Increasing Reading Skills in Rural Districts:
A Case Study of Three Schools

August, 2010

NIFDI Technical Report 2010-2

Prepared by Jean Stockard, Ph.D.
Director of Research
National Institute for Direct Instruction
Eugene, Oregon
jstockard@nifdi.org
877-485-1973

Increasing Reading Skills in Rural Districts: A Case Study of Three Schools

Executive Summary

Recent studies of rural education have called for identifying ways to help rural schools improve teachers' pedagogical skills and student achievement (Arnold, Newman, Gaddy, & Dean, 2005, p. 18). This report addresses that charge. It examines changes in reading skills and achievement of elementary students in three rural Midwestern school districts as they implemented the Direct Instruction (DI) reading curriculum, *Reading Mastery (RM)* with support from the National Institute for Direct Instruction (NIFDI). All of the districts were in sparsely populated areas and had higher proportions of at-risk students (measured by receipt of free or reduced lunch or minority status) than in the state as a whole.

A quasi-experimental, longitudinal design was used to examine the impact of *RM* on students' reading skills. Data from two sets of student cohorts were compared to national and state-wide data: 1) cohorts that began school before *RM* was implemented and did not start the curriculum until first grade or later, termed "partial exposure cohorts;" and 2) cohorts that had *RM* beginning in kindergarten with teachers fully complying with the NIFDI model, termed "full exposure cohorts." There were no significant differences between the cohorts in their at-risk status.

Students' reading skills were measured with two curriculum-based measures taken from the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) system: Nonsense Word Fluency (NWF), assessed from the middle of kindergarten through the beginning of second grade, and Oral Reading Fluency (ORF), assessed from the middle of first grade through the end of third grade. The ORF scores were translated into reading Lexile scores to provide comparability from one year to the next. Differences between the two cohort groups and a national sample were examined with comparisons of means, calculations of effect sizes, and comparisons to benchmarks established by the DIBELS system. In addition, differences in growth in reading skills of students in the two cohort groups were examined with multivariate linear growth modeling. All of the analysis techniques produced the same conclusion. By the middle of kindergarten, those in the full exposure cohorts had significantly higher NWF and ORF Lexile scores than students in the partial exposure cohort and scores that were equal to or significantly higher than those in the national sample. Differences remained strong and significant through the primary years, and effect sizes surpassed usual standards of educational significance. (See Figures 1 and 2.)

Data were available for one district on the percentage of fourth graders who met the established state reading standards over a five year period. Before exposure to the curriculum (in 2004-05) the district's fourth graders were less likely than those in the state as a whole to meet the established benchmarks, but within a few years they were more likely to do so. The change in the district's scores over time, relative to changes in the state were statistically significant and educationally important (d = .31). (See Figure 3.)

Figure 1: NWF Scores by Group, K-First Grade

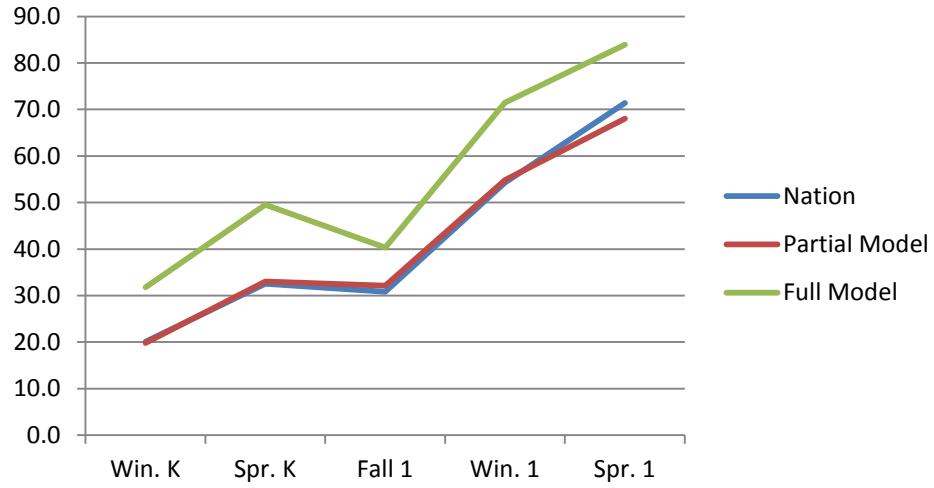


Figure 2: Oral Reading Fluency Lexiles by Group, Grades 1-3

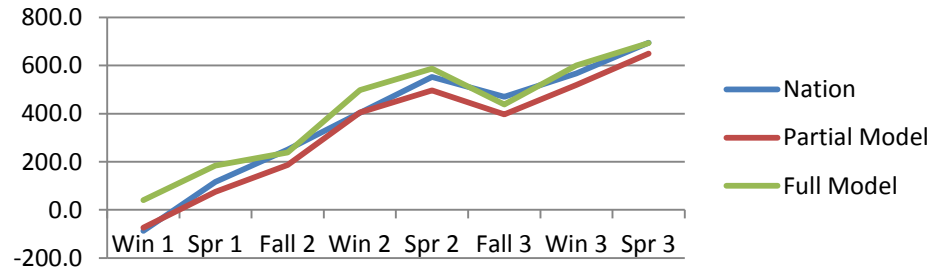
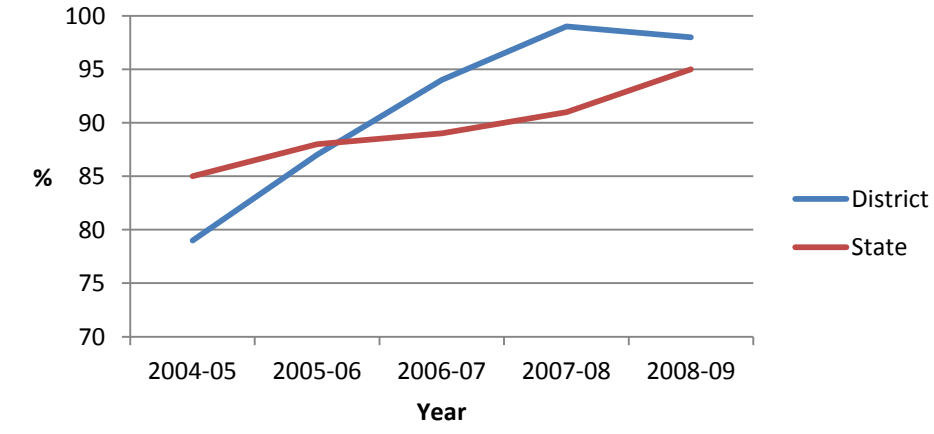


Figure 3: Percentage of Fourth Graders Meeting State Standards by Year



Increasing Reading Skills in Rural Districts:
A Case Study of Three Schools

The majority of school districts in the nation are relatively small. Close to half (47.5 percent) have fewer than 1000 students, and almost three-quarters (71.5 percent) have fewer than 2500 students (NCES, 2008). Like larger districts around the country, these small districts, almost all of which are in rural areas, are required to meet state and national standards regarding increased achievement. Small districts have some characteristics, such as limited bureaucracy and the ability to develop personal relationships among staff, students, and parents, which might help in promoting these changes (Barley and Beesley, 2007; Stern, 1995). At the same time, however, small districts may face special challenges in meeting new standards and requirements, especially in areas such as staff training, scheduling for special needs, and curriculum development. While large districts may have curriculum specialists charged with providing such on-site support to teachers, smaller districts may be less likely to have these resources (McNeil, 2009). In an analysis of the research literature regarding rural education, Arnold and associates (2005) noted this concern. They suggested that identifying ways to help rural schools improve teachers' "pedagogical skills in ways that have the greatest impact on student achievement" should be a priority area of research in rural education (Arnold, Newman, Gaddy, & Dean, 2005, p. 18).

This paper addresses that priority. We examined changes in reading skills through the primary grades of students in three rural, Midwestern districts that occurred after the implementation of a highly structured and explicit reading curriculum with associated support and guidance. Using a quasi-experimental design we compared reading skills of students who had the curriculum from the beginning of kindergarten with a) students who began the curriculum after the kindergarten years and b) national and state level data. Our analysis addresses the issues raised by Arnold and colleagues by providing an example of how rural districts in a sparsely populated state can support teachers' pedagogical development and how this support can translate into higher student achievement.

Related Literature

The scholarly literature that underlies our work comes from analyses of 1) research on the needs of rural schools, 2) the most effective curricular approaches for promoting student literacy, and 3) the role of implementation support in helping teachers deliver curricula.

