Technical Note

Volume 3, Number 1 September 2014 Steven M. Urdegar, Ph.D, Director

Achieve 3000:
An Analysis of Usage and Impact, 2013-14

At A Glance

This analysis of the dose response and impact of Achieve 3000 examined the reading achievement of students with disabilities who worked with the application during the 2013-14 school year. The analysis compared participating students' posttest scores, at each of three levels of activity completion, to the posttest scores of a reference group, controlling for usage, initial ability, and demographic differences; and also compared their performance with that of similar students in similar schools who did not use the software. The findings indicate that the application did not improve the achievement of the students who used it.

Background

Achieve 3000 is an online differentiated reading program for students with disabilities and English Language Learners in grades 6-8. The software delivers differentiated assignments at 12 different reading levels. The software features internal assessments that continuously gauge students' reading levels, provides feedback to teachers, and automatically adapts content as Lexile levels change. Students practice 20 minutes per day, five days a week. The purpose of this paper is examine the extent of the impact of the Achieve 3000 program on the M-DCPS students with disabilities in traditional (non-charter) M-DCPS schools who used it during the 2013-14 school year.

Methods

The district's Office of Program Evaluation conducted a study to examine students' usage of Achieve 3000 and to gauge its impact on students' achievement scores. The study was guided by a series of questions:

1. **To what extent was Achieve 3000 used by students during the 2013-14 school year?**
2. **Did students who used the software more frequently score higher on standardized achievement tests than students who were typical users?**
3. **Did students who used the software score higher on standardized achievement tests than similar students in similar schools who did not use the software?**

Office of Assessment, Research, and Data Analysis
1450 NE 2nd Avenue, Suite 208, Miami, Florida 33132
(305) 995-2943 Fax (305)995-7521

Data were gathered from two sources to address the research questions: (a) usage information provided by the software vendor and (b) student demographic and assessment data maintained on the district's data warehouse.

- **Usage**

  The sample for the study included all students with disabilities in grades 6 through 8 in middle schools and K-8 centers who used the Achieve 3000 software during the 2013-14 school year. The identifying information in the vendor-provided files was first validated against district records. Then, two measures of usage were obtained: (b) total hours used and (b) number of multiple choice activities completed. In separate analyses, all records with zero usage were removed. Then, hours of usage was sorted within grade and classified in four bands, based on percentile: Low (0 to 39.99), Typical (40.00 - 59.99), High (60.00 - 89.99), and Max (90.00 - 100.00). These bands were defined to provide for inferential comparisons between targeted percentiles of usage located at the midpoint of each band within the distribution: Low (20^{th}), Typical (50^{th}), High (75^{th}), and Max (95^{th}). Analyses conducted for this section were limited to descriptive statistics.

- **Dose Response**

  A predictive correlational design (Tuckman, 1999) was used to gauge the impact of usage of the Achieve 3000 program on students' achievement. The sample was the same as that used in the analysis of usage except that only students who completed one or more activities were included. Students who did not have valid Florida Comprehensive Assessment Test 2.0 (FCAT 2.0) pre- and post- test scores at consecutive grades, were excluded from the analysis.

  FCAT 2.0 is a criterion referenced test designed to measure students' mastery of the state's Next Generation Sunshine State Standards (NGSSS) and is the primary accountability measure used by the state of Florida through 2013-14. It was administered statewide to students in Reading (Grades 3 through 10) during April of each school year. Students' performance on FCAT 2.0 is measured in scale scores (i.e., equal units of achievement amenable to mathematical manipulation and specifically designed to compare individuals and groups) and reported in achievement levels that range from 1 (low) to 5 (high).

  The analysis compared students' posttest scores at each of the three levels of Achieve 3000 activity completion (Low, High, and Max) to the posttest scores of a reference group of students with "Typical" activity completion, controlling for their usage time, initial ability, and demographic characteristics. As the number of students who used the software declined sharply with increasing usage, usage was transformed to restore normality using a $base_2$ logarithmic transformation.

  Separate regression analyses at each grade were used to predict the influence of usage time, demographic characteristics, pretest, and activity completion on the students'

posttest scores. Dichotomous variables were defined for three levels of activity completion (i.e., Low, High, and Max) and for eight demographic variables (i.e., Female, Black, Free/Reduced Price Lunch eligible, English Language Learner status, Over Age for Grade, and three separate indicators for the primary exceptionalities [a] Autistic Spectrum Disorder and [b] Other Health Impaired, and [c] Hard of Hearing/Specific Learning Disability. Interactions between each of the activity completion levels and the pretest were also defined to account for the possibility that the effect of activity completion varied with the level of the pretest.

