MEMORANDUM May 16, 2017

TO: Mechiel Rozas, Secondary Director of ESL & ESOL

FROM: Carla Stevens
Assistant Superintendent, Research and Accountability

SUBJECT: **ACHIEVE3000®, 2015-2016**

The 2015–2016 academic year was the third year that the Houston Independent School District offered the differentiated reading instruction program, Achieve3000®, to its 9th and 10th grade students. While there were 22,175 9th graders and 16,160 10th graders nested in the district's 9th- and/or 10th-grade serving schools, only 1,660 9th graders and 1,265 10th graders (1) completed at least one activity during the academic year, (2) had an average first-try assessment score on the post-reading LevelSet™ test of 75 percent or higher, and (3) had a STAAR reading or English I score in the prior grade. The attached evaluation compares 2015–2016 STAAR outcomes between these students and their same-grade peers in both the same schools and across schools in the district. It employs analytic methods that control for bias in students' self-selection into the use of Achieve3000.

Key findings include:

- Taking into account student- and school-level factors, the average 9th grade student who utilized the Achieve3000 tool achieved 50 scale score points more on the STAAR EOC English I assessment than did his same-grade peer who did not or could not make use of Achieve3000 (i.e., 3982 vs. 3932). This difference **IS ALMOST EQUIVALENT** to the Achieve3000 group having a mean scale score at the Level II Recommended standard, the standard above the minimum Level II Satisfactory standard.
- Taking into account student- and school-level factors, the average 10th grade student who utilized the Achieve3000 tool achieved 69 scale score points more on the STAAR EOC English II assessment than did his same-grade peer who did not or could not make use of Achieve3000 (i.e., 4007 vs. 3938). This difference **IS EQUIVALENT** to the Achieve3000 group having a mean scale score at the Level II Recommended standard, the standard above the minimum Level II Satisfactory standard.

Further distribution of this report is at your discretion. Should you have any further questions, please contact me at 713-556-6700.

CJS

Attachment

cc: Grenita Lathan
Sam Sarabia
Annie Wolfe

HISD | Research and Accountability
ANALYZING DATA, MEASURING PERFORMANCE.

EVALUATION REPORT
BUREAU OF PROGRAM EVALUATION
Volume 10, Issue 1, December 2016

Achieve3000® impacts on students' STAAR EOC English I and English II performance for the 2015–2016 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 cohorts of 9th and 10th graders and assesses the impact deriving from the use of Achieve3000's reading solutions on students' scale score performance on the State of Texas Assessments of Academic Readiness End-Of-Course (STAAR EOC) for the English I and English II exams. Utilizing the inverse-probability-weighted regression adjusted (IPWRA) estimation method to account for the non-randomness of student self-selection into Achieve3000 usage, the results of this evaluation may be viewed as causal. Findings suggest that students' use of Achieve3000's web-based software leads to higher performance on the STAAR EOC exams than non-use of Achieve3000. Controlling for student- and school-level factors, the treatment effect among 9th-grade Achieve3000 students was 50 scale score points higher than the potential outcome mean among their non-Achieve3000 peers on the English I assessment. The corresponding gain for 10th-grade Achieve3000 students over their non-Achieve3000 peers was 69 scale score points on the English II assessment. Based on these findings, it is recommended that teachers in schools that have Achieve3000 licenses actively encourage the completion of Achieve3000 exercises.

Background

Student proficiency in reading literacy remains a perennial concern of educators and education policymakers in the United States. Despite research findings, however, that even students from the most suboptimal backgrounds can be taught to read well (Snow, Burns, and Griffin, 1998), a significant proportion of the nation's children leave school without the literacy skills needed to be successful in a postindustrial society (Aud, Wilkinson-Flicker, Kristopovich, Rathbun, Wang, and Zhang, 2013). Indeed, the most recent results highlighting U.S. children's performance in reading revealed that only 35 percent of fourth graders, 36 percent of eighth graders, and 38 percent of 12th graders performed at or above proficient level (Snyder, de Brey, & Dillow, 2016).

Because classroom instruction tends to focus on materials that are grade-level specific, struggling readers do not benefit as much as their non-struggling peers. As a result, the goal of improving reading proficiency among elementary and secondary students through means other than formal education in the classroom has proliferated in recent years. In particular, school districts around the country are actively engaged in getting students to read outside of school hours, which, even net of school effects, is associated with better performance on exams and greater odds of obtaining a university degree (Cheung and Andersen, 2003). More specifically, a number of districts, including the Houston Independent School District (HISD), are utilizing self-directed reading tools that employ differentiated instruction that takes into account students' different abilities.