Rural Education Research

Observers of rural education research have termed it "scant," noting that the area has received much less attention than urban education (Mulkey, 1993, Stern, 1994, cited by Sherwood, 2000). The most comprehensive recent review of work on rural education is undoubtedly that of Arnold, Newman,

Gaddy, and Dean (2005). Systematically examining studies that appeared over a number of years, Arnold and associates developed a research agenda for rural school improvement including improving students' "opportunity to learn," promoting teacher quality and professional development, and strengthening the capacity of rural schools and districts to improve student achievement.

In a rather interesting exchange, commentators challenged this agenda, implying that it ignored the "meaningfulness" of rural life (Howley, Theobald, & Howley, 2005). In a response, one of the original authors and a colleague restated the underlying premise of their review:

> [A]ll children and schools, including those in rural settings, deserve access to the very best information about high-quality and effective schooling….It is no longer adequate, in this day and age of research sophistication, to argue the value and success of rural education based solely on belief in and passion for rural communities. Nor is it adequate to argue that rural education is too unique to be the subject of rigorous research, or that scientific inquiry and sound decision-making are not relevant to rural education and communities. We must move beyond these beliefs and philosophical conjecture toward more rigorous research-based knowledge that gives us the information needed to direct and improve rural educational systems. (Cicchinelli & Dean, 2005, pp. 1, 2)

Our work shares this philosophical orientation and research aim, focusing on the improvement of reading skills in rural settings.

Reading Research

Reading experts stress that early development of reading has a long-term impact on a child's future. They have coined the term "Matthew effect," using the Biblical quotation that the "rich get richer and the poor get poorer," to describe the cumulative effects of good or poor reading skills on later academic success. A large body of empirical evidence demonstrates that children who can read fluently in first grade have much more success throughout their school careers. Early reading fluency results in exposure to much greater volume of material, and thus also produces a strikingly greater accumulation of vocabulary, language skills and bodies of knowledge (Cunningham & Stanovich, 1997, 1998; Francis, et al, 1998; Gough & Juel, 1991; Juel, 1988; Stanovich, 1986).

An even larger body of research, spanning several decades, has documented the importance of systematic and explicit instruction in promoting reading achievement (e.g., Adams, 1990; Anderson, Hiebert, Scott, & Wilkinson, 1985; Baker, Kameenui, Simmons, & Stahl, 1994; Bond & Dykstra, 1967 Chall, 1967; Foorman, 1995; Fukkink & deGlopper, 1998; Grossen,1997; ICHHD, 2000; Juel & Minden-Cupp, 2000; Murphy, 2004; National Reading Panel, 2000; NICHHD, 1996; Pflaum, Walberg, Karigianes, & Rasher, 1980; Smith et al., 2001; Snider, 1990; Snow, Burns, & Griffin, 1998; Stanovich, 1994). Meta-analyses show that curricula that embody these elements consistently result in larger achievement gains. Our analysis focused on the implementation of one well-established systematic and explicit reading program, *Reading Mastery (RM)*, the elements of which are described more fully in the

methodology section below. A number of studies indicate that students who receive reading instruction in this approach have higher levels of reading achievement and stronger growth in reading skills over time than students in other curricula (e.g., Adams & Engelmann, 1996; AFT, 1998; Beck & McCaslin, 1978; Borman, Hewes, Overman, & Brown, 2003; Hattie, 2009; Herman et al., 1999).

Supporting Teachers

Numerous educational researchers have highlighted the importance of technical assistance in promoting teachers' skills and the fidelity with which they implement curricula. The literature increasingly recognizes that teaching is a highly technical and involved process and that training and support are crucial for developing and honing excellent teaching skills. Studies also suggest that this assistance should be ongoing and intensive, ideally involving on-site support (Blakeley, 2001; Berends, Bodilly, & Kirby, 2002; Bodilly, 1998; Bodilly, Glennan, Kerr, & Galegher, 2007). Such support may be especially important for systematic and explicit curricula, such as *RM*, which involve a broad array of behaviors and actions for complete implementation (Engelmann & Engelmann, 2004). As would be expected, studies have found that the gap between students in DI programs, such as *RM*, and those in traditional programs is greater for students of teachers who implemented DI with higher fidelity (Carlson & Francis, 2002; Gersten, Carnine, Zoref, & Cronin, 1986). Similarly, studies that have focused only on students receiving Direct Instruction have found the highest achievers in classrooms or schools that have higher levels of fidelity of implementation (Benner, Nelson, Stage, & Ralston, 2010; Gersten, Carnine, & Williams, 1982; Ross et al., 2004).

However, reflecting the general body of educational research, all but one of these studies occurred in urban settings, and the exception (Benner et al, 2010) combined data from rural and urban schools. As noted above, rural areas may experience unique issues in providing on-going technical support and assistance, and thus it is important to examine attempts to improve student achievement and assist teachers in these settings. Our study begins to fill that gap, by examining changes in student achievement in three rural school districts as they implemented an explicit, structured curriculum with technical assistance. Our analysis provides an example of how smaller districts in isolated regions of the country can help their students meet universal standards of achievement. Based on the literature reviewed above, we expected that students who received the more structured and explicit curriculum would have higher reading achievement than the national norm, but that these differences would be greatest for students who were most fully exposed to the model.

Methodology

In the sections below we describe the sample used in the study; the procedures, including the curriculum and implementation support provided to the schools; and then the measures and analysis techniques that were used.

Sample

The sample included students from three rural districts in a Midwestern state. In 2009 fewer than 2 million people lived in the state, most of them in rural areas. The largest city had a population of less than 500,000, and more than half of the state's residents lived in communities smaller than 30,000 population. In general, the state was sparsely populated, with a population density of 22.3 people per square mile compared to 79.6 people for the nation as a whole.

All of the districts served students in a central small community and surrounding towns and rural areas and had more students with an at-risk status than in the state as a whole. District A was based in a community with about 9,000 residents and served approximately 1700 students in 2008-09. It included a preschool; two elementary schools, one serving grades K-2 and the other serving grades 2-4; a middle school; and a high school. In 2008-09 the district had a slightly higher percentage of students receiving free and reduced lunch than in the state as a whole (42% versus 39%) and a percentage of Hispanic students that was twice that of the state (18% vs. 9%). District B was based in a community of 7,800 people and had about 2,200 students enrolled in the 2008-09 school year. The district included a preschool, three elementary schools, a junior high school and a high school. Compared to the state as a whole, the district had slightly more students in poverty (42%) and significantly more minority students, almost all of whom were Hispanic (25%). District C was the smallest of the districts in the analysis, based in a community of 1000 residents and serving about 300 students in the 2008-09 year. Students came from three different communities and the surrounding rural areas, studying in two school sites, one serving students in grades K-3 and the other serving students in grades 4-12. In 2008-09 over half of the students (54%) qualified for free or reduced lunch, substantially more than in the state as a whole, but there were no Hispanic or ELL students.

The districts fully implemented the new curriculum in different years: the fall of 2007 for District A, the fall of 2004 for District B, and the fall of 2006 for District C. In addition they differed slightly in the years for which data were available. In order to maximize sample size we combined data for the three districts and compared results for two cohorts of students: a) those with full exposure to the curriculum, starting kindergarten in the first year of implementation or later (n=887), and b) those with less than full exposure, beginning school prior to the first year of implementation (n= 802). Full details on years of implementation and data availability are given in Appendix A. The results did not differ when data were

analyzed separately for each site or with more discrete separation of cohorts, and these results are available upon request from the author.

Table 1 compares the at-risk status of students in the cohort groupings for each site, using two standard measures: the proportion of students receiving free or reduced lunch and the proportion of racial-ethnic minorities. We also combined these indicators, looking at the proportion of students with either of these risk factors (Panel 3 in Table 1). The fourth panel reports the sample size for each cohort-group combination, and the fifth panel reports the results of two way analyses of variance, with site and cohort (full versus partial exposure) as factors. There were no significant interaction effects and no significant differences between the cohort groups on any of the variables, indicating that there were no differences between the cohort groups in their at-risk status. There were, however, significant differences between sites in the proportion of minority students, reflecting the much lower percentage in District C.

