



YEAR THREE REPORT: EVALUATION STUDY OF THE WRITING ROAD TO READING

Gary Bitter, Mary Aleta White
Arizona State University

Arizona State University conducted the third of a four-year quasi-experimental study for Spalding in the 2008-2009 school year. This study involved 44 teachers and 996 students in the final study sample. In this third year, the study followed the 2006-2007 kindergarten students into second grade in the same 5 experimental and 6 control schools. The following research questions continue to guide the study design and methods:

1. Do children who participate in *Spalding's Writing Road to Reading* program demonstrate significant learning gains in reading skills?
2. How does the reading skill attainment of children participating in *Spalding's Writing Road to Reading* program compare to that of children participating in other core reading programs?
3. How well do teachers implement *Spalding's Writing Road to Reading* program in their varied classrooms?

METHOD

To study teacher implementation, researchers utilized a uniform quantitative instrument to measure how *The Writing Road to Reading* was being implemented in the experimental classrooms. In order to measure program implementation, researchers collected data through classroom observations using observation protocols designed to measure constructs such as classroom management, adherence to program philosophy, and strategies for spelling, writing, and reading content. Both experimental and control teachers also completed a survey questionnaire that provided a variety of background information including degrees, certifications, endorsements, and professional development activities over the past ten years. Other items included length of time implementing reading programs, materials used, assessment practices, and the number of years teaching at the current grade level.

For student measures, researchers employed the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) as the primary measure to assess changes in students' reading skills during the 2008-2009 school year. The DIBELS measures used in this study were primarily Oral Reading Fluency and Retell Fluency. All participating students were first tested within the first three weeks of the 2008-2009 school year, again in December, 2008 or January, 2009, and finally in May or early June, 2009.

Participants

This study was conducted in 11 diverse Arizona schools with a total of 991 participating second-grade students at the first benchmark test, and increasing slightly to 996 total students by the year-end test. The experimental students can be further separated into two groups: the first group includes those students who were also in the kindergarten study (total of 351 at year-end), and the second group was all students in the grade level (540 at year end or 189 new experimental students by year-end). Table 1 lists the study schools, number of classes and number of students at year-end.

Table 1: *Schools Included in the Analysis*

Group	Name of School	# Total 2nd graders	# of classrooms
Experimental	Alhambra	76	3
	Bret Tarver	128	5
	CTA-Liberty	141	5
	Gallego	90	4
	Valley Academy	105	4
Total	5	540	21
Control	#1 – M	103	5
	#2 – N	104	5
	#3 – O	84	4
	#4 – P	22	1
	#5 – Q	111	5
	#6 - R	32	3
Total	6	456	23

There is a slight decrease in the control sample size between Year 1 and Year 3. It is the result of a reorganization of one control school which had 4 classrooms averaging 23 students in previous years. Project researchers met with school district staff and received approval to “substitute” another control school with very similar demographics in the fourth and final year of the study.

Table 2 presents the student distribution information for the experimental and control groups at the start of each school’s academic year. Compared to Year 1 and Year 2 data, there has been a steady decline in the number of students identified as English Language Learners in both the experimental and in the control groups. On the other hand, the SES measure, free and reduced lunch has increased in control schools from the first year, so that overall 60% of participating students in the control group qualified as low-income compared to 38% for the experimental group. This differential has been constant and has widened since Year 1. The overall percentage of minority students by group has been higher in the experimental group during Years 1 and 3.

Table 2: *Demographics of the Experimental and Control Groups (percentages)*

Group	% Girls	% ELL	% Hispanic	% Minorities	% F/RL
Experimental	48.7	24.6	51.4	66.8	38.1
Control	50.0	27.3	47.9	62.3	59.9

PROGRAM IMPLEMENTATION, RESULTS, & DISCUSSION

The classroom observations were the primary measure for program implementation. The goal for observations is to see consistent Spalding instruction across grade levels and schools. In terms of experimental teachers’ performance, the year-end researchers’ overall observation protocol results showed that at least 81% of program practices were satisfactorily implemented by experimental teachers with 11% of experimental teachers’ behaviors needing further refinement. The final observation summary showed that in the area of program philosophy and spelling, most teachers were successfully adhering to *The Spalding Method*.

According to control teacher questionnaires, all second-grade control schools used either Houghton or Harcourt reading program. These programs were evaluated by the Arizona Department of Education as core reading programs under Reading First. Control teachers received from 2 to 5 hours of inservice training on these publishers’ materials.

Student Performance Results

Table 3 displays the comparative performance of the Spalding and the control students on the DIBELS that were administered in the Fall of 2008, Winter, and Spring of 2009.

Table 3: *Comparative Mean Scores of Spalding and Control Second-Grade Students on the DIBELS (Fall 2008, Winter 2009, Spring, 2009)*

		Experimental	Longitudinal Group	Control	Difference
Fall, 2008	NWF	84.88*	87.05	67.27	17.61
	WUF	44.50*	44.90	32.59	11.91
	ORF	73.63*	77.06	49.71	23.92
	RTF	28.59*	29.74	15.42	13.17
Winter, 2009	ORF	98.59*	101.65	66.33	32.26
	RTF	38.72	41.09	25.39	13.33
Spring, 2009	ORF	109.96**	111.17	87.48	22.48
	RTF	44.83**	45.82	33.47	11.36

*p<.05

**p<.01

Similar to last year, Spalding students had consistently higher *mean* values on all DIBELS areas, which provides preliminary evidence that Spalding has been more effective than other methods used in the control schools in teaching those reading skills. The students who have participated in the study since Year 1 have scores that are higher than or equal to the overall experimental students.

