



**YEAR TWO REPORT: EVALUATION STUDY OF
THE WRITING ROAD TO READING**

First Grade Study

Gary Bitter, Mary Aleta White

Arizona State University

September, 2008

MARY LOU FULTON COLLEGE OF EDUCATION
TECHNOLOGY BASED LEARNING AND RESEARCH
ASU SkySong
1475 N. Scottsdale Rd., Suite 200 Scottsdale, AZ 85257-3538
Phone # (480) 884-1694 Fax # (480) 884-1888
<http://tblr.ed.asu.edu>

Table of Contents

List of Tables	3
INTRODUCTION	4
RESEARCH DESIGN	4
METHOD	
Teacher Measures	4
Student Measures	5
Participants	5
PROGRAM IMPLEMENTATION, RESULTS, & DISCUSSION	
Implementation by Treatment Teachers	7
Student Performance Results	7
Longitudinal Student Analysis	11
Connecting Teachers to Students	13
SUMMARY	13

List of Tables

Tables

Table 1	Schools Included in the Analysis	5
Table 2	Student Distribution by Student Level Variables	6
Table 3	Demographics of the Experimental and Control Groups (percentages)	6
Table 4	Comparative Mean Scores of Spalding and Control First Grade Students on the DIBELS (Fall 2007, Winter 2008, Spring, 2008)	8
Table 5	Comparative DIBELS Statistics of the Spalding and Control First Grade Students (Fall 2007, Winter and Spring, 2008)	9
Table 6	First Grade Students' Mean Post- Scores on DIBELS	10
Table 7	DIBELS Cut Scores for Bench Marks	11
Table 8	First Grade Students From Year 1 By School	11
Table 9	Comparative Mean Scores of Cohort vs Non-Cohort First Grade Students on the DIBELS	12
Table 10	Comparative Mean Scores of Spalding Kindergarten Students, Year 1 and Year 2	12
Table 11	Demographics of Participating Teachers	13

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

Gary Bitter, Mary Aleta White
Arizona State University

INTRODUCTION

This project was initiated in 2006-2007 because Spalding Education International has seen positive results in test scores at schools where the Spalding *Writing Road to Reading* program has been implemented. The research project is being conducted to validate the effectiveness of this program in teaching reading to children from varied backgrounds who attend different types of schools.

RESEARCH DESIGN

Arizona State University conducted the second of a four-year quasi-experimental study for Spalding in the 2007-2008 school year. This study involved 47 teachers and 1,002 students in the final study sample. In this second year, the study followed the 2006-2007 kindergarten students into first grade in the same 5 treatment and 6 control schools.

The following research questions continue to guide the study design and methods:

1. Do children who participate in *The 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* compare to that of children participating in other, more traditional reading programs?
3. How well do teachers implement *The Writing Road to Reading* in their varied classrooms?

METHOD

This section presents descriptions of the different study components including measures, participants, and program procedures.

Teacher Measures

The study evaluated teachers' implementation of Spalding's *Writing Road To Reading*. Researchers utilized a uniform quantitative instrument to measure how *The Writing Road to Reading* was being implemented in the treatment classrooms. In order to measure program implementation, researchers collected data through classroom observations using observation protocols. Four researchers, in teams of two or as a whole group, visited the teacher classrooms three times per year and observed individual teachers to ensure inter-observer agreement and reliability. The observation protocol was designed to measure constructs such as classroom management, adherence to program philosophy, and strategies for spelling, writing, and reading content. Classroom observations lasted for approximately 45 minutes to one hour and focused on whole group instruction. Both treatment and control teachers 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.

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 2007-2008 school year. Researchers selected the DIBELS assessment because it has broad visibility, acceptance in the field, and it demonstrates high technical merit. The Arizona State Department of Education has adopted DIBELS as the assessment for its Reading First program.

Administration periods for DIBELS occurred at the beginning, middle, and end of the school year, and subtests are designed for administration across multiple years. The DIBELS measures used in this study were Letter Naming Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Oral Reading Fluency, and Retell Fluency. Some, but not all, schools tested for Word Use Fluency. All participating students were first tested within the first three weeks of the 2007-2008 school year as required by the Arizona State Department of Education, again in December, 2007 or January, 2008, and finally in May or early June, 2008.