Table 1

At-Risk Variables by Full DI Status and Site

Proportion Minority			
	Full DI		
	No	Yes	Total
District A	0.32	0.30	0.31
District B	0.31	0.32	0.32
District C	0.09	0.02	0.07
Proportion Free or Reduced Lunch			
	Full DI		
	No	Yes	Total
District A	0.47	0.47	0.47
District B	0.46	0.51	0.50
District C	0.47	0.43	0.45
Proportion Any At Risk Status			
	Full DI		
	No	Yes	Total
District A	0.52	0.54	0.52
District B	0.55	0.61	0.60
District C	0.49	0.43	0.47
Sample Size			
	Full DI		
	No	Yes	Total
District A	577	149	726
District B	150	692	842
District C	75	46	121
Analysis of Variance Results			

	Minority		FRL		Any At Risk Status	
	F-ratio	Sig.	F-ratio	Sig.	F-ratio	Sig.
Site	15.82	<.001	0.22	0.80	2.71	0.07
Full DI	0.66	0.42	0.02	0.90	0.05	0.82
Interaction	0.45	0.64	0.41	0.67	0.69	0.50

Secondary comparative data came from two sources, one national in scope and one state-wide in nature. The first was data from all schools participating in the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) system in the 2001-2002 academic year (Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002). While the participating schools represented all areas of the country, they may not be representative of the nation, and, as noted by Good et al (2002), may, because of their investment in the measurement system, be more likely than other districts to be invested in the improvement of early reading. The second source of comparative data was the state department of education for the school districts. All schools in the state were required to participate in the statewide data gathering procedure, and thus this source should provide reliable comparative information for the state as a whole.

Procedures

As noted above, all of the sites in the rural sample implemented the Direct Instruction (DI) reading program, *Reading Mastery*. The Direct Instruction model was first developed almost four decades ago based on work with preschoolers in an "at-risk" population (Engelmann, 2007). All of the DI programs seek efficiency and effectiveness of instruction through program design, organization of instruction, and positive student-teacher interaction. The approach attempts to control all the major variables that impact student learning through the placement and grouping of students into instructional groups, the rate and type of examples presented by the teacher, the wording that teachers use to teach specific concepts and skills, the frequency and type of review of material introduced, the assessment of students' mastery of material covered and the responses by teachers to student's attempts to learn the material. The programs are constructed according to a small step design that teaches isolated skills and concepts in separate tracks that are systematically integrated with skills and concepts in other tracks in increasingly sophisticated applications. For this reason, lessons do not focus on a single skill or topic. Instead, only about 10% of a lesson's contents are new. The rest of the lesson is devoted to reviewing and applying skills and concepts that were introduced in previous lessons. Placement in the program is a critical factor in the program's success as appropriate placement allows students to both learn new concepts and skills each day (Collins & Carnine, 1988; Engelmann, 2007; Engelman & Carnine, 1982; Gersten, Darch, & Gleason, 1988; Huitt, Monetti, & Hummel, 2009).

The schools received support for implementation of the curriculum from the National Institute for Direct Instruction (NIFDI). NIFDI is a not-for-profit corporation associated with the original developer of

DI and is dedicated to providing school districts with training and support. The NIFDI model encompasses the elements found in the implementation research literature to be especially effective and important in technical support (Fixsen, et al, 2005). An Implementation Manager (IM) trains teachers, assistants, and coaches. All teachers receive pre-service training and coaching until they teach each program to a minimum adequate level of fidelity. Teachers continue to receive in-service coaching to improve implementation fidelity. The IM is typically at a school about 35 days during the school year, working in classrooms with the teachers and presenting in-service sessions. In addition, the IM reviews reports of students' lesson progress on a weekly basis, following up with conference calls to address any problems a classroom may be experiencing. During the second year of implementation, teachers who perform well are identified as coaches and are deployed to work with other teachers in the school. Beginning in the third year NIFDI support is gradually phased out so that schools can become more self-sufficient.

The NIFDI model was implemented in the rural sites in this study in the same manner in which it is implemented in other areas with one exception. In contrast to implementations in larger cities, the pre-service training for teachers was sometimes held in a central location, requiring teachers to travel elsewhere in the state to receive training. However, as with implementations in other settings, the implementation managers made regular in-person visits to each school, observing classrooms and helping teachers improve their skills, and weekly reports of student progress were reviewed and discussed. In recent years as technology has become more advanced, the in-person visits have been supplemented with observations and conferences utilizing web-based cameras and communications software such as Skype.

Measures and Analysis

Our primary measures of reading skills came from the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) system. The DIBELS measures incorporate assessments of various elements of reading development including children's ability to recognize letters and link sounds and letters. DIBELS provides a way for teachers to have regular, systematic, and efficient assessments of children's skills, with repeated short testing sessions during the school year and comparison of these assessments to established benchmarks that indicate if children are making the progress needed to achieve the goal of reading fluently by the end of grade 3. All measures result in numeric scores that indicate the number of correct responses given in one minute (DIBELS, 2008; Good, Simmons, & Kame'enui, 2001; Good, Simmons, & Smith, 1998; Hasbrouck & Tindal, 2006; Kaminski and Good, 1996). The districts administered the assessments to all students at the times specified by the DIBELS guidelines – at the beginning, middle and end of the school year – and made the data available to the researcher. As noted above, the years for which data were available for each student cohort varied slightly from one site to another, and these patterns are summarized in Appendix A.

Two DIBELS measures were used as indicators of children's reading development: Nonsense Word Fluency (NWF), which measures the ability to read phonetic nonsense words and was assessed from the middle of kindergarten through the beginning of second grade; and Oral Reading Fluency (ORF), which measures the rate at which children can correctly read connected text in grade-level materials and was assessed from the middle of first grade through the end of third grade. Although the ORF ostensibly measures decoding and fluency, research indicates that these scores are highly associated with measures of reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Good, et al, 2001). Because the connected text used for the measure of oral reading fluency is taken from grade-level material, comparisons of ORF scores from one year to the next may not provide the most optimal picture of changes in skills over time. To compensate, we transformed the ORF scores into Lexiles, a developmental scale of reading that ranges from less than zero for those who are just beginning to read to above 1700L for advanced readers. Thus it adjusts for the different content used in the ORF at each grade level (MetaMetrics 2009).

We analyzed the DIBELS data in three ways: First we examined simple descriptive statistics (means and standard deviations) for the measures at each testing period, using inferential tests and effect sizes (Cohen's d) to examine the differences between the cohorts with full exposure to the curriculum and those with less exposure as well as the national sample. Second, we calculated the percentage of students in each set of cohorts who met the benchmarks set by the DIBELS system for indicating that children were making adequate academic progress to prevent future academic difficulties and those who were at risk of future problems, using chi-square statistics to compare the percentages for each cohort and the national sample. Bonferroni corrections were used to adjust for multiple tests of significance with each set of comparisons. Finally, we used linear growth modeling to examine variations in trends in the growth of reading skills over time for students in the two cohort groups. In these models we compared the pace of growth of students with full and partial exposure to the curriculum, controlling for their at-risk status and school site. Three incrementally more complex models were compared: 1) a model that included only changes over time; 2) a model that added dummy variables for at-risk status, site, and cohort; and 3) a model that added the interaction of time and at risk status and cohort. Comparison of these models let us assess the extent to which students' skill growth over time varied by cohort, site, and at-risk status.

In addition to the formative, curriculum-based DIBELS measures, we were able, for one site (District B), to examine changes in the percentage of fourth grade students who met or exceeded state standards in reading as determined by the state's testing program. Given the amount of time that the curriculum had been implemented at this site were able to compare the percentage of students who met or exceeded the standards for students in three groups: 1) those with no exposure to *Reading Mastery* (fourth graders in 2004-05), 2) those with some exposure (fourth graders from 2005-06 through 2007-08), and 3)

those who had experienced *RM* since beginning kindergarten (fourth graders in 2008-09). We computed an effect size describing the magnitude of the difference between the district performance and that of the state and tested the hypothesis that changes over time in the district were greater than in the state as a whole. (Details on the computations are in Appendix B.)

Results

Sections below summarize the analysis of DIBELS reading scores and the state mandated achievement tests.

Nonsense Word Fluency

Table 2 reports the average Nonsense Word Fluency scores of students in the national sample and the two cohort groups at each of the testing periods. The first panel gives descriptive statistics for each group and the second panel gives the inferential results and Cohen's d, a standard measure of effect size, for each pair of comparisons. The cohorts that did not begin the curriculum until after their kindergarten year had NWF scores very similar to the national sample. In contrast, the cohorts that experienced the full model (starting the program in kindergarten) had NWF scores that were significantly higher than both the national sample and the other set of cohorts at all testing periods, even with the application of the Bonferroni correction for multiple tests of significance. All of the associated effect sizes surpassed the .25 mark typically seen as educationally significant, ranging from .30 (at the start of grade 2) to .76 (at the middle of kindergarten).