- **Impact**

A non-equivalent groups quasi-experimental design (Campbell & Stanley, 1963) was used to gauge the impact of the program on students' achievement. The sample was the same as was used in the analyses of dose response except that only students who completed a sufficient number Achieve 3000 activities to achieve a median of 40 at each grade were included.

A comparison group was also defined by matching to each member of the program group on eight student-level variables (i.e., Pretest, Female, Black, Free/Reduced Price Lunch eligible, English Language Learner status, Over Age for Grade, and three separate indicators for the primary exceptionalities [a] Autistic Spectrum Disorder, [b] Other Health Impaired), and [c] Hard of Hearing/Specific Learning Disability six school-level variables (i.e., Percent of students who are Black, Hispanic, Eligible for Free/Reduced Price Lunch, and Proficient in Reading; and Latitude and Longitude), and an index of comparability produced from those variables.[A] Students who were exposed to the program in a quantity insufficient to be included in the analysis or who did not attend the same school during October and February of the 2013-14 school year were excluded from both groups.

Matching was conducted using Multivariate and Propensity Score Matching Software with Automated Balance Optimization (Mebane & Sekhon, 2011; Sekhon, 2011) in R version 3.0.2 (R Development Core Team, 2013). Matching was conducted within grade and without replacement. As such, the matching procedure yielded balanced groups of matched students at each grade. Nevertheless, independent sample t-tests conducted on all of the individual-level and school-level variables within each grade level identified significant differences for Pretest, Female, and Black in seventh grade indicating that the treatment group was comprised of students who were initially lower achieving, and that it was not possible to draw statistically equivalent matches.

Separate regression analyses, conducted at each grade, were used to compare the difference in the groups' posttest scores controlling for the influence of the pretest and demographic predictors previously identified. Interactions between the program indicator and the pretest were also defined to account for the possibility that the effect of the program varied with the level of the pretest.

Results

- **Usage**

Non-zero usage was sorted within grade and classified in four bands, based on percentile, with midpoints as follows: Low (20th), Typical (50th), High (75th), and Max (95th). These bands were centered at the 20th, 50th, 75th, and 95th percentiles, respectively. Table 1 lists for each grade: the total number of students, the hours used, and number of multiple choice activities completed by students at the midpoints of the second and fourth bands of usage.

Table 1. Achieve 3000 Usage Metrics by Grade

	Time			Multiple Choice Activities		
		Percentiles			Percentiles	
Grade	n	50	95	n	50	95
6	798	2.96	24.97	674	9.00	99.40
7	827	3.31	17.84	712	10.00	84.35
8	876	3.32	17.05	748	12.00	76.55
Total	2,501	3.27	18.66	2,134	10.00	84.00

The table shows that the program was used by around 800-900 students per grade level. during the 2013-14 school year. However, half of the students used the software for less than 3.27 hours all year, and 5% used it for more than 18.66 hours. Of those students, approximately 700 at each grade completed at least one activity. Half of those students completed fewer than 10 activities all year, and 5% completed more than 84. Both usage and completion declined with grade.

- **Dose Response**

The predictive correlational design was applied using separate regression analyses conducted by grade, which compared the students' posttest scores at different levels of activity completion controlling for usage time, demographic characteristics and baseline achievement. Usage time was subjected to a base_2 logarithmic transformation to restore normality.

Three dummy variables were created for Low, High, and Max levels of activity completion, with typical activity completion serving as the reference group and eight demographic variables (i.e., Female, Black, Free/Reduced Price Lunch eligible, English Language Learner status, Over Age for Grade, and three separate indicators for the primary exceptionalities [a] Autistic Spectrum Disorder and [b] Other Health Impaired, and [c] Hard of Hearing/Specific Learning Disability were included in the analysis.

The results of this analysis are shown in Table 2, which lists for each predictor, the statistics for the unstandardized (B) coefficients and their significance, and the standardized coefficients (β) for each grade.