Participants

This study was conducted in 11 diverse Arizona schools with a total of 1,055 participating first grade students at the first benchmark test, and declining to 1,002 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 433 at year-end), and the second group was of all students in the grade level (115 new experimental students by year-end). A quasi-experimental design was used to assign schools as a control or treatment school. Schools were matched on socioeconomic status of students, class size, race/ethnicity/gender of students, and geography. 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 1 st graders	# of classrooms
Experimental	Alhambra	88	3
	Bret Tarver	126	7
	CTA-Liberty	138	5
	Gallego	92	4
	Valley Academy	104	4
Total	5	548	23
Control	#1 – M	92	4
	#2 – N	104	4
	#3 – O	86	4
	#4 – P	70	3
	#5 – Q	121	5
	#6 - R	34	4
Total	6	507	24

Classes in the treatment condition (23) used the Spalding curriculum an average of 90 minutes each day, while control classes (24) used their standard core programs. Twenty-three first grade teachers participated in the study as a treatment group, while 24 teachers were in the control group for a total of 47 teachers. Teacher class size averaged 24 students in the experimental group and 22 in the control group. As an incentive for participation, treatment teachers received materials and training in *The Writing Road to Reading* program without charge. Control teachers each received \$200 gift certificates to a bookstore for classroom materials.

Table 2 presents the student distribution information for the treatment and control groups at the start of each school’s academic year. The school districts provided data for 1,077 students: 552 in the treatment group and 525 students in the control group. Table 3 provides the same data but uses percentages to demonstrate the student demographics.

Table 2: Student Distribution by Student level variables

		Treatment (n = 552)	Control (n = 525)	Overall (n = 1,077)
Gender	Female	263	259	522
	Male	289	266	555
Ethnicity	Asian	38	15	53
	Black	17	17	34
	Hispanic	262	289	551
	Native Am	12	8	20
	White	223	195	418
SES	F/R	217	314	531
Language Ability	ELL	179	229	408

Compared to year 1 data, the distribution of students by gender, race, SES, and language ability is very similar. A little more than half of the students (52%) were male, and approximately half (48%) were female. In the second year of the study, there were more EL students in the control schools, which increased the average percentage of ELL students to 38% of participants as compared to 33% in year 1. The same trend occurred in the SES measure, free and reduced lunch. The rates increased in control schools from the previous year, so that overall 49% of participating students across treatment conditions qualified as low-income compared to 47% in year 1.

Table 3: Demographics of the Experimental and Control Groups (percentages)

Group	% Girls	% ESL	% Hispanic	% Minorities	% F/RL
Experimental	47.2	32.4	47.8	59.6	39.3
Control	49.3	43.6	55.0	62.9	59.8

PROGRAM IMPLEMENTATION, RESULTS, & DISCUSSION

This section presents the results of the classroom implementation study, the student assessment scores, and an analysis of teacher characteristics as they relate to student achievement.

Implementation by Treatment Teachers

The classroom observations were the primary measure for classroom implementation. The goal for observations is to see consistent Spalding instruction across grade levels and schools. Observers noticed an increase in consistency within and across the five schools by the fourth-quarter observations. This is attributed to consistent use of the *Teacher Guides*.

In terms of treatment teachers' performance, the year-end researchers' overall observation protocol results showed that at least 87% of program practices were satisfactorily implemented by treatment teachers with 9% of treatment 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*. In program philosophy, observers noted a slight decline in the number of teachers who consistently encouraged higher-level thinking in the writing lesson. There was an increase however in the use of higher-level thinking in the reading lesson. Last, compared to the previous round of observations, there was a decline in the number of classrooms that demonstrated the connection between spelling, writing, and reading.

During the year, there was a decline in modeling composing sentences that demonstrate usage and meaning of unfamiliar words. There were also teachers needing to refine coaching as children identify and label three mental actions (24%). Observers also saw a decline in the number of classrooms where children read a decodable book in unison. In all instances this was because of poor time management.

Student Performance Results

The DIBELS tests administered to first grade students in Fall 2007 were Letter Name Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Word Use Fluency (WUF). Two treatment schools and two control schools did not administer the Word Use Fluency in Fall of 2007.

The DIBELS tests administered in winter 2008 were the same as the fall with the exception of Word Use Fluency and the addition of Oral Reading Fluency (ORF) and Retell Fluency (RF). One experimental and two control schools did not administer the Retell Fluency in Winter, 2008. During the spring testing, all schools used the same testing categories as they did in the Winter.

Table 4 displays the comparative performance of the Spalding and the control students on the DIBELS that were administered in the Fall of 2007, Winter, and Spring of 2008.