Table 2

Nonsense Word Fluency Scores National Sample and Cohorts by Testing Period

Means, S.D., and N's

	Nation			Partial Model			Full Model		
	Mean	S.D.	N	Mean	S.D	N	Mean	S.D.	N
Win. K	20.1	17.8	13,221	19.8	13.5	410	31.8	17.9	792
Spr. K	32.5	22.5	39,169	33.0	18.4	535	49.5	26.9	788
Fall 1	30.8	22.5	36,708	32.1	20.0	531	40.3	25.8	715
Win. 1	54.3	28.5	37,473	54.8	25.5	516	71.5	34.1	707
Spr. 1	71.4	34.6	36,834	68.0	31.7	526	83.9	37.2	697
Fall 2	----	----	----	64.2	30.7	520	74.0	35.1	650

Inferential Tests

	Partial vs. National Data			Full vs. National Data			Partial vs. Full Model		
	z	p	d	z	P	d	T	p	d
Win. K	-0.34	0.73	-0.02	18.47	<.0001	0.65	12.95	<.0001	0.76
Spr. K	0.47	0.64	0.02	21.21	<.0001	0.69	13.24	<.0001	0.73
Fall 1	1.32	0.19	0.06	11.30	<.0001	0.39	6.31	<.0001	0.36

Win. 1	0.42	0.67	0.02	16.07	<.0001	0.55	9.82	<.0001	0.56
Spr. 1	-2.26	0.02	-0.10	9.54	<.0001	0.35	8.09	<.0001	0.46
Fall 2	----	----	----	----	----	----	5.06	<.0001	0.30

Note: National data were obtained from Good et al, 2002, Table 3 and were not reported for second graders. Z-tests examine the null hypothesis that the sample value (the cohort mean) equals the population (national) mean. T-tests examine the null hypothesis that the mean value for the two cohorts are equal. All t-tests were adjusted for equality of variances between the groups, and probabilities are two-tailed. The Bonferroni corrected p value for .05 is .01 for each of the five comparisons between the cohort groups and the national data and .008 for the six comparisons between the two cohort groups. D values are Cohen's d, calculated as the difference between the means divided by the common standard deviation.

Tables 3 and 4 compare the NWF scores to benchmarks established by the DIBELS system. Table 3 gives data regarding the percentage of students in each group who met the NWF benchmark indicating adequate academic progress, and Table 4 reports the data regarding students predicted to be at risk of future academic difficulties. The first panel of each table gives percentages of students in these categories and the second panel reports the results of chi-square tests comparing each pair of results at each testing period. The findings parallel those in Table 2. At each testing period students in the cohorts with full exposure to the curriculum were significantly more likely than those with only partial exposure and those in the national sample to reach the established benchmarks regarding adequate academic progress and significantly less likely to be considered at risk. (The only exception was the comparison of at risk status in the fall of second grade for students in the two cohort groups.) Comparisons of students in the national sample and the partial exposure cohort varied across the testing period, with somewhat greater percentages of the partial exposure group meeting benchmarks at the earlier testing periods and slightly fewer doing so at the later testing periods.

Table 3

Percentage of Students Meeting NWF Benchmarks for Adequate Progress, National Sample and Cohorts by Testing Period

% Students Meeting Benchmarks			
	Nation	Partial	Full
Winter - K	60.0	69.8	89.4
Spring - K	60.0	66.9	86.3
Fall - 1	58.0	63.1	75.5
Winter - 1	49.0	50.8	70.3
Spring - 1	71.0	65.0	80.9
Fall - 2	---	61.7	71.5
Chi-Square Values Paired Comparisons			

	Partial v. Nation		Full v. Nation		Full v. Partial	
	Chi-Sq.	p	Chi-Sq.	P	Chi-Sq.	p
Winter - K	16.3	<.0001	285.12	<.0001	72.8	<.0001
Spring - K	10.7	0.001	227.01	<.0001	70.8	<.0001
Fall - 1	5.6	0.018	90.14	<.0001	22.5	<.0001
Winter - 1	0.7	0.403	128.32	<.0001	48.3	<.0001
Spring - 1	9.1	0.003	33.3	<.0001	39.5	<.0001
Fall – 2	---	---	---	---	12.6	0.0004

Note: National level data taken from Good, et al, 2002, Tables 5 and 6. Degrees of freedom for all chi-square tests equaled 1. Computations for the comparisons of the partial and full cohort data with national data used expected frequencies calculated from the national data. Computations for the comparisons of the full and partial cohort data were calculated in the usual manner for contingency (cross-tabulated) data. The Bonferroni corrected p value for .05 is .01 for each of the five comparisons between the cohort groups and the national data and .008 for the six comparisons between the two cohort groups.

Table 4

Percentage of Students at Risk Based on NWF Benchmarks, National Sample and Cohorts by Testing Period

% Students Meeting Benchmarks

	Nation	Partial	Full
Winter - K	18.0	14.6	2.1
Spring - K	18.0	13.8	3.7
Fall - 1	18.0	12.4	8.0
Winter - 1	15.0	14.5	6.2
Spring - 1	6.0	6.1	2.6
Fall - 2	---	7.9	7.7

Chi-Square Values Paired Comparisons

	Partial v. Nation		Full v. Nation		Full v. Partial	
	Chi-Sq.	p	Chi-Sq.	p	Chi-Sq.	p
Winter - K	3.15	0.0759	134.86	<.0001	70.26	<.0001
Spring - K	6.30	0.0121	109.47	<.0001	45.74	<.0001
Fall - 1	11.16	0.0008	48.71	<.0001	6.80	0.0090
Winter - 1	0.09	0.76	42.71	<.0001	23.46	<.0001
Spring - 1	0.01	0.92	14.43	0.0001	9.37	0.0022
Fall - 2	---	---	---	---	0.01	0.9029

Note: National level data taken from Good, et al (2002), Tables 5 and 6. Degrees of freedom for all chi-square tests equaled 1. Computations for the comparisons of the partial and full cohort data with national data used expected frequencies calculated from the national data. Computations for the comparisons of the full and partial cohort data were calculated in the usual manner for contingency (cross-tabulated) data. The Bonferroni corrected p value for .05 is .01 for each of the five comparisons between the cohort groups and the national data and .008 for the six comparisons between the two cohort groups.

Table 5 reports the results of the growth curve analysis of NWF scores, utilizing data from the two sets of cohorts in the rural districts. Linear growth models provide a more parsimonious examination of the changes in reading skills over time than the comparisons in Tables 2 through 4 and also can control for students' at-risk status and differences between the districts. The bottom panel of Table 5 describes the models that were tested and gives the -2 Log Likelihood (-2LL) statistic, commonly used as a measure of fit. The -2 LL can be compared across models, with the difference having a chi-square distribution. These differences and the associated degrees of freedom are also in the bottom panel of Table 5. They indicate that, while Model 2 provided a significantly better fit than Model 1, Model 3 was not significantly better than Model 2. In other words, a model that included interactions between time and cohort and between time and at-risk status did not provide a significantly better fit to the data. Thus coefficients are reported only for Model 2.

Table 5

Growth Curve Model Results, Nonsense Word Fluency, Mid-K to Beginning of 2nd

Parameter Estimates, Model 2			
Fixed Effects			
	est.	s.e.	prob.
Intercept	21.50	0.99	<.0001
Time	9.95	0.17	<.0001
At risk	-9.36	0.94	<.0001
District B	10.90	1.17	<.0001
District C	3.72	1.91	0.05
Full Model	8.55	1.12	<.0001
Random Effects			
Between persons (11)	131.5	11.9	<.0001
Within persons (12)	45.9	3.1	<.0001
In growth rates (22)	18.7	1.5	<.0001
Residual	291.6	6.0	<.0001
Models and Fit Statistics			
	1	2	3

Intercept	X	X	X
Time	X	X	X
At risk		X	X
District B		X	X
District C		X	X
Full Model		X	X
Time * at risk			X
Time * Full Model			X
-2 LL	66550.8	66190.1	66188.2
Ch. In -2 LL	-----	360.7	1.9
df change	-----	4	2
prob.	-----	<.001	n.s.

The coefficients for the analysis of NWF scores in Model 2 are given in the top panel of Table 5. The random coefficients were all statistically significant, indicating that students' scores differed from one student to another, from one time period to another, and that students had different rates of growth. All of the fixed coefficients were statistically significant. They indicate significant increases in NWF scores from one period to another, lower scores for students at risk (through free and reduced lunch and/or minority status), and higher scores for students in Districts B and C than in District A. As expected, students in cohorts with full exposure to DI had significantly higher NWF scores than students in other cohorts throughout the testing period.

To summarize, the analysis of NWF DIBELS measures indicated that students with full exposure to the curriculum, beginning in the kindergarten years, had higher average scores than students in the nation as a whole and those in the same districts who with only partial exposure to the curriculum. These results appeared at all time points. Students with full exposure were also more likely to reach the DIBELS benchmark of making adequate progress and less likely to be categorized as being at risk of future failure. Multivariate analyses of growth over time indicated that the advantage appeared by the first testing period in the middle of kindergarten and remained relatively constant through the beginning of second grade.