Table 2. Dose Response: Effect of Usage and Multiple Choice Activity Completion on the Reading Posttest

	Post Grade (2014)					
	6		7		8	
Predictor	B	β	B	β	B	β
Intercept	200.95 ***		208.41 ***		210.54 ***	
Black	-3.46 **	-.09	3.56 **	-.08	-3.85 **	-.09
English Language Learner	--	--	-4.47 **	-.09	--	--
Female	2.82 **	.08	--	--	--	--
Free/Reduced Price Lunch	-4.07 **	-.08	--	--	--	--
Over Age	-2.82 **	-.08	-4.36 ***	-.11	--	--
Pretest	0.72 ***	.66	0.71 ***	.62	0.81 ***	.68
Usage (hours)[a]	1.28 *	.12	1.24 *	.11	0.18	.02
Low	1.58	.05	1.64	.04	1.71	.05
High	-1.80	-.05	-2.65	-.07	0.12	.00
Max	-3.24	-.06	-4.08	-.07	0.18	.02
Usage Mean (hr.)	7.47		6.08		6.27	
N	610		643		685	
R^2	.52		.50		.50	

Note. The intercept is the value of the posttest when all the predictors are zero and the B (β) coefficient for each predictor is the impact of a one-point change in that predictor on the posttest when both the predictor and the posttest are in original (standard deviation) units. The practical significance of R^2, the proportion of variance in the posttest explained by the model, has been classified . Cohen (1988) as .02 (weak), .13 (moderate), and .26 (strong). All predictors are dichotomous except pretest which is continuous and expressed as a deviation from its sample mean values. Cells displayed as dashes represent predictors that were not entered into the regression model when the model was fitted. The number of multiple choice activities completed are based on their percentile rank within the sample: Low (0 to 39.99), Typical (40.00 - 59.99), High (60.00 - 89.99), and Max (90.00 - 100.00) with Typical designated as the reference group.

[a] Base 2 logarithmically transformed to restore normality. Regression coefficients produced from predictors transformed in this manner give the impact of each doubling of the predictor on the posttest.

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

The B coefficient for each predictor gives the impact of a one-point change in that predictor on the posttest, when both the predictor and the posttest are in original units. For example, in the sixth grade, a one scale-score point change in the pretest predicts a 0.72 scale score point change in the posttest.[1] Because the B for the pretest is measured in scale scores and the B for usage is measured in hours, the two coefficients can't be compared. A β coefficient also gives the impact of the predictor on the posttest, but because it is unitless, it can be compared with other β coefficients.

For example, in the sixth grade, Black, Over Age, and Free/Reduced eligibility are each shown to have a similar effect on the posttest. The table shows that generally students who score low on the pretest, are classified as Black, or are English Language Learners, overage for grade, or eligible for Free/Reduced Price Lunch tend to score lower than students not so

[1] In grades (6 -8), a one scale point increase corresponds to a Lexile™ gain of around 100 (Knutson, 2006) and represents somewhat more than one month of growth (Florida Department of Education, 2012).

classified. Examination of the relative strength of those effects reveal pretest to be the strongest, followed by Black, Over Age, and English Language Learner. No other significant demographic effects were found.

With regard to total hours used, the table shows a significant positive effect in sixth and seventh grade, and no significant effect in eighth grade. In grade 6, doubling the usage from 7.47 to 14.94 hours predicts a 1.28 scale score point increase on the posttest, and redoubling the usage from 14.94 hours to 29.88 hours predicts an additional 1.28 scale score point increase. In grade 7, doubling the usage from 6.08 to 12.16 hours predicts a 1.24 point scale score point increase, and redoubling the usage from 12.16 to 24.32 hours predicts an additional 1.24 scale score point increase. A significant effect for Activity Completion over and above Usage was not found at any grade.

- **Impact**

The impact analysis compared the performance of a group of students who completed a median of 40 multiple choice activities to a group of students with no exposure to the program who were matched to the program group on eight individual-variables, six school-level variables, and an index of comparability produced from those variables. Separate full regression analyses, conducted at each grade, were used to compare the difference in the groups' posttest scores controlling for the influence of the pretest and demographic predictors previously identified. Interactions between the program indicator and the pretest were also defined to account for the possibility that the effect of the program varied with the level of the pretest. Table 3 lists for each predictor the statistics for the unstandardized (B) coefficients and their significance, and the standardized coefficients (β) for each grade.

