Technical Note

Volume 3, Number 2 September 2014 Steven M. Urdegar, Ph.D, Director

My Virtual Reading Coach:
An Analysis of Usage and Impact, 2013-14

At A Glance

This analysis of the dose response and impact of My Virtual Reading Coach examined the reading achievement of struggling readers 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 posttest scores of a reference group, controlling for 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

My Virtual Reading Coach (MVRC) is an online program for students who have been identified as struggling readers. It is used as an intervention within the Response to Intervention (RtI) framework, as well as for students with disabilities. The software addresses reading sub-skills (i.e., comprehension, fluency, phonemic awareness, phonics, and vocabulary) and offers multiple approaches in several of them. Pre-recorded podcasts use speech pathologists and reading coaches to show the placement of the tongue, teeth, and lips in order to guide struggling readers. The software provides diagnostics in eye tracking and reading sub-skills, automatic progress monitoring, and individualized student learning plans. Students in grades K-12 with reading levels from pre-primer to grade 12 are expected to work with the software 30 minutes per day, four days a week. This purpose of this paper is to examine the usage, and to analyze the impact of the MVRC program on the M-DCPS students who used it during the 2013-14 school year.

Methods

MVRC is an online program for struggling readers who are within the RtI framework or who are classified as students with disabilities. Although the software may be used by students in grades K-12, the vast majority of the users were in elementary grades. The district's Office of Program Evaluation conducted a study to examine students' usage of MVRC and to assess its impact on elementary students' reading achievement scores. The study was guided by a series of questions:

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

1. **To what extent was MVRC 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?**

Data were gathered from three sources to address the research questions: (a) usage information provided by the software vendor, (b) student demographic and assessment data maintained on the district's data warehouse, and (c) an RtI database supplied by the Office of Exceptional Student Education.

- **Usage**

  The sample for the study included all students in grades K through 5 at traditional schools who used the MVRC software during the 2013-14 school year. The identifying information in the vendor-provided files was first validated against district records. Then, vendor-supplied usage in hours 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 MVRC program on students' achievement. The sample was the same as was used in the analysis of usage except that students were excluded from the analysis if they did not have valid pre- and post- test scores at consecutive grades. The results of two different achievement measures were used in this analysis: (a) the Stanford Achievement Test, Tenth Edition (SAT-10) and (b) the Florida Comprehensive Assessment Test 2.0 (FCAT 2.0). The SAT-10 served as the posttest in Grades 1-2 and the FCAT 2.0 served as the posttest in 3-5. The SAT-10 served as the pretest in Grades 1-3 and the FCAT 2.0 served as the pretest in Grades 4 - 5.

  The SAT-10 is a standardized norm-referenced test designed to measure students' performance in comparison to a national normative sample. Students' performance is measured in scale scores that are equal units of achievement that vertically align across grades, are amenable to mathematical manipulation, and specifically designed to compare individuals and groups. The SAT-10 is administered locally to all students in Grades K through 2 during the spring of each school year.

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 usage (Low, High, and Max) to the posttest scores of a reference group of students with "Typical" usage, controlling for their initial ability, and demographic characteristics.

Separate regression analyses at each grade were used to predict the influence of demographic characteristics, pretest, and usage on the students' posttest scores. Dichotomous variables were defined for three usage levels (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, [b] Gifted, and [c] the eight remaining exceptionalities combined). Interactions between each of the activity completion levels and the pretest were also defined to account for the possibility that the effect of usage 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 used the software enough to achieve a median of 10 hours at each grade, and who had an RtI designation or who were classified as students with disabilities were included.

A comparison group was also defined by matching to each member of the program group on the same eight student-level variables, six school-level variables, 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. Independent sample t-tests conducted on all of the individual-level and school-level

variables within each grade level did not identify any significant school or individual differences at any grade, indicating that comparison group was statistically equivalent on all the matching variables.

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 and the hours used by students at the midpoints of the second and fourth bands of usage.

Table 1. MVRC Usage by Grade

		Percentiles	
Grade	n	50	95
K	120	4.62	31.53
1	164	9.84	50.39
2	213	5.13	34.79
3	332	6.35	30.11
4	126	2.72	26.14
5	97	4.63	24.39
Total	1,052	5.80	35.41

The table shows that the program was used by around 100 to 165 students at Grades K, 1, 4, and 5 and by around 210 to 335 students at Grades 2 and 3. Half of the students used the software for less than 5.80 hours all year, and 5% used it for more than 35.41 hours.

- **Dose Response**

The predictive correlational design was applied using regression analysis. Separate regression analyses conducted by grade compared the students' posttest scores at different levels of activity completion controlling for usage time, demographic characteristics and baseline achievement. Three dummy variables were created for Low, High, and Max levels usage with typical usage serving as the reference group. Eight student-level 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, [b] Gifted, and [c] the eight remaining exceptionalities combined) were included in the analysis.

Table 2 lists for each predictor, the statistics for the unstandardized (B) coefficients and their significance, and the standardized coefficients (β) for each grade. Statistics on the quality of the model, R^2 and the sample size, N, are found at the bottom of the table.