For the past two years, HISD has targeted raising the literacy skills of high school freshmen and sophomores using Achieve3000's LevelSet™ assessment. Rooted in the Lexile Framework,

Achieve3000 assessments measure the difficulty of texts and students' reading ability on the same scale to accurately match readers to appropriate texts (Achieve3000, 2016). As students use the program, they are challenged to reach ever-higher levels of reading and writing ability. The rationale behind this approach is the idea that, because students have individual backgrounds and needs, their reading instruction should be differentiated. A one-size-fits-all approach, as is taught in the traditional classroom, would necessarily leave behind the most vulnerable students, exacerbating present trends that show general poor reading proficiency of U.S. children. The aim of universal compulsory education, though, is to ensure that all students, regardless of their different origins and experiences, are prepared for the demands of a postindustrial knowledge economy.

Previous district and peer-reviewed evaluations have revealed appreciable Lexile and literacy gains across the elementary and secondary years (Borman, Park, and Min, 2015; Reeves, 2014; Torres, 2015). The largest gains, according to Achieve3000 (2015), were experienced by students who (1) completed at least two activities per week, (2) scored 75 percent or higher on the post-reading exercise assessments, (3) scored two or more years below grade level on the initial LevelSet, and (4) were English Language Learners (ELL).

Research Questions

The present evaluation utilizes the 2015–2016 administrative student-level data from HISD to assess the efficacy of 9th and 10th grade students' use of Achieve3000 differentiated reading to improve their reading ability relative to a comparison group of their peers. Specifically,

1. Is students' use of Achieve3000 associated with, net of previous years' measured reading ability and a host of student- and school-level factors, higher scores on the State of Texas Assessments of Academic Readiness End-of-Course (STAAR EOC) English I (for 9th graders) and English II (for 10th graders) exams?

The STAAR EOC English assessments are suitable for the present study because they test students' reading and vocabulary development, as well as comprehension of literary and informational texts, all of which are indicative of increased literacy, and which exhibit a proficiency in abilities needed to be successful in the current and future economy.

Data and Method

Sample

For the 2015–2016 academic year, Achieve3000 was implemented in 38 of 54 HISD schools serving 9th or 10th grade students. While there were 22,175 9th graders and 16,160 10th graders nested in the district's 9th and/or 10th grade serving schools, the corresponding number of same-grade students in Achieve3000 schools was 20,327 and 15,146, respectively. Of these, only those who (1) completed at least one activity during the academic year, (2) had an average first-try assessment score on the post-reading LevelSet test of 75 percent or higher, and (3) had a STAAR reading score in 8th grade (for the 9th-grade sample being assessed on the STAAR EOC English I exam) or a STAAR EOC English I score in 9th grade (for the 10th-grade sample being assessed on the STAAR EOC English II exam), constituted the treatment group.

Ninth and 10th grade students in Achieve3000 schools who failed to meet these criteria, as well as all 9th and 10th grade students from the district's other schools serving 9th and 10th graders, constituted the potential control group. This group was similarly restricted to those who had STAAR assessment scores in both 2014–2015 and 2015–2016. The final analytic sample, then, consisted of 1,660 9th-grade treatment group students nested in 36 high schools, 8,683 9th-grade control group students nested in 44 schools, 1,263 10th-grade treatment group students nested in 35 schools, and 8,952 10th-grade control group students nested in 44 schools.

Measures

Dependent Variables. This evaluation used students' scale scores on any version of the STAAR EOC English I and English II exams, administered in the spring of 2016, as the main outcome variables. While the majority of students in the two samples took the standard STAAR test, 266 9th-graders and 142 10th-graders took the STAAR L (Linguistic Accommodations for English Language Learners) and 353 9th graders and 336 10th-graders took the STAAR A (Accommodations for Students with Disabilities or Accommodations in Unexpected or Emergency Situations).

Independent Variables. Because the relationship between treatment (completion of at least one Achieve3000 activity during the academic year with attendant average first-try reading comprehension score of 75 percent or higher on the LevelSet post-reading exercise assessments) and STAAR EOC English assessment outcomes may depend on students' previous reading ability or other

demographic or school factors, this evaluation controlled for these factors. Specifically, the 2015–2016 9th-grade sample controlled for students' 2014–2015 8th-grade STAAR reading scale scores, and the 2015–2016 10th-grade sample controlled for students' 2014–2015 9th-grade STAAR EOC English I scale scores.

Additional student-level controls were entered for race/ethnicity, gender, current age, economic disadvantage, at-risk for dropout, English Language Learner (ELL) status, special education status, magnet student status, gifted/talented student status, and homeless student status (for more information on how these variables were operationalized, see **Appendix A**).