Table 4: Comparative Mean Scores of Spalding and Control First Grade Students on the DIBELS (Fall 2007, Winter 2008, Spring, 2008)

		Experimental	Control	Difference
Fall, 2007	LNF	45.74**	42.18	3.56
	PSF	47.60***	35.55	12.05
	NWF	48.87***	34.70	14.17
	WUF	35.82***	18.80	17.02
Winter, 2008	PSF	51.91***	45.31	6.61
	NWF	63.12**	56.73	6.39
	ORF	54.92***	31.36	23.61
	RF	20.72***	11.82	8.87
Spring, 2008	PSF	52.94***	48.95	4.01
	NWF	78.29*	72.69	5.65
	ORF	73.33***	53.51	19.87
	RF	30.65***	19.76	10.92

*p<.05, **p<.005, ***p<.001

Similar to last year, Spalding students had consistently higher *mean* values on all DIBELS areas, which indicate that Spalding has been more effective than all the other methods used in the control schools in teaching those reading skills. Table 5 provides further descriptive statistics on the tested areas. Students in both groups improved in reading skills by the end of year two; however, in every category the treatment group students had higher mean scores than control group students by an average of 10 points. Although some mean differences are small (e.g., LNF, PSF), the significance value is large because the issue is whether the null hypothesis (no difference in scores) is unlikely to be true. The smaller the p value, the more convincing is the rejection of the null hypothesis. In this study, a two sample t-test was used to simply answer a question about the means of the two independently observed samples.

One possible reason for the smaller mean score differences of Letter Naming Fluency and Phoneme Segmentation is that reading programs used in both control and experimental schools teach these skills. However, the largest mean score difference is in Oral Reading Fluency. Spalding's *Writing Road to Reading* teaches short and long vowel sounds and sounds of letter combinations from the beginning, enabling Spalding students to independently segment and read more words. Other score differentials are hypothesized as being due to Spalding's emphasis on multisensory word meaning/usage and comprehension strategies such as summarizing (retelling).

Table 5:
Comparative DIBELS Statistics of the Spalding and Control First Grade Students (Fall 2007, Winter and Spring, 2008)

	Fall, 2007				Winter, 2008				Spring, 2008			
	LNF	PSF	NWF	WUF	PSF	NWF	ORF	RF	PSF	NWF	ORF	RF
Spalding, Experimental Schools												
N, Valid	548	548	547	284	498	499	498	373	515	514	517	396
Mean	45.74	47.60	48.87	35.82	51.91	63.12	54.92	20.72	52.94	78.29	73.33	30.65
Std. Devia	18.7	16.0	27.9	16.8	13.7	35.4	38.4	15.8	13.6	39.8	39.9	16.7
Skewness	-.019	-.702	1.156	-.087	-.824	.622	.754	.746	.068	-.029	.392	.534
Kurtosis	.040	.974	1.379	.424	1.427	-.244	-.056	.158	4.58	-.927	-.326	.141
Max	107	110	139	89	92	142	199	71	144	151	195	92
Control Schools												
N, Valid	507	507	507	300	475	475	453	270	484	484	484	279
Mean	42.18	35.55	34.70	18.80	45.31	56.73	31.36	11.82	48.95	72.69	53.51	19.76
Std. Devi	18.7	16.8	25.8	15.4	15.5	30.0	32.2	15.4	14.2	33.2	39.3	13.4
Skewness	.185	-.550	1.391	.347	-.359	.996	1.745	1.638	-.509	.412	.999	.848
Kurtosis	-.239	-.362	2.862	-.967	2.422	.784	3.112	1.926	2.14	-.484	.671	.444
Max	99	76	147	59	133	142	199	69	107	142	215	60

As shown in Table 6 below, both groups tested above their current grade level on the posttest assessment.

Table 6
First-grade Students' Mean Post-Scores on DIBELS

		Mean Test Scores		
		Treatment	Control	DIBELS End of First- grade Low risk score
		(n = 517)	(n = 484)	
Phoneme Segmentation Fluency (PSF)	Spring test	52.94	48.95	35
Nonsense Word Fluency (NWF)	Spring test	78.29	72.69	50
Oral Reading Fluency (ORF)	Spring test	73.33	53.51	40

According to DIBELS decision rules, at the end of first grade student scores on the ORF reading level is most important. Based on their research, high scores on ORF (40 or more words correct per minute) should also mean high scores for PSF and NWF skills as well. The researchers note, “Students who meet the end of first grade benchmark goal on ORF have odds of 75 – 92 percent of achieving the second grade goal for more common patterns of performance.” Spalding students in this study exceeded the low risk reading level by twice the level: 33.33 points above the low risk indicator as opposed to 13.51 points above for the control group.

As shown in Table 7, additional analyses of the extent to which treatment students experienced learning gains 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. As can be seen in Table 7, Spalding participants experienced significant gains in reading performance from the beginning, to the middle and end of the school year.