Oral Reading Fluency Lexiles

Table 6 gives the descriptive statistics for the Oral Reading Fluency Lexile scores for students in both cohorts and the national sample at each testing period (Panel 1) and the associated tests of significance and effect sizes (Panel 2). Students in the partial implementation cohorts had significantly lower Lexiles than the national group in six of the eight testing periods. In contrast, those experiencing the full implementation had lower scores than the national group in only the fall administrations and this difference was significant (with the Bonferroni correction) in the fall of third grade. At all testing periods

the full implementation cohorts had higher scores than the partial implementation cohorts, and these differences were significant (with the Bonferroni correction) in all but two periods in third grade.

Table 6

Oral Reading Fluency Lexile Scores, Winter, 1st Grade through Spring, 3rd Grade, National Sample and Cohort Groups

Means, S.D., and T-tests									
	National Sample			Partial Model			Full Model		
	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.	N
Win 1	-87.6	78.7	37410	-74.6	282.7	512	40.3	299.9	704
Spr 1	116.2	72.8	37017	74.3	307.9	524	184.3	289.1	695
Fall 2	250.9	157.6	15494	186.7	222.7	520	238.7	233.9	648
Win 2	403.2	212.3	16841	404.8	263.6	505	497.1	274.8	378
Spr 2	552.6	235.2	16805	496.5	265.8	514	586.3	277.6	375
Fall 3	469.7	214.3	10941	396.3	248.6	490	438.1	258.5	360
Win 3	566.4	225.4	12318	518.4	273.4	363	599.9	261.5	207
Spr 3	694.7	236.8	12531	648.9	281.6	367	692.8	248.5	204
Inferential Tests									
	Partial Model vs. Nation			Full Model vs. Nation			Full versus Partial Model		
	z	p	d	z	P	d	t	sig.	d
Win 1	3.76	0.0002	0.07	43.13	<.0001	0.68	6.82	<.001	0.39
Spr 1	-13.17	<.0001	-0.22	24.66	<.0001	0.38	6.39	<.001	0.37
Fall 2	-9.29	<.0001	-0.34	-1.98	0.05	-0.06	3.85	<.001	0.23
Win 2	0.16	0.87	0.01	8.60	<.0001	0.39	5.06	<.001	0.34
Spr 2	-5.40	<.0001	-0.22	2.78	0.005	0.13	4.88	<.001	0.33
Fall 3	-7.58	<.0001	-0.32	-2.79	0.005	-0.13	2.38	.02	0.16
Win 3	-4.06	<.0001	-0.19	2.14	0.032	0.14	3.48	.001	0.30
Spr 3	-3.71	0.0002	-0.18	-0.12	0.900	-0.01	1.93	.05	0.17

Note: National data were obtained from Good et al, 2002, Tables 3 and 4. Z-tests examine the null hypothesis that the sample value (the cohort mean) equals the population (national) mean. T-tests examine the null hypothesis that the mean value for the two cohorts are equal. All t-tests were adjusted for equality of variances between the groups, and probabilities are two-tailed. The Bonferroni corrected p value for .05 is .006 for each set of 8 comparisons. D values are Cohen's d, calculated as the difference between the means divided by the common standard deviation.

Tables 7 and 8 report data regarding the extent to which students met benchmarks for adequate academic progress (Table 7) and being at-risk for future difficulties (Table 8), and the results parallel those reported in Table 6. Students in the partial implementation cohort were significantly less likely than

students in the nation as a whole to be making adequate progress and significantly more likely to be categorized "at risk" in five of the eight testing periods. In contrast, students in the full implementation cohort were more likely than students in the national sample to be making adequate progress and less likely to be at risk, with statistically significant differences in most comparisons. All of the comparisons between students in the full and partial cohorts indicated that those exposed to the curriculum beginning in kindergarten were more likely to be making adequate progress and less likely to be at risk. As with the analysis of average scores, differences between the groups were smaller at the fall testing before instruction began.

Table 7

Comparisons of Percentage of Students Reaching DIBELS ORF Benchmarks for Adequate Progress by Group and Testing Period

Percentage of Students Meeting ORF Benchmarks			
Test Period	Nation	Partial	Full
Winter 1	61	60	77
Spring 1	65	59	79
Fall 2	60	52	61
Winter 2	60	63	78
Spring 2	61	52	73
Fall 3	60	47	54
Winter 3	60	52	70
Spring 3	60	54	68

Chi-Square Values Paired Comparisons						
	Partial v. Nation		Full v. Nation		Partial v. Full	
Test Period	Chi-sq.	sig.	Chi-sq.	sig.	Chi-sq.	sig.
Winter 1	0.15	0.70	79.73	<.0001	42.2	<.0001
Spring 1	8.38	0.004	63.56	<.0001	60.3	<.0001
Fall 2	13.47	0.0002	0.54	0.46	10.2	.001
Winter 2	1.86	0.17	49.78	<.0001	22.3	<.0001
Spring 2	17.71	<.0001	21.95	<.0001	39.5	<.0001
Fall 3	34.83	<.0001	5.10	0.02	4.3	.04
Winter 3	10.89	0.001	7.89	0.005	17.6	<.0001
Spring 3	6.11	0.01	4.97	0.03	10.6	.001

Note: Degrees of freedom for all chi-square tests equaled 1. Computations for the comparisons of the partial and full cohort data with national data used expected frequencies calculated from the national data. Computations for the comparisons of the full and partial cohort data were calculated in the usual manner for contingency (cross-tabulated) data. The Bonferroni corrected p value for .05 is .006 for each set of 8 comparisons.

Table 8

Percentage of Students at Risk Based on ORF Benchmarks, National Sample and Cohorts by Testing Period

% Students Meeting Benchmarks			
Test Period	Nation	Partial	Full
Winter 1	12.0	10.4	4.7
Spring 1	12.0	17.0	7.1
Fall 2	19.0	22.2	14.9
Winter 2	26.0	23.2	14.0
Spring 2	21.0	27.0	17.0
Fall 3	19.0	26.4	21.8
Winter 3	19.0	25.7	13.5
Spring 3	16.0	20.4	12.3

Chi-Square Values Paired Comparisons						
	Partial v. Nation		Full v. Nation		Full v. Partial	
Test Period	Chi-Sq.	p	Chi-Sq.	p	Chi-Sq.	p
Winter 1	1.31	0.25	35.13	<.0001	13.96	0.0002
Spring 1	12.19	0.0005	15.99	<.0001	28.92	<.0001
Fall 2	3.37	0.07	7.03	0.008	10.01	0.002
Winter 2	1.99	0.16	28.19	<.0001	11.77	0.001
Spring 2	11.18	0.001	3.65	0.06	12.28	0.001
Fall 3	17.60	<.0001	1.90	0.17	2.34	0.13
Winter 3	10.58	0.001	4.03	0.04	11.62	0.001
Spring 3	5.25	0.02	2.13	0.14	5.99	0.014

Note: National level data taken from Good, et al (2002), Tables 6 and 7. Degrees of freedom for all chi-square tests equaled 1. Computations for the comparisons of the partial and full cohort data with national data used expected frequencies calculated from the national data. Computations for the comparisons of the full and partial cohort data were calculated in the usual manner for contingency (cross-tabulated) data. The Bonferroni corrected p value for .05 is .006 for each set of 8 comparisons.

Table 9 gives the results of the growth model analysis of changes in the Lexile scores over time. The model fit statistics shown in the bottom panel of Table 9 indicate that Model 2 provided a significantly better fit than Model 1 and that Model 3, which includes interactions of group and at-risk status with time, fit significantly better than Model 2. Coefficients for both Model 2 and Model 3 are given in the top panel of the table. As with the analysis of NWF, the random coefficients were all statistically significant, indicating that students' scores differed from one student to another, from one time period to the next and that students had different rates of growth. Fixed coefficients associated with

the main effects were similar between the two models, indicating that the increase in scores from one time period to the next was statistically significant, that at-risk students had significantly lower scores, and that students in the full exposure cohorts had significantly higher scores. The interaction effects included in Model 3 indicated that the at-risk students had a slightly greater increase in scores over time and that those in the full exposure cohorts had slightly smaller increases over time. As shown in Table 6, however, even with this slightly greater rate of growth over time, those in the partial cohorts still had lower scores than those with full exposure at the end of third grade. Unlike the results with NWF, there were no significant differences in Lexile scores or changes in these scores between the students in the three districts.