Finally, controls were also included to account for variation at the school level. Among these variables was whether a student was in a school (1) where the percent black exceeded the district high school mean, (2) where the percent economically disadvantaged exceeded the district high school mean, (3) where the percent at-risk exceeded the district high school mean, (4) that received the top 25 percent closing performance gap distinction, (5) that received the top 25 percent student progress distinction, (6) that received campus postsecondary readiness distinction, (7) that received campus reading/English Language Arts (ELA) academic achievement distinction, and (8) that received an accountability rating of "Improvement Required." A continuous control for the school mobility rate was also included.

Analytic Strategy

Causal Inference. To test the causal effects of Achieve3000 usage, the inverse-probability-weighted regression-adjusted (IPWRA) estimates were obtained (Imbens and Wooldridge, 2009). IPWRA is a doubly robust quasi-experimental statistical method that seeks to overcome what Holland (1986) stated is the "fundamental problem of causal inference," namely, that, since one can observe at most a single outcome given the exposure to a specific treatment or level of intervention, it is impossible to directly observe causal effects. In a three-step process, Achieve3000 usage was regressed on the student- and school-level factors listed above to obtain the propensity of being treated. The inverse of these propensity scores was then calculated. Finally, the inverse probabilities were used as weights in a model that regressed STAAR EOC English achievement on treatment and student- and school-level factors (Appendix A provides a more comprehensive explanation of the analytic strategy used here).

Results

Descriptive Statistics. Before examining the main analyses of this evaluation, summary statistics of the control and treatment groups were observed separately for the 9th and 10th grade samples. **Table 1** (page 4) and **Table 2** (page 5) show the means and standard deviations for student-level demographic characteristics, school-level factors, and reading/English-language ability in the previous year. The pre-test scores of 9th graders on their 8th grade STAAR Reading assessment were higher for Achieve3000 students—i.e., those in the treatment group—than for those in the comparison group (Table 1). The difference was statistically significant. While the pre-test scores of 10th graders on their 9th grade STAAR EOC English I tests were also higher those in the treatment group relative to their peers in the control group, this difference was statistically nonsignificant (Table 2).

Notably, and regardless of the sample, fewer whites and Asians/Pacific Islanders were Achieve3000 students than non-Achieve3000 students, and blacks and Hispanics were slightly overrepresented in the treatment group, relative to both their control group percentages and their percentages among all control and treatment group students in the sample. In addition, a larger percentage of Achieve3000 students were classified as economically disadvantaged. For the 9th-grade sample, larger percentages of Achieve3000 students were either magnet or gifted/talented students, while smaller percentages of Achieve3000 students were, for the 9th-grade sample, at-risk, limited English proficient (LEP), or special education, or, for the 10th-grade sample, homeless students.

Trends varied across both samples with respect to school-level variables. For the 9th-grade sample, smaller percentages of Achieve3000 students than control students attended a school where the percent of black and at-risk students exceeded the district high school mean. The differences were statistically significant. Greater percentages of Achieve3000 students than control group students attended schools that received the top 25 percent closing performance gaps distinction, the campus postsecondary readiness distinction, and the campus reading/English Language Arts (ELA) academic achievement distinction. This suggests that the average school attended by those engaging and excelling at Achieve3000 activities has scores that were above the district average on the key state accountability indicators than the average school attended by those students not engaging and excelling at Achieve3000 activities. The majority

Table 1. Summary Statistics by Treatment Group Status for the 2015-2016 9th Grade Sample.

	Total N = 10,354		Control N = 8,694		Treatment N = 1,660		*Mean Diff. & Sig.*		
	Mean	SD	Mean	SD	Mean	SD		*t*	χ^2
Student-Level Variables									
Pre-test Score on STAAR 8th Reading	1650.0	144.3	1641.3	146.5	1695.9	122.5	-54.6	***	
Race/Ethnicity (%)									
White	7.7	26.7	8.1	27.3	5.8	23.3	-2.3		**
Black	23.9	42.7	23.9	42.6	24.0	42.7	0.2		
Hispanic	63.4	48.2	62.8	48.3	66.5	47.2	3.7		**
Asian/Pacific Islander	4.0	19.5	4.2	20.0	2.8	16.4	-1.4		**
Other Race	1.0	9.8	1.0	9.9	0.9	9.5	-0.1		
Gender (%)									
Female	49.5	50.0	49.3	50.0	50.7	50.0	1.4		
Male	50.5	50.0	50.7	50.0	49.3	50.0	-1.4		
Current Age (in Years)	15.1	0.6	15.1	0.6	15.0	0.5	-0.1	***	
Economically Disadvantaged (%)	78.4	41.2	77.9	41.5	81.0	39.2	3.1		**
At-Risk (%)	42.5	49.4	44.6	49.7	31.8	46.6	-12.8		***
LEP (%)	13.6	34.3	15.1	35.8	5.9	23.6	-9.2		***
Special Education Student (%)	5.9	23.5	6.5	24.7	2.3	15.0	-4.3		***
Magnet Student (%)	24.9	43.2	22.8	42.0	35.4	47.8	12.5		***
Gifted/Talented Student (%)	20.3	40.2	19.6	39.7	24.0	42.7	4.5		***
Homeless Student (%)	10.3	30.4	10.4	30.5	9.8	29.7	-0.6		
School-Level Variables									
% Black Exceeds District HS Mean (%)	36.7	48.2	39.1	48.8	24.6	43.1	-14.5		***
% Economically Disadvantaged Exceeds District HS Mean (%)	61.1	48.7	61.1	48.8	61.3	48.7	0.2		
% At-Risk Exceeds District HS Mean (%)	57.4	49.5	60.0	49.0	43.9	49.6	-16.1		***
Top 25% Closing Performance Gaps Distinction (%)	20.1	40.1	18.2	38.6	30.1	45.9	11.9		***
Top 25% Student Progress Distinction (%)	27.7	44.7	28.0	44.9	25.8	43.8	-2.3		
Campus Postsecondary Readiness Distinction (%)	41.0	49.2	40.1	49.0	45.9	49.8	5.8		***
Campus Reading/ELA Academic Achievement Distinction (%)	34.0	47.4	32.4	46.8	42.2	49.4	9.7		***
Campus Mobility Rate (%)	18.0	11.3	18.6	11.5	15.2	9.7	-3.3	***	
School Rated "Improvement Required" (%)	14.4	35.1	14.6	35.3	13.6	34.2	-1.0		