Table 7: DIBELS cut scores for bench marks

		LNF	PSF	NWF	ORF
Beginning					
	DIBELS Benchmark	37	35	24	-
	Treatment	45.7	47.6	48.8	-
	Control	42.2	35.5	34.7	
Middle					
	DIBELS Benchmark	-	35	50	20
	Treatment	-	51.9	63.1	54.9
	Control	-	45.3	56.7	31.4
End					
	DIBELS Benchmark	-	35	50	40
	Treatment	-	52.9	78.3	73.3
	Control	-	48.9	72.7	53.5

Longitudinal Student Analysis

This study was conducted in the second of a four-year quasi-experimental study. In this second year, the study reported on all first grade students in Spalding schools during the 2007-2008 school year. Additional analysis was conducted on the 433 first grade students who were in the kindergarten cohort reported in the 2006-2007 study. Table 8 lists the experimental study schools, the total number of study students at year-end, and the kindergarten cohort.

Table 8: First grade students from year 1 by school at year-end

Group	Name of School	Total 1 st graders (07-08)	Total Kindergarten cohort (06-07)	# 1 st graders from initial kindergarten cohort	Retention rate from Year 1
Experimental	Alhambra	88	89	78	88%
	Bret Tarver	126	123	85	69%
	CTA-Liberty	138	114	98	86%
	Gallego	92	95	75	79%
	Valley Academy	104	117	97	83%
Total	5	548	538	433	

The next table displays the comparative performance of Spalding students in the two-year cohort against the smaller number of students who were not in the kindergarten cohort. As would be expected, cohort students had higher scores.

Table 9: Comparative Mean Scores of Spalding Cohort vs non-cohort First Grade Students on the DIBELS (Fall 2007, Winter 2008, Spring, 2008)

		Non-Cohort 1 st Graders	Longitudinal Cohort	Score Difference
Fall, 2007	LNF	41.54	46.83	5.29
	PSF	45.38	47.83	2.45
	NWF	45.75	49.81	4.06
	WUF	35.84	35.85	0.01
Winter, 2008	PSF	48.2	52.98	4.78
	NWF	57.41	64.06	6.65
	ORF	49.26	55.79	6.53
	RF	17.98	21.1	3.12
Spring, 2008	PSF	51.44	53.36	1.92
	NWF	71.02	78.89	7.87
	ORF	64.57	74.69*	10.12
	RF	23.34	31.81**	8.47

*p<.05

**p<.01

In an effort to examine threats to the study's internal validity (such as the Hawthorne effect), an additional analysis was conducted to compare the 2007-2008 kindergarten student achievement with that of the 06-07 kindergarten cohort. During the 2007-2008 year, the kindergarten students, teachers, and administrators were not participants in the study; however, data was collected on a sample of 07-08 kindergarten students.

Although kindergarten teachers and classrooms were not the grade level under study, student performance slightly exceeded the kindergarten cohort of 06-07. The chart below provides mean scores on the final testing categories. The schools were identically matched.

Table 10: Comparative Mean Scores of Selected Spalding Kindergarten Students on the year-end DIBELS from two year study (Spring, 2007 = Year 1; Spring, 2008 = Year2)

		Year 1 ('07) Kindergarten	Year 2 ('08) Kindergarten	Difference
Year End Testing Results	Letter Name	48.43	49.35	0.92
	Phoneme Segmentation	50.09	51.49	1.4
	Nonsense Word	45.92	50.41*	4.49
	Word Use	37.51	41.46*	3.95

*p<.05

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 of the experimental group is higher than for the control group: 45 compared to 36. The experimental group also had more teaching experience. The average number of years teaching was 13 for the experimental group and 9.5 for the control group. The experimental group obtained their initial degrees from 1971 to 2006 while the control group obtained their initial degrees between 1978 and 2006. The following table displays additional demographics for the participating teachers.

Table 11: *Demographics of Participating 1st Grade Teachers*

Group	Average age	# Females	# Males	# Minority	# White	Avg # years teaching
Experimental	45	22	1	1	22	13
Control	36	23	1	2	22	9.5

A correlational analysis revealed that the additional years of teaching was not associated with higher levels of student assessment results. In fact, the number of years teaching had a weak, negative correlation to student achievement for the experimental group ($r=-.224$). Mean scores for teachers with 3-5 years of experience were higher than all other categories.

There wasn't such a consistent pattern when analyzing the number of years a teacher has used Spalding and the resultant student achievement profile ($r=-.138$). Future studies may include regression analysis for examining the relationship between student achievement and teacher preparation.

SUMMARY

According to the year two 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. This achievement pattern is augmented by looking at the 2007-2008 kindergarten students. Their year-end scores exceed those of their counterparts in the year one study. These preliminary findings suggest that use of *The Writing Road to Reading* curriculum is an effective method for enhancing performance on critical early literacy skills.