Table 2. Dose Response Analysis

	Post Grade (2014)									
	1		2		3		4		5	
Predictor	B	β	B	β	B	β	B	β	B	β
Intercept	542.77 ***		568.23 ***		182.25 ***		193.96 ***		188.21 ***	
Black	--	--	--	--	--	--	-5.09 *	-.14	--	--
Free/Reduced Price Lunch	--	--	--	--	--	--	-9.07 *	-.15	--	--
Gifted	--	--	22.51 **	.14	8.23 *	.08	--	--	--	--
Over Age	-34.38 *	-.17	--	--	--	--	-7.28 **	-.21	0.72 ***	.66
Pretest	0.73 ***	.65	0.53 ***	.71	0.45 ***	.85	0.65 ***	.59	-2.81	-.08
Students with Disabilities	--	--	-13.69 *	-.12	-8.99 ***	-.17	--	--	10.76 *	.18
Low	-15.78	-.13	7.39	.09	2.56	.06	-0.98	-.03	3.07	.08
Medium	17.31	.14	11.54 *	.13	0.64	.01	3.95	.11	2.23	.04
High	20.88	.10	6.70	.05	-6.39 [a]	-.08	2.37	.04	--	--
Pretest Mean	468.39		538.4		573.7		176.5		182.4	
N	120		262		127		104		104	
R^2	.46		.63		.65		.51		.45	

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). Demographic predictors are dichotomous, while the pretest is continuous and expressed as a deviation from its sample mean value. Cells displayed as dashes represent predictors that were not entered into the regression model when the stepwise rules for model fitting were applied.

[a]A statistically significant negative interaction for **High x Pretest** was found, which indicates that the effect of High vs. "typical" usage on students' posttest scores is significantly positive for students with pretest scores in stanines 1-2, not significant for students with pretest scores in stanine 3, and significantly negative for students with pretest scores in stanines 4-9; when all the other predictors in the model are taken into account.

* 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 third grade, a one scale-score point change in the pretest predicts a 0.85 scale score point change in the posttest. Because the B for the pretest is measured in scale scores and the B for each usage band is measured in hours, the two coefficients can't be compared. A β coefficient also gives the impact of a predictor on the posttest, but because it is unitless, it can be compared with other β coefficients. For example, in the second grade, Gifted and classification as a Student with Disability are each shown to have similar but opposite effects on the posttest.

Table 2 shows that generally students who scored low on the pretest, or who are classified as Black, English Language Learners, overage for grade, or eligible for Free/Reduced Price Lunch tend to score lower than students not so classified. Examination of the relative strength of those effects reveal pretest to be the strongest, followed by Black, Over Age, and English Language Learner.

With regard to dose response, the table shows a significant positive effect for High vs. typical usage was found in second grade. In third grade, a non-significant effect for Max vs. typical usage accompanied by a significant interaction was found, which indicates that effect of Max vs. typical usage varied with baseline achievement. The effect of High vs. typical usage on students' posttest scores was significantly positive for students with pretest scores in the bottom 10^{th} percentile and significantly negative for students with pretest scores above the 25^{th} percentile. No other significant dose response effects were found.

- **Impact**

The impact analysis compared the performance of a group of students who used the software for a median of 10 hours to a group of students with no exposure to the program who were matched to the program group on nine individual-level variables, six school-level variables, and an index of comparability produced from those variables.

Students who were exposed to the program, but did not meet the criteria for inclusion, or who did not attend the same school during October and February of the 2013-14 school year, or who were not participants in either the RtI process or who were not classified as students with disabilities, were excluded from both groups.

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. Statistics on the quality of the model, R^2 and the sample size, N, are found at the bottom of the table.

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

	Post Grade (2014)									
	1		2		3		4		5	
Predictor	B	β	B	β	B	β	B	β	B	β
Intercept	508.50 ***		566.35 ***		176.77 ***		195.21 ***		193.45 ***	
Student										
Black	--	--	--		--	--	-9.92 ***	-.28	-10.80 ***	-.31
Female	--	--	9.21 *	.11	--	--	--	--	--	--
Gifted	--	--	--	--	15.75 **	.15	--	--	--	--
Free/Reduced Price Lunch	--	--	-0.43	.00	--	--	--	--	-16.47 **	-.20
Over Age	--	--	--	--	--	--	--	--	10.60 ***	.29
Students with Disabilities[a]	--	--	0.46	.00	--	--	--	--	--	--
Pretest	0.49 ***	.36	0.65 ***	.79	0.38 ***	.64	0.80 ***	.64	0.85 ***	.68
Program	-9.55	-.12	-4.13	-.05	-0.35	-.01	0.16	.00	-2.65	-.08
School										
Hispanic	--	--	0.25 *	.17	--	--	--	--	--	--
Reading Proficiency	0.56 *	.21	--	--	--	--	--	--	--	--
N	95		135		189		109		85	
R^2	.16		.64		.44		.46		.58	

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 student-level predictors except pretest are dichotomous and all continuous predictors are expressed as deviations from their sample mean values. School level predictors as expressed as percentages. Cells displayed as dashes represent predictors that were not entered into the regression model when the stepwise rules for model fitting were applied.

[a] Excludes students classified as Gifted or Autistic Spectrum Disorder

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

Table 3 shows that students who used the program did not have significantly different reading scores than the comparison group at any grade. No significant interactive effects were found. Generally students who score low on the pretest or were eligible for Free/Reduced Price Lunch tend to score lower than students not so classified.

Discussion

The Office of Program Evaluation conducted an analysis of the dose response and impact of My Virtual Reading Coach. It examined the reading achievement of students who were participating in the Response to Intervention (RtI) process, or who were classified as 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 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 approximately 100-300 students per grade for around 5.80 hours per year. However, greater usage did not result in improved achievement. When compared with a group of students who did not use the program, no significant effect on achievement was found. 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.

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.