Note: The statistical significance for all treatment-control group differences are evaluated using a *t* test for continuous measures and a χ^2 for binary measures.

of these differences were also statistically significant.

Conversely, for the 10th-grade sample, smaller percentages of control students than Achieve3000 students attended a school where the percent of black, economically disadvantaged, and at-risk students exceeded the district high school mean. Control group students also constituted a smaller percentage of 10th-grade students who attended a school with an accountability rating of "Improvement Required." Greater percentages of control group students than Achieve3000 students attended schools that received the top 25 percent closing performance gaps distinction, the campus postsecondary readiness distinction, and the campus reading/English Language Arts (ELA) academic achievement distinction. This suggests that the average school attended by those not engaging Achieve3000 activities has scores that were above the district average on the key state accountability indicators than the average school attended by those students engaging and excelling at Achieve3000 activities. All differences were statistically significant at the $p < .001$ level.

Figure 1 (page 6) shows the control and treatment group means for the 9th-grade STAAR EOC English I and 10th-grade STAAR EOC English II outcomes. Achieve3000 9th graders outperformed their non-Achieve3000 peers. The mean difference, significant at the $p < .001$ level, was greater than 200 scale score points on the STAAR EOC English I assessment. Similarly, Achieve3000 10th graders

Table 2. Summary Statistics by Treatment Group Status for the 2015-2016 10th Grade Sample.

	Total N = 10,214		Control N = 8,949		Treatment N = 1,265		*Mean Diff. & Sig.*		
	Mean	SD	Mean	SD	Mean	SD		*t*	χ^2
Student-Level Variables									
Pre-test Score on STAAR EOC English I	3952.6	601.6	3951.4	618.8	3963.0	461.8	-11.6		
Race/Ethnicity (%)									
White	8.9	28.4	9.5	29.3	4.4	20.6	-5.1		***
Black	24.2	42.8	23.8	42.6	27.1	44.5	3.3		**
Hispanic	61.7	48.6	61.1	48.7	65.5	47.6	4.4		**
Asian/Pacific Islander	4.3	20.4	4.6	21.0	2.2	14.7	-2.4		***
Other Race	0.9	9.6	1.0	9.8	0.8	8.9	-0.2		
Gender (%)									
Female	50.6	50.0	50.9	50.0	48.7	50.0	-2.2		
Male	49.4	50.0	49.1	50.0	51.3	50.0	2.2		
Current Age	16.1	0.7	16.1	0.7	16.0	0.6	-0.1	***	
Economically Disadvantaged (%)	73.8	44.0	72.3	44.7	84.3	36.4	12.0		***
At-Risk (%)	42.0	49.4	42.4	49.4	39.4	48.9	-3.0		*
LEP (%)	9.9	29.9	10.2	30.3	7.6	26.5	-2.6		**
Special Education Student (%)	6.3	24.3	6.7	25.1	3.1	17.3	-3.6		***
Magnet Student (%)	33.9	47.3	34.1	47.4	32.5	46.9	-1.5		
Gifted/Talented Student (%)	17.4	37.9	17.7	38.1	15.5	36.2	-2.1		
Homeless Student (%)	8.8	28.3	8.4	27.7	11.7	32.2	3.4		***
School-Level Variables									
% Black Exceeds District HS Mean (%)	35.1	47.7	34.3	47.5	40.7	49.1	6.4		***
% Economically Disadvantaged Exceeds District HS Mean (%)	59.5	49.1	58.0	49.4	70.7	45.5	12.7		***
% At-Risk Exceeds District HS Mean (%)	54.8	49.8	53.2	49.9	66.0	47.4	12.7		***
Top 25% Closing Performance Gaps Distinction (%)	22.8	42.0	23.4	42.3	18.8	39.1	-4.6		***
Top 25% Student Progress Distinction (%)	30.7	46.1	31.5	46.5	24.8	43.2	-6.7		***
Campus Postsecondary Readiness Distinction (%)	44.6	49.7	47.4	49.9	24.9	43.2	-22.5		***
Campus Reading/ELA Academic Achievement Distinction (%)	37.0	48.3	39.5	48.9	19.2	39.4	-20.4		***
Campus Mobility Rate (%)	17.0	11.2	16.7	11.3	19.3	10.5	2.7	***	
School Rated "Improvement Required" (%)	12.6	33.1	11.9	32.4	17.3	37.9	5.5		***