Table 3. Regression Analysis of the Effects of the Program on the Posttest

	Post Grade (2014)					
	6		7		8	
	B	β	B	β	B	β
Intercept	201.73 ***		208.97 ***		214.52 ***	
Black	--	--	--	--	-4.79 **	-.09
English Language Learner	--	--	--	--	-5.22 **	-.09
Over Age	--	--	-4.24 **	-.10	--	--
Pretest	0.73 ***	.68	0.69 ***	.61	0.74 ***	.70
School Free/Reduced Price Lunch[a]	-0.16 ***	-.12	-0.17 **	-.12	--	--
Program	-1.76	-.05	-3.75 **	-.10	-1.52	-.04
Program (S.E.)	1.26		1.40		1.34	
Usage Mean (hr.)	15.81		13.81		12.64	
N	395		387		373	
R^2	.52		.52		.56	

Note. All variables are dichotomous except pretest which is expressed as a deviation from its sample mean. Each unstandardized (B) coefficient gives the influence on of a unit change in the predictor on the criterion. Each standardized (β) coefficient gives the influence of a one standard deviation change in the predictor on the criterion. The intercept gives the value of the criterion when all the predictors are zero.

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

The table shows that students in Grade 7 who used the program at designated levels had significantly lower reading scores than the comparison group with no other significant program effects observed at Grades 6 and 8. Although, a dose-response analysis identified a significant dose response effect for Grades 6 and 7, a 10 and 38-fold increase in dosage, respectively, would be needed to produce significant improve achievement in those grades.[B]

Discussion

The Office of Program Evaluation conducted an analysis of the dose response and impact of Achieve 3000. It examined the reading achievement of special education students who worked with the application during the 2013-14 school year. The analysis compared participating students' posttest scores, at each of three levels of usage, to the pretest scores of a reference group, controlling for initial ability and demographic differences; and also compared their performance with similar students in similar schools who did not use the software.

Findings indicate that the software was typically used by around 800 students per grade for around 3.25 hours to complete 10 multiple choice activities. However, when compared with a group of students who did not use the program, no significant effect on achievement was found. Although, a dose response analysis identified significant dose response effect for Grades 6 and 7, very large increases in dosage would be needed to significantly improve achievement in those grades. These findings indicate that the application cannot be considered to have improved the achievement of the students who used it.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching*. Boston: Rand McNally.

Florida Department of Education (2012). Florida's school grading system 2011-12 changes. Paper presented at the Annual Meeting of the Florida Organization of Instructional Leaders (Lake Mary, Florida), May 10, 2012. Retrieved September 8, 2014 from http://www.floridafoil.com/wp-content/uploads/2011/11/School-Grading-Changes_FOIL_051012_final.pdf

Knutson, K.A. (2006). Because you can't wait until spring: Using the SRI to improve reading performance. Professional Paper. New York: Scholastic, Inc. Retrieved August 31, 2014 from http://teacher.scholastic.com/products/sri_reading_assessment/pdfs/SRI_ProfPaper_ImprovePerformance.pdf

Mebane, W., & Sekhon, J.S. (2011) Generic optimization using derivatives: *The Rgenoud package for R. Journal of Statistical Software, 42*(11), 1-26. Retrieved, July 14, 2009, from http://sekhon.berkeley.edu/papers/MatchingJSS.pdf

R Development Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria: ISBN 3-900051-07-0. Retrieved, May, 2 2014, from http://cran.cnr.berkeley.edu/bin/windows/base/R-3.0.2-win.exe

Sekhon, J.S. (2011) Multivariate, and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42*(7), 1-52. Retrieved, July 14, 2009, from http://sekhon.berkeley.edu/papers/MatchingJSS.pdf

Tuckman, B.W. (1999). *Conducting educational research*. Belmont, CA: Wadsworth Group/Thompson Learning.

[A] The index of comparability used in the matching process was the natural logarithm of the likelihood ratio of the expected probability that a given student was a member of the program group, as estimated by separate logistic regression procedures conducted at each grade, based on students' individual demographic characteristics and baseline achievement, and their school's demographic characteristics and geographic location.

[B] At Grade 6, a significant impact would be indicated by a B coefficient of 2.47 (1.96 X 1.26) and require a change of 4.23 points from the current estimate of -1.76. The change can be produced through a 3.30 (4.23 ÷ 1.28) unit change in dosage, which because usage was transformed, corresponds to 9.8 ($2^{3.30}$) fold increase in dosage over the mean usage of 15.81 hours. At Grade 7, a significant impact would be indicated by a B coefficient of 2.73 (1.96 X 1.40) and require a change of 6.48 points from the current estimate of -3.75. The change can be produced through a 5.23 (6.48 ÷ 1.24) unit change in dosage, which because usage was transformed, corresponds to 37.5 ($2^{5.23}$) fold increase in dosage over the mean usage of 13.81 hours.