Table 9

Growth Curve Model Results, Reading Lexiles, Mid-1st to End of 3rd

Parameter Estimates, Models 2 and 3

	Model 2			Model 3		
Fixed Effects						
	est.	s.e.	prob.	est.	s.e.	prob.
Intercept	34.39	13.63	0.01	36.14	14.47	0.0126
Time	105.46	1.08	<.0001	103.98	1.87	<.0001
At risk	-116.30	13.69	<.0001	-150.07	15.06	<.0001
District B	14.63	17.72	0.41	19.59	17.74	0.27
District C	34.77	27.33	0.20	34.29	27.32	0.21
Full DI	74.40	17.11	<.0001	101.06	18.04	<.0001
Time * at risk	-----	-----	-----	11.25	2.12	<.0001
Time * Full DI	-----	-----	-----	-9.86	2.11	<.0001
Random Effects						
Between persons (11)	67374.0	2966.1	<.0001	66863.0	2931.6	<.0001
Within persons (12)	-1817.5	331.7	<.0001	-1634.6	317.6	<.0001
In growth rates (22)	215.5	57.8	<.0001	153.4	55.2	0.003
Residual	14299.0	323.5	<.0001	14321.0	324.1	<.0001

Models and Fit Statistics

	1	2	3
Intercept	x	x	x
Time	x	x	x
At risk		x	x
Gering		x	x
Loup City		x	x
Full DI		x	x
Time * at risk			x
Time * Full DI			x

-2 LL	81506.4	81376.8	81322
Ch. In -2 LL	-----	129.6	54.8
df change	-----	4	2
prob.	-----	<.001	<.001

Note: The three way interaction of time, at risk status, and Full DI was not statistically significant.

To summarize, as with the analysis of NWF scores, in comparison to the students with only partial exposure to the curriculum, those who experienced the full implementation of the curriculum had significantly higher average ORF Lexile scores, were more likely to reach the benchmark of adequate progress and less likely to be considered at risk of future academic difficulties. Differences were also significant in most comparisons with the national sample. The multivariate analysis of growth over time, which controlled for at-risk status and differences between the sites, indicated that these differences appeared at the first testing period, but declined slightly over time (a slightly lower rate of growth for those in the full implementation cohorts).

Performance on Statewide Assessments

Table 10 reports the percentage of fourth graders who met the established state reading standards in District B and the state as a whole from 2004-05 to 2008-09. In the earliest year none of the fourth graders had any exposure to *RM*, while in the last year the fourth graders had been exposed to the curriculum since beginning kindergarten. Thus the data allow us to compare students' performance on the statewide test before any exposure to the curriculum, with partial exposure during their primary grades, and with full exposure.

Students in both the district and the state improved their performance over time, but the increase was greater for students in District B than in the state as a whole. Before implementation of the curriculum the performance of District B's fourth graders was lower than that of students in the state as a whole, but by 2006-07, their performance was better than that of students state-wide. The fifth column of Table 10 compares the district values to those of the state using a z-score, or standard deviation unit score, which is equivalent to an effect size. It can be seen that the effect sizes were negative for the year with no DI exposure and the next year, but then became positive, as students had more exposure. Using the formula described in Appendix B we calculated a t-test examining the null hypothesis that the changes in the district over time, relative to the changes in the state, were equal to zero. The results indicate that these changes were statistically significant ($t = 2.52$, $df = 259$, $p = .006$ [one-tailed]), and the associated effect size of .31 for these changes would be considered educationally important. In other words, the change from 2004-05 to 2008-09 in the percentage of students in District B who passed the state exam was significantly greater than the change in this percentage in the state as a whole.

Table 10

Percentage of Fourth Graders Meeting State Reading Standards, by Exposure to Curriculum, District B and State

Year	% - District	% - State	N	Z – score	Grades Exposed to RM	Type of Exposure
2004-05	79	85	117	-0.18	0	None
2005-06	87	88	147	-0.05	gr 3-4	Partial
2006-07	94	89	156	0.14	gr 2-4	Partial
2007-08	99	91	144	0.26	gr 1-4	Partial
2008-09	98	95	144	0.14	gr K-4	Full

Summary and Discussion

This report examined changes in reading skills and achievement among elementary students in three rural, K-12 districts in one sparsely populated state. All of the districts implemented Direct Instruction's *Reading Mastery*, a highly explicit and structured reading curriculum with implementation support from the National Institute for Direct Instruction. The districts ranged in size from about 200 to over 2000 students. All of the districts had somewhat more students receiving free or reduced lunch than in the state as a whole, and two of the districts had substantially more Hispanic students. A quasi-experimental longitudinal design was employed, comparing achievement of students in cohorts that experienced full implementation of the curriculum with those who experienced the curriculum during only part of their primary years as well as with data from national and state-wide samples.

Analyses of the curriculum-based DIBELS measures from kindergarten through third grade indicated that students with full exposure to the program had significantly higher scores, were more likely to be meeting benchmarks indicating adequate progress, and were less likely to be termed at risk than those with less exposure. In most comparisons, those with full exposure also had significantly greater skills than the national sample. Smaller differences tended to occur in the fall testing, after the summer recess and before the beginning of instruction during the school year. However, the differences grew as instruction continued. In most cases the effect sizes comparing students with full exposure to those in the other groups surpassed the usual criterion of educational importance. Similar results appeared with the analysis of fourth grade students' performance on statewide reading assessments, with significantly significant changes from before implementation of the changes to the time when they were fully established.

Most important, the results illustrate how rural school districts with limited staff can provide consistent support to teachers and improve student achievement. The changes were incorporated within the usual public school system without lengthening the school day or setting up alternative systems, such as a charter school organization. As described above, the NIFDI support model includes a gradual phase-out of outside support, with training for local staff to assume the coaching and support roles originally provided by outside consultants. One of the sites in the study (District B) has already moved to this phase, and the others are in the process of doing so, providing models of how the innovations can be independently continued. Thus the experiences of the schools in this study can help counter the fears that some have expressed regarding the way in which rural districts might meet legislated demands for increased student achievement. (See McNeil, 2009 for a discussion of these concerns.)

It should be stressed, however, that the most positive results of the implementations only occurred for students who were exposed to the full model – who began the curriculum in kindergarten and when their schools and teachers fully complied with the instructional tenets. This result replicates other studies that have found the greatest achievement gains for students occur with the highest fidelity to the Direct Instruction model (Benner, Nelson, Stage, & Ralston, 2010; Carlson & Francis, 2002; Gersten, Carnine, & Williams, 1982; Gersten, Carnine, Zoref, & Cronin, 1986; Ross et al., 2004). In addition, the results provide support for studies showing that achievement gains are greatest when the program has been fully stabilized with a school (Engelmann & Engelmann, 2004; Stockard, 2010). Thus, practitioners should be cautioned to exercise patience when implementing *Reading Mastery* and other Direct Instruction programs. Our results suggest that the programs can produce significantly higher reading achievement, but that these changes will most likely appear after teachers have fully learned the new curriculum and with students who are exposed to the full model, beginning their work with the program in the kindergarten years.

Our analysis had several advantages. For instance, we were able to look at reading achievement of students over multiple years and we replicated our findings across three different sites. We used several different statistical techniques, including multivariate analyses that controlled for differences between sites and students' at-risk status, obtaining equivalent results with all analyses. We were also able to compare data for students with varying levels of exposure to the program and to compare our results to data for national and state-wide samples.

At the same time, limitations of our study suggest directions for further research. Our analysis was limited to three districts in one state. Although the districts were in very different areas of a relatively large state, it would be important to replicate this study in other areas. Other studies should examine other measures of reading achievement. While our analysis of fourth graders' scores on the state assessment provided encouraging results, even longer-term analyses of students' performance would be important. In

addition, our national sample was limited to schools that participated in the DIBELS system and only dealt with one year of data. A broader national sample would be preferable. Other aspects of our results deserve additional attention. For instance, follow-up studies should examine the elements of the NIFDI model of support that contributed to the positive results. Such studies should try to identify the elements of the model that are most important and develop ways that others could incorporate it. In addition, the lagging performance of those with only partial exposure to the program should be examined to see if there might be ways to enhance their achievement even if they begin the program at later grades.

While our results involved only three districts in one state, they replicated findings that have been reported in numerous other settings regarding the efficacy of *Reading Mastery* and the ways in which the NIFDI model of implementation support can promote higher student achievement. The results also demonstrated the ways in which three small districts, all in relatively isolated regions of the country, could implement structured and explicit curricula and promote strong achievement gains that persisted through the early elementary years. We suggest that the example of these districts and the dedication and hard work of their teachers and administrators can begin to answer the call of Arnold and colleagues for ways to help rural schools improve teachers' "pedagogical skills in ways that have the greatest impact on student achievement" (Arnold, et al, 2005, p. 18).

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: The MIT Press.

Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle: Educational Achievement Systems.

American Federation of Teachers. (1998). *Building on the best, learning from what works: Seven promising reading and language arts programs*. Washington, D.C.: AFT.

Anderson, R. C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A.G. (1985). *Becoming a nation of readers: The report of the Commission on Reading*. Washington, D.C.: National Institute of Education.