Note: The statistical significance for all treatment-control group differences are evaluated using a *t* test for continuous measures and a χ^2 for binary measures.

did better than their control group peers by 70 scale score points. This difference was statistically significant at the $p < .001$ level. Despite these apparent appreciable gains in English and reading ability due to Achieve3000 usage, they do not take into account either selection bias or demographic and other control factors. To assess the causal impact of treatment group status, net of other factors, IPWRA estimates are examined.

Achieve3000 Analysis. The main analyses of this evaluation compare Achieve3000 students to non-Achieve3000 students. For both the 9th-grade STAAR EOC English I and 10th-grade STAAR EOC English II post-test outcomes, three models were estimated. Model 1 was unconditional and only included the treatment indicator as a predictor. Model 2 included the treatment indicator while controlling for student-level demographic factors. Model 3, building on Model 2, further conditioned on school-level factors.

Table B1 (see Appendix B, page 11) shows the IPWRA estimates of the effect of Achieve3000 on 9th-grade students' performance on the 2015-2016 STAAR EOC English I assessment. The average treatment effect (ATE) was about 217 scale score points, an effect that was statistically significant. After controlling for student-level demographic factors, the ATE was attenuated slightly, though the advantage of Achieve3000 usage was still appreciable, at about 63 scale score points, and statistically significant. Further controlling for school-level factors, the potential outcome mean

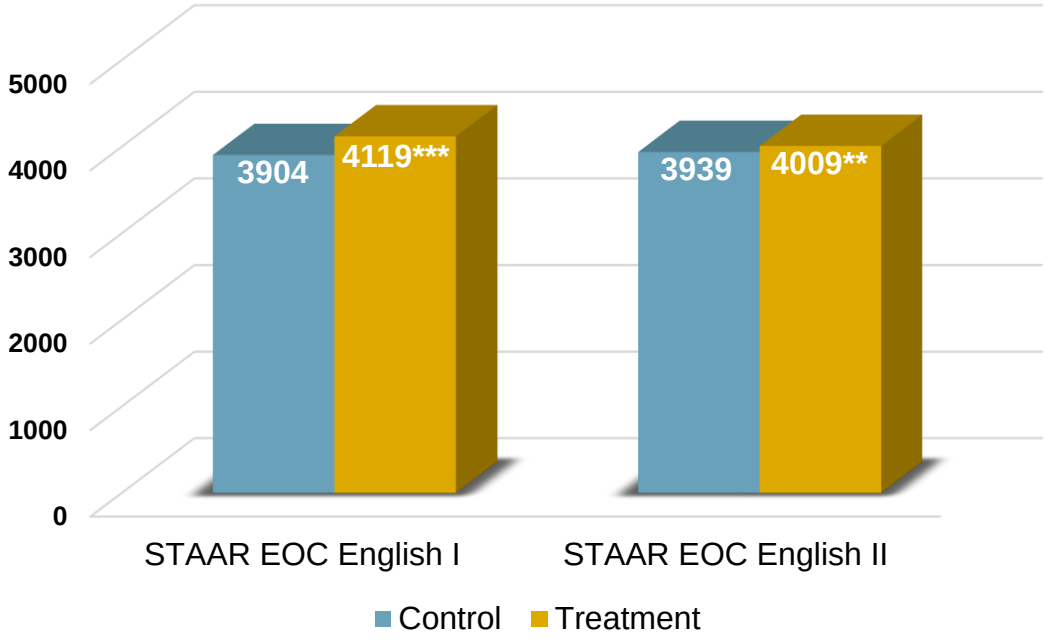

Figure 1. Post-Test Score Comparisons between Control and Treatment Groups on the STAAR 2015–2016 English I (9th Graders) and English II (10th Graders) Assessments.

scale score was about 3932 for the control group and the ATE was 50 scale score points higher than that.

