50***	63.23***
	(4.29)	(1.12)
Log-likelihood	-11303.35	-9101.62

Note: The restricted model coefficients are net of only the number of activities covariate, whose effect was also allowed to vary by school. The full model coefficients are net of all controls listed in the Data and Method section of this research brief, including the pre-treatment Lexile score. Again, the effect of the number of activities completed was allowed to vary by school. Maximum likelihood estimation was used to obtain the estimates. Standard errors are in parenthesis.
*p < .05, **p < .01, ***p < .001; two-tailed tests.