Baker, S. K. Kameenui, E.J., Simmons, D.C., & Stahl, S.A. (1994). Beginning reading: Educational tools for diverse learners. *School Psychology Review, 23,* 372-391.

Arnold, M.L., Newman, J.H., Gaddy, B.B., & Dean, C.B. (2005). A look at the condition of rural education research: Setting a direction for future research. Journal of Research in Rural Education, 20 (6), 1-25.

Barley, Z.A., & Beesley, A.D. (2007). Rural school success: What can we learn? Journal of Research in Rural Education, 22 (1): 1-16.

Beck, I.L., & McCaslin, E.S. (1978). *An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs*. LRDC Report No. 1978/6 Pittsburgh.

Benner, G.J., Nelson, J.R., Stage, S.A., & Ralston, N.C. (2010). The influence of fidelity of implementation on the reading outcomes of middle school students experiencing reading difficulties. *Remedial and Special Education Online First*, published February 12, 2010.

Berends, M., Bodilly, S.J., & Kirby, S.N. (2002). *Facing the challenges of whole-school reform: New American schools after a decade*. Santa Monica, CA: Rand.

Blakeley, M. R. (2001). A survey of levels of supervisory support and maintenance of effects reported by educators involved in Direct Instruction implementations. *Journal of Direct Instruction*, 1 (2), 73-83.

Bodilly, S.J., Glennan, T.K., Jr., Galegher, J.R. & Kerr, K.A. (2004). Introduction: Framing the problem. In T.K. Glennan, Jr., S.J. Bodilly, J.R. Galegher, K.A. Kerr (Eds.). *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions, ,* (pp. 1-35). Santa Monica, CA: Rand.

Bodilly, S. J. (1998). *Lessons from New American Schools' Scale-up Phase*. Santa Monica, California: Rand.

Bond, G., & Dykstra, R. (1967). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly, 2,* 5-142.

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73(2), 125-230.

Carlson, C. D., & Francis, D.J. (2002). Increasing the reading achievement of at-risk children through Direct Instruction: Evaluation of the Rodeo Institute for Teacher Excellence (RITE). *Journal of Education for Students Placed at Risk, 7* (2), 141-166.

Chall, J.S. (1967). *Learning to read: The great debate.* New York: McGraw Hill.

Cicchinelli, L. F., & Dean, C. B. (2005, December 31). It's all about the quality of advice, guidance, and research for rural educators: A rejoinder to Howley,Theobald, and Howley. *Journal of Research in Rural Education, 20*(19)

Collins, M. & Carnine, D. (1988). Evaluating the field test revision process by comparing two versions of a reasoning skills CAI program. *Journal of Learning Disabilities, 21*, 375-379.

Cunningham, A.E. & Stanovich, K.E. (1990). Assessing print exposure and orthographic processing skill in children: A quick measure of reading experience. *Journal of Educational Psychology, 82,* 733-740.

Cunningham, A. E. & Stanovich, K. E. (1998). What reading does for the mind. *American Educato**r,*** (Spring/Summer), 1-8.

DIBELS (2008). DIBELS data system: Using data to improve achievement for each and all. https://dibels.uoregon.edu/, accessed December 17, 2008.

Engelmann, S. (2007). *Teaching needy kids in our backward system: 42 Years of trying*. Eugene, Oregon: ADI Press.

Engelmann, S.E. & Carnine, D. (1982). *Theory of instruction: Principles and applications*. New York: Irvington Publishers.

Engelmann, S.E. & Engelmann, K.E. (2004). Impediments to scaling up effective comprehensive school reform models. In T.K. Glennan, Jr., S.J. Bodilly, J.R. Galegher, K.A. Kerr (Eds.). *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions* (pp. 107-133). Santa Monica, CA: Rand.

Fixsen, D.L., Naoom, S.F., Blasé, K.A., Friedman, R.M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Foorman, B. R. (1995) Research on "the great debate": Code-oriented versus whole language approaches to reading instruction. *School Psychology Review, 24*, 376-392.

Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C., Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37-55.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fukkink, R.G., & deGlopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research, 68,* 450-469.

Gersten, R.M., Carnine, D.W., & Williams, P.B. (1982). Measuring implementation of a structured educational model in an urban school district: An observational approach. *Educational Evaluation and Policy Analysis 4(1)*, 67-79.

Gersten, R., Carnine, D., Zoref, L. & Cronin, D. (1986). A multifaceted study of change in seven inner-city schools. *The Elementary School Journal 86* (3), 257-276.

Gersten, R., Darch, C., & Gleason, M. (1988). Effectiveness of a Direct Instruction academic kindergarten for low-income students. *Elementary School Journal, 89*(2), 227-40.

Gough, P. B. & Juel, C. (1991). The first stages of word recognition. In: L. Rieben & C.A. Perfetti (eds.) *Learning to read: Basic research and its implications* (pp. 47-56). Hisllsdale, NJ: Erlbaum.

Grossen, B. (1997). *A synthesis of research on reading from the National Institute of Child Health and Human Development*. Eugene, Oregon: University of Oregon.Francis, et al, 1998;

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5* (3), 257-288.

Good, R. H., Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *Educational and Child Psychology, 15* (1), 56-70.

Good, R.H., Wallin, J., Simmons, D.C., Kame'enui, E. J., & Kaminski, R.A. (2002). System-wide percentile ranks for DIBELS benchmark assessment (Technical Report 9). Eugene, Oregon: University of Oregon.

Hasbrouck, J. & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59* (7), 636-644.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

Herman, R., Aladjam, D., McMahon, P., Masem, E., Mulligan, I., Smith, O., O'Malley, A., Quinones, S., Reeve, A., & Woodruff, D. (1999). *An educator's guide to schoolwide reform*. Washington, D.C.: American Institutes for Research.

Howley, C. B., Theobald, P., & Howley, A. A. (2005, December 31). What rural education research is of most worth? A reply to Arnold, Newman, Gaddy, and Dean. *Journal of Research in Rural Education, 20*(18).

Huitt, W.G., Monetti, D.M., & Hummel, J.H. (2009). Direct approach to instruction. In C.Reigeluth & A. Carr-Chellman (Eds.), *Instructional-Design Theories and Models: Volume III, Building a Common Knowledge Base* (pp. 73-98). Mahwah, NJ: Lawrence Erlbaum.

Institute of Child Health and Human Development (ICHHD). (2000). *Report of the national reading panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, D. C.: U. S. Government Printing Office.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80 (4),* 437-447.

Juel, C. & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly, 35,* 458-492.

Kaminski, R. A. & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25,* 215-227.

McNeil, M. (2009). Rural Areas Perceive Policy Tilt. Education Week, September 2, published online, August 28 (http:www.edweek.org/ew/articles/2009/09-02/02stim-rural.h29.html).

MetaMetrics (2009). *Linking DIBELS oral reading fluency with the Lexile framework for reading.* Durham, NC: MetaMetrics, Inc.

Mulkey, D. (1993). *Education in the rural south: Policy issues and research needs*. Mississippi State, MS: Sourthern Rural Development Center.

Murphy, Joseph (2004). *Leadership for Literacy: Research-Based Practice, PreK-3*. Sage: Corwin.

National Institute of Child Health and Human Development (NICHHD). (1996). Thirty years of NICHD research: What we now know about how children learn to read. *Effective School Practices, 15,* 33-46.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, D.C.: National Institute of Child Health and Human Development.

National Center for Education Statistics (2008)., Common Core of Data (CCD), "Local Education Agency Universe Survey," 1979-80 through 2006-07, NCES, Washington, D.C. Department of Education.

Pflaum, S., Walberg, H.J., Karigianes, M.L., & Rasher, S.P. (1980). Reading instruction: A quantitative analysis. *Educational Researcher, 9,* 12-18.

Ross, S.M., Nummery, J.A., Goldfeder, E., McDonald, A., Racho, R., Hornbeck, M., & Fleishman, S. (2004). Using school reform models to improve reading achievement: A longitudinal study of Direct Instruction and Success for All in an urban district. *Journal of Education for Students Placed at Risk, 9,* 357-368.

Sherwood, T. (2000). Where has all the "rural gone? Rural education research and current federal reform. *Journal of Research in Rural Education, 16* (Winter), 159-167.

Smith, S., Simmons, D., Gleason, M., Kameenui, E., Baker, S., Sprick, M., Gunn, B., & Thomas, C. (2001). An analysis of phonological awareness instruction in four kindergarten basal reading programs. *Reading and Writing Quarterly, 17,* 25-50.

Snider, V.E. (1990). Direct Instruction reading with average first graders. *Reading Improvement, 27,* 143-148.

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

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21* (4), 360-407.

Stanovich, K.E. (1994). Romance and reality. *The Reading Teacher, 47,* 280-291.