Table B2 (see Appendix B, page 11) shows the IPWRA estimates of the effect of treatment group status on 10th grade students' 2015-2016 STAAR EOC English II assessment. Similar to the results shown in Table 3, use of Achieve3000 led to a nearly 73 scale score point advantage over non-use of Achieve3000 before controlling for any other factors. This difference was statistically at the $p <$ .001 level. Net of student-level demographic factors, the ATE was reduced slightly to about 68 scale score points, which was still appreciable as well as statistically significant. When both student- and school-level factors were controlled for, the potential outcome mean among control group students was 3938 scale score points while the ATE was about 69 scale score points higher than that.

Discussion

This evaluation sought to understand how students' use of Achieve3000 reading solutions improved their literacy, which includes a host of skills such as literary analysis, comprehension of informational and expository texts, the writing process and writing conventions, and the rules of syntax and grammar. Utilizing an analytic method that addressed the selection bias inherent in a completely voluntary self-directed differentiated improvement instrument such as Achieve3000's LevelSet, the results revealed statistically significant gains in students' performance on the STAAR EOC English tests. Specifically, HISD 9th graders who used Achieve3000 had a 50 scale score point advantage over their peers who did not use Achieve3000 on the EOC English I assessment, net of student demographics, family background, and school-level factors. Given the potential outcome mean for the comparison group of about 3932 scale score points, a 50-point advantage is almost equivalent to reaching the final recommended level II performance standard of academic readiness, a standard to be set for all students beginning in the 2021–2022 academic year (Texas Education Agency, 2016). The average 9th grade Achieve3000 student, then, is ahead of schedule in reaching the higher reading standard to which the district is moving incrementally.

Not unlike the gains to English-language ability seen among 9th grade Achieve3000 students, those witnessed among 10th graders, which was on the order of 69 scale score points, were also the difference between meeting the 2015–2016 level II satisfactory progression standard of academic readiness and the 2021–2022 final recommended level II performance standard of academic readiness. A 69 scale score point advantage among

Achieve3000 10th graders over the potential outcome mean among the comparison group of 10th graders of about 3938 scale points suggests that the former group are well ahead in meeting the higher reading standards currently set for the 2021–2022 academic year (Texas Education Agency, 2016).

To the extent that the knowledge economy depends on how successfully engaged students are with the reading and writing demands of secondary, and eventually postsecondary, education, these gains, derived from the use of Achieve3000, are both needed and desirable for a broader set of students. The acquisition of increasingly abstract concepts and the ability of students to analyze and synthesize information from multiple sources improve the aim of the district to produce graduates who are more competitive in a global context (OECD, 2000). Indeed, improved literacy contributes to individuals' personal development inasmuch as individuals parlay such skills to enhance their participation in society and their employability in the labor market. In the aggregate, then, improved literacy has important consequences for both labor force participation and unemployment rates. The more students gain in reading competencies and literacy skills, the higher the labor-force participation rate will be and the lower the unemployment rate will be.

The more immediate concern of district policymakers with respect to literacy is focused on its impact on the likelihood of higher academic achievement, school completion, and postsecondary enrollment and attendance. While it is true that literacy skills are primarily a result of formal education, to a considerable degree they may be augmented by factors outside the classroom such as self-directed differentiated programs like Achieve3000 that actually do what they are designed to do when used with fidelity. However literacy is increased, though, it is doubtless that it leads to higher measured academic ability, which is associated with higher attainment. Not only do students excel in the classroom, but they pursue ever-higher levels of education beyond the secondary level, whether that education is traditional postsecondary attendance at a college or university or whether it is the initiation into a trade or guild.

Limitations

The findings of this evaluation notwithstanding, there are a few limitations that could attenuate the effects of Achieve3000 on student outcomes. First, it would be beneficial to know the extent to which teacher attributes account for students' English-language ability. In particular, how, if at all, do students' primary English teachers change the strong causal impact of Achieve3000 on their achievement on the STAAR English assessment? Unfortunately, because mobility between classrooms can be considerable, particularly at the beginning of the academic year, it is difficult, and therefore not feasible, to parse teacher effects from Achieve3000 effects. It is hoped that school effects are highly correlated with teacher effects such that the absence of the latter in the models shown in this paper is inconsequential. It cannot be gainsaid, however, that controlling for teacher effects is also desirable as it is certainly possible that some teacher effects are not correlated with school effects.

Two other limitations relate to the services provided by the vendor: days of on-site professional development and/or online workshops for teachers and the access capacity at each Achieve3000 school. The former allows for an assessment of whether Achieve3000 effectiveness is a consequence of the amount of formal training provided to teachers. For instance, some schools received more days of on-site training and no online workshop training while other schools received fewer days of on-site professional development supplemented by online workshop training. Might these differences in professional development between schools attenuate the Achieve3000 impact on student outcomes? While such information was available in prior academic years, however, no such data was made available for the 2015–2016 academic year.