In addition to measures of statistical significance, researchers frequently calculate and report measures of practical significance, known as the effect size. The effect size is a way to help educators decide whether a statistically significant difference between programs translates into a meaningful difference—one that would justify a program adoption for instance. There are different ways to measure effect sizes. One commonly used measure is called Cohen’s *d*. Cohen’s *d* using a pooled standard deviation was computed for DIBELS ORF scores at the end of each benchmark assessment. The effect size for Fall 2008 was .7; for Winter 2009 it was .8; and in Spring 2009 the effect size is .6. This means that the intervention has a positive, medium effect on student achievement and is a more effective reading program. Converting the score to percentiles would mean the average student in the Spalding sample, at the end of the year, would score higher than 73% of the control sample.

As shown in Table 4 below, additional analyses of the extent to which experimental students experienced learning gains by the end of second grade as well as between the beginning and the middle of the school year show that they exceeded the DIBELS decision rules benchmarks for achievement at each testing period.

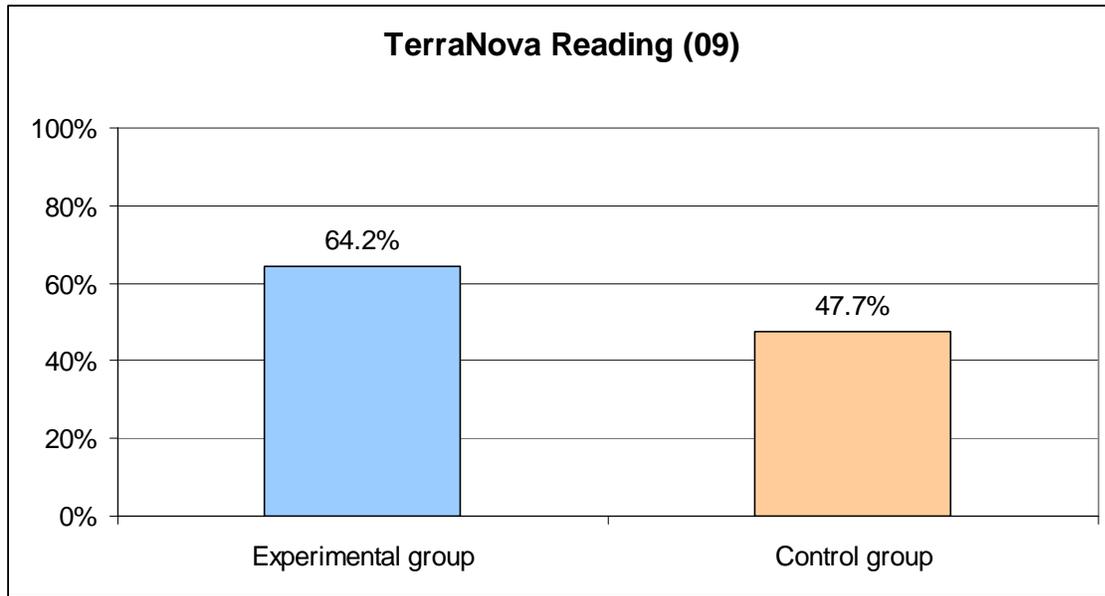
Table 4: *Second-grade Students’ Mean Post-Scores on DIBELS*

		Mean Test Scores			
		Spalding Experimental Schools	Spalding Longitudinal Group	Control Schools	DIBELS End of Second-grade Low risk score
		(n = 512)	(n = 363)	(n = 439)	
Oral Reading Fluency (ORF)					
	Spring test	109.96	111.17	87.48	90

As can be seen, Spalding participants experienced significant gains in reading performance from the beginning, to the middle and end of the school year. Unfortunately, by the middle and ending benchmarks in second grade, the average control student is not meeting the DIBELS assessment for low-risk scores.

Another analysis of reading achievement was available in this year’s study because all second grade students are required to complete the state’s norm-referenced achievement test, TerraNova. The chart below represents a sample of the study students (three control and three experimental schools) and their average NCE score on the TerraNova reading portion. As would be expected from reviewing the DIBELS scores, the Spalding students’ NCE scores were significantly higher than the control students on the state test ($p < .01$).

Chart 1: Student NCE reading scores from Spring, 09 AZ TerraNova exam



Connecting Teachers to Student Data

As noted earlier, teacher survey data is available for both the experimental and control group. Demographic data from the surveys show that the average teacher age between the two groups is fairly close: 40 for the experimental group and 36 for the control group. The experimental group had slightly more teaching experience (9 versus 7).

Teacher experience was re-coded into a range of beginning or little experience to those who have taught for over 15 years. The average ORF scores of students within the five categories vary widely. That is, additional years of teaching was not associated with higher levels of student assessment results. The same was true for the number of years using the Spalding method.

SUMMARY

According to the year three results, students who used *The Writing Road to Reading* continue to demonstrate statistically significant learning gains as measured by DIBELS. In addition, their scores were significantly higher than control group student scores again this year. Since both the control groups and the experimental groups used detailed teacher guides evaluated by NCLB for research-based reading components, theoretically, they should have produced similar results. This was not the case. These preliminary findings are strongly suggesting that use of *The Writing Road to Reading* curriculum is an effective method for enhancing performance on critical early literacy skills.