Stanovich, K.E. & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*, 402-433.
Stern, J.D. (1994). *The condition of education in rural schools*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.
Stern, J.D. (1995). Reflections of a recently retired federal analyst in rural education. *The Rural Education Newsletter*. Supplement (Spring), 3-6.
Stockard, J. (2010). Direct Instruction and First Grade Reading Achievement: The Role of Technical Support and Time of Implementation, Journal of Direct Instruction, forthcoming.

Appendix A

Timing of Implementations of *Reading Mastery* and NIFDI Support for Each District

As noted in the text, the three districts described in this report varied in the year in which they began to implement *Reading Mastery* and the grades exposed to the curriculum. District B began the curriculum under NIFDI's guidance in 2004-05 and began gathering DIBELS data in the previous academic year (2003-04). District C began its implementation in 2006-07 and began gathering DIBELS data two years earlier, in 2004-05. District A began using the DI curriculum in 2005-06 and gathering DIBELS data in 2004-05. However, this district did not fully comply with the NIFDI model until the 2007-08 year. (One of the key elements of good implementations is providing time for teachers to practice their teaching formats. Such practice allows the teachers to increase their presentation skills and give full attention to their students during lessons. District A's administrators did not allow time for this practice until the 2007-08 year, and thus this time-point is used to signify full implementation in the analysis.) Table A-1 provides details on the grouping assigned to each student cohort, the year they began kindergarten, and their exposure to the full curricular model.

Table A-1

Study Cohorts, Kindergarten Year and Grades Exposed to Curriculum by Study Site

District A			
Cohort Group	Kindergarten Year	Grades with Full Model	n
Full DI	2007-08	K to 1	149
Partial DI	2006-07	1 to 2*	168
Partial DI	2005-06	2 to 3*	193
Partial DI	2004-05	3 to 4**	216
District B			
Cohort Group	Kindergarten Year	Grades with Full Model	n
Full DI	2007-08	K to 1	183
Full DI	2006-07	K to 2	194
Full DI	2005-06	K to 3	165
Full DI	2004-05	K to 4	150
Partial DI	2003-2004	1 to 5	150
District C			
Cohort Group	Kindergarten Year	Grades with Full Model	n
Full	2007-08	K to 1	25
Full	2006-07	K to 2	21
Partial	2005-06	1 to 2	40
Partial	2004-05	1 to 3	35

*Students in these groups were exposed to *Reading Mastery* beginning in kindergarten, but the schools did not fully embrace the curriculum model until they began grade 1 (for those beginning kindergarten in 2006-07) or grade 2 (for those beginning kindergarten in 2005-06).

**Students in this group were exposed to *Reading Mastery* beginning in grade 1, but the schools did not fully implement the model until their third grade year.

Appendix B

Statistical Calculations

This appendix describes the way in which three statistical calculations given in the body of the text were computed: the effect sizes comparing district values to state values, effect size for change in the district over time relative to changes in the state, and t-tests examining the null hypothesis that changes in the district over time equaled those in the state. Table 10 provided data for these calculations and is reproduced below as Table B-1.

Table B-1

Percentage of Fourth Graders Meeting State Reading Standards, by Exposure to Curriculum, District B and State

Year	% - District	% - State	N	Z – score	Grades Exposed to RM	Type of Exposure
2004-05	79	85	117	-0.18	0	None
2005-06	87	88	147	-0.05	gr 3-4	Partial
2006-07	94	89	156	0.14	gr 2-4	Partial
2007-08	99	91	144	0.26	gr 1-4	Partial
2008-09	98	95	144	0.14	gr K-4	Full

Effect Sizes Comparing District Values and State Values

Educational researchers often use Cohen's d, a measure of effect size, to describe the magnitude of an effect. Cohen's d is calculated as

$(M1 - M2)/ s.d.$ (1)

Where Mi = mean of a group and s.d. = the common standard deviation.

The resulting value simply tells us the magnitude of the difference between two groups in standard deviation terms. A value of 1.0 indicates that the means differ by an entire standard deviation; a value of .50 indicates that they differ by one-half of a standard deviation.

While d is used to compare average values between two groups, z-scores are used to compare the scores of a sample with a population. The formula for a z-score is

$Z = (M1 - \mu)/ \sigma$ (2)

Where μ = the population mean and σ = the population standard deviation.

The resulting value tells us the magnitude of the difference between a sample and a population in standard deviation terms. As with Cohen's d, a z value of 1.0 indicates a difference of one standard deviation; a value of .50 indicates a difference of one-half of a standard deviation, etc..

The fifth column of Table B-1 reports z-scores comparing the results for the district to those for the state as a whole for each year. Because we wish to compare the values for the district (a sample) to the state as a whole (the population), we need to compute the standard deviation for the state. This can be done with the binomial distribution. Translating the percentages to proportions (by simply dividing by 100),

$$\sigma = \sqrt{pu * qu}, \quad (3)$$

where p_u = the proportion in the population, and $q_u = (1- p_u)$.

As an example, consider the computations for 2004-05. For this year, pu = .8487, qu = .1513.

$$\text{Thus } \sigma = \sqrt{(.85) * (.15)} = \sqrt{.1284} = .3583 \quad (4)$$

Substituting in the formula for standard (z) scores, where .7859 is the sample value M (or p_s) (for District B) and .8487 is μ (or p_u),the sample for the population,

$$z = (.7859 - .8487)/.3583 = -.0628/.3583 = -.1752. \quad (5)$$

This value (rounded to two significant digits) is given in the first line of Table B-1 and Table 4. It indicates that in 2004-05 the percentage of fourth graders in District B who met or exceeded state standards was .18 of a standard deviation less than the state. Similar calculations were completed for each year. Note that the z-scores are interpreted in precisely the same as one would use in comparing the averages or proportions in two independent samples – a difference in standard deviation terms. Thus, one can treat these z-scores as effect sizes; they are equivalent to Cohen's use of standard deviation units as an effect size for differences between means, but involve comparing a sample mean to a population mean.

Effect Sizes for Changes Over Time in a District Controlling for Changes in the State

While the results given in the top panel of Table B-1 provide a snapshot of achievement in each year relative to the state, the question of greater interest is the extent to which changes occurred over time in the district. To be most accurate we also need to control for changes that occurred within the state as a whole. Specifically, we want to know the extent to which a district's performance changed over time relative to the performance of the state.

A simple way to describe these changes is to compare the z-scores from one year to another. In other words, we can simply calculate the change in the standard deviation scores. Again, we build on the standard formula for Cohen's d, where

$$d = (M1 – M2)/ s.d. \quad (6)$$

Because the standard deviation for z-scores is, by definition, 1.0,

$$d_z = (Z1 – Z2). \quad (7)$$

Thus, the effect size that describes the change in a district relative to changes in the state can be calculated by comparing the z-scores using equation (2) above, for two different years.

Again the data in Table B-1 can illustrate. For District B in 2004-05, the z-score comparing the proportion of 4th graders meeting or exceeding standards with the state data was -.18, indicating that the proportion was .18 standard deviations below the proportion for the state as a whole. In 2009 the z score for this comparison was +.14, indicating that the proportion was .14 of a standard deviation higher than the score for the state.

The difference of these scores can be easily calculated

$d_z = .14 - (-.18) = +.32.$ (8)

From 2004-5 to 2008-9, the proportion of fourth graders meeting or exceeding state standards increased by .32 of a standard deviation relative to changes in the scores of fourth graders throughout the state.

One could, of course, simply look at the change in scores over time and compute an effect size with this information (using the formula $d = (M_{t1} - M_{t2})/s.d.$). This is the type of information one would get in a classic pretest posttest design. However, as is noted in classical experimental design theory, this design lacks internal validity because there is no control group – no indication of possible changes that might be occurring apart from an experimental intervention. Comparing the changes in the district with those in the state as a whole, as occurs with formula (7) provides a comparison to such a control group.

The data in Table B-2 illustrate the importance of including such a comparison. The data report the results of a simulation involving changes in achievement over time. We assume that the proportion of students meeting criteria in an imaginary district changed from .3 in year 1 to .7 in year 10. If only these data were considered, the effect size reflecting the change would be .87 (=(.7-.4)/.46 = .4/.46). This indicates a large change, one that would be considered educationally significant.

The data within the body of Table B-2 provide information on different possible changes within the state over the ten year time period. For instance, the first line reports the situation where .50 of the students in the state met criteria in Year 1 and the same proportion met criteria in Year 10. The third and fourth columns report the population standard deviation for years 1 and 10 (using equation (4) above); and the fifth and sixth columns give the z-scores comparing district values with the state values for each year, using the formula in equation (2) above. For Year 1, the district proportion of .3 compared to a state proportion of .5 