The latter limitation related to services provided by the vendor, i.e., platform access at each site, coupled with grade-specific enrollment numbers, would allow for an assessment of the extent to which potential availability constraints might prevent Achieve3000 usage by those who desire to use it and, hence, attenuate the effect of Acheive3000 on student outcomes. If a school has a limited number of platforms, but a glut of students attempting to use them, overall Achieve3000 usage could be high but would vary across individuals such that some would have fewer completed exercises than they might have had at a site with more platforms. This under-utilization of Achieve3000, if it exists and if it could be accounted for, could potentially reveal an underestimation of the impact of Achieve3000 given that the numbers of students is large enough. The more students engage the tool, the more they are likely to improve those skills in which they once underperformed.

Conclusion

Consistent with the findings of the two previous academic years, the present evaluation found that Achieve3000's self-directed differentiated reading

solutions are beneficial to students' literacy, where literacy is understood to be comprised of skills aimed at analysis and synthesis of printed (or electronic) texts in a process of personal development that has long-term consequences relative to achievement, years of schooling, and labor force participation. The clear implication of this is that the district should not only continue to employ this tool, as it will likely remain a boon to 9th and 10th grade students' academic achievement, but it might also consider purchasing additional licenses to extend Achieve3000 availability to elementary and middle school grades as well. At the school level, English-Language Arts, Reading, and English teachers, after they have undergone professional development training on Achieve3000 reading solutions, should actively encourage their students to engage the tool at every opportunity, particularly as a supplementary resource to enhance literacy skills beyond the classroom. Such encouragement should occur on a regular basis so that student knowledge of Achieve3000 becomes second nature. The greater the degree of awareness that the Achieve3000 tool is one that students can utilize at any time, the greater will be the potential benefit to those who take advantage of it. The ultimate result across the district could be a dramatic rise in the aggregate literacy rate, fulfilling in one aspect the aim of the district to produce graduates who have the skills necessary to compete with their peers around the world.

References

Achieve3000. (2015). National Lexile study 2014–2015. Retrieved from http://doc.achieve3000.com/marketing/A15-004_NationalLexile_2014-15_Final.pdf

Achieve3000. (2016). Achieve3000's LevelSet Assessment frequently asked questions. Retrieved from http://doc.achieve3000.com/article/LexileFAQ/LexileFAQ.pdf.

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., & Zhang, J. (2013). *The condition of education 2013* (NCES 2013-037). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

Borman, G. D., Park, S. J., & Min, S. (2015). The district-wide effectiveness of the Achieve3000 program: A quasi-experimental study (ED558845). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

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

Cheung, S. Y., & Andersen, R. (2003). Time to read: Family resources and educational outcomes in Britain. *Journal of Comparative Family Studies*, 413–433.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Huber, P. J. (1967, June). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1, pp. 221-233). Berkeley, CA: University of California Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5–86.

OECD. (2000). *Literacy in the information age: Final report of the international adult literacy survey*. Paris: OECD.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, D.C.: National Academies Press.

Snyder, T. D., de Brey, C., & Dillow, S. A. (2016). *Digest of education statistics 2014* (NCES 2016-006). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

Reeves, B. R. (2014). What was the impact of the Achieve3000 literacy program on student reading performance, 2013–2014. Houston, TX: Houston Independent School District.

Texas Education Agency. (2016, November 1). STAAR raw score conversion tables for 2015–2016. Retrieved from http://tea.texas.gov/Student_Testing_and_Accountability/Testing/State_of_Texas_Assessments_of_Academic_Readiness_(STAAR)/STAAR_Raw_Score_Conversion_Tables_for_2015-2016/

Torres, D. D. (2015). Achieve3000 impacts on student reading and STAAR EOC English I, English II, and Biology exams for the 2014–2015 academic year. Houston, TX: Houston Independent School District.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–830.

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

Independent Variables

There were four race/ethnicity variables, each of which was dummy coded: Black (0 = non-Black, 1 = Black), Hispanic (0 = non-Hispanic, 1 = Hispanic), Asian/Pacific Islander (0 = non-Asian/Pacific Islander, 1 = Asian/Pacific Islander), and other race (0 = non-other race, 1 = other race). Whites served as the reference group. Gender was coded 0 if a student was female and 1 if a student was male. Current age was an integer centered at the grand mean. Economic disadvantage was determined by whether a student did not qualify for reduced or free lunch or receive other family aid or services (coded 0) or did quality for reduced or free lunch or receive other family aid or services (coded 1). At-risk status was dummy coded (0 = not-at-risk, 1 = at-risk) based on whether a student met one or more of thirteen (13) criteria under TEC §29.081, Compensatory and Accelerated Instruction. English Language Learner (ELL) status was coded 0 for non-ELL students and 1 for ELL students. Dummy variables were also included for whether a student was categorized as a special education student (0 = non-special education student, 1 = special education student), magnet student (0 = non-magnet student, 1 = magnet student), gifted/talented student (0 = non-gifted/talented student, 1 = gifted/talented student), or homeless student (0 = non-homeless student, 1 = homeless student).

Analytic Strategy

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. The counterfactual, or potential outcomes, framework requires that we ask what an individual's outcome would have been had they been exposed to a condition other than that to which they were exposed. That is, we want to know, for all subjects in a study, their potential outcomes under all possible treatment assignments.

The implication of the fundamental problem of causal inference, then, is that the researcher is presented with a missing data problem. To resolve this problem, assumptions must be made about the data to fill in the missing values. To arrive at an unbiased and consistent estimate of the average treatment, or causal, effect, it is sufficient to assume that individuals in the treatment group, on average, are identical to individuals in the control group with respect to potential outcomes. Most often this condition may be achieved by random assignment of individuals to treatment and control.

Sometimes, however, random assignment is not possible, and all that the researcher has are actually observed outcomes. In this case, a sufficient condition that must hold to arrive at an unbiased and consistent estimate of the average treatment, or causal, effect is that of *ignorability*. Specifically, the potential outcomes under the treatment or control condition must be jointly independent of assignment to the treatment or control condition. In the present study, this simply means that students' use or non-use of Achieve3000 must be independent of their likelihood of achieving increased reading acuity. Since in observational studies ignorability seldom holds without adjustments, this evaluation conditions on students' demographic and school-level factors. Selecting on enough observable covariates in this way allows one to confidently assert the nonexistence of any unobservable covariates that affect both the treatment assignment and the potential outcomes.

Use of the treatment effects estimator also requires the assumptions that each individual has a positive probability of receiving, net of covariates, each treatment level (the overlap assumption), and that the potential outcomes and treatment or control status of each individual are unrelated to the potential outcomes and treatment or control status of all other individuals in the population (the independent and identically distributed [i.i.d.] sampling assumption; Imbens and Wooldridge, 2009; Wooldridge, 2010).

IPWRA. 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). The IPWRA estimator uses three steps to estimate treatment effects. First, employing a logit model, the treatment model estimates, for each student in the sample, the probability of treatment assignment (i.e., the propensity score is calculated). The propensity scores allow for students to be matched within each treatment level. As long as each student has a positive probability of receiving each treatment level, the inverse weights will not be too large. Second, an ordinary least squares (OLS) regression is fitted, in which the inverse of the estimated propensity scores are used as weights on the treatment dummy, controlling for covariates, to estimate the treatment-specific predicted outcomes for each student in the sample. Third, the average treatment effect is computed by contrasting the weighted average of the predicted outcomes of the treatment and control groups. Because treatment assignment is assumed to be independent of the potential outcomes, net of covariates, this three-step process provides consistent estimates of the average treatment effect. To correct for the three-step process, Huber-White standard errors are reported (Huber, 1967; White, 1980).

Appendix B

Table B1. Treatment Level Predicted Scale Score Outcome Deriving from the Inverse-Probability-Weighted Regression-Adjusted Estimator.

	2016 STAAR EOC English I Scale Score		
	Model 1	Model 2	Model 3
ATE			
Treatment			
1 vs. 0	216.52***	62.99***	50.17***
	(13.38)	(11.47)	(12.07)
POmean			
Treatment			
0	3903.20***	3930.06***	3931.55***
	(6.56)	(6.13)	(6.12)

Note: Coefficients under Model 1 are unconditional, while those under Model 2 are net of all student-level controls and those under Model 3 are net of all student- and school-level controls. Robust standard errors are in parentheses.

*p < 0.05, **p < 0.01, ***p < 0.001; two-tailed tests.

Table B2. Treatment Level Predicted Scale Score Outcome Deriving from the Inverse-Probability-Weighted Regression-Adjusted Estimator.

	2016 STAAR EOC English II Scale Score		
	Model 1	Model 2	Model 3
ATE			
Treatment			
1 vs. 0	72.52***	67.83***	68.59***
	(16.86)	(13.77)	(15.45)
POmean			
Treatment			
0	3939.17***	3939.96***	3938.16***
	(7.76)	(7.32)	(7.33)

Note: Coefficients under Model 1 are unconditional, while those under Model 2 are net of all student-level controls and those under Model 3 are net of all student- and school-level controls. Robust standard errors are in parentheses.

*p < 0.05, **p < 0.01, ***p < 0.001; two-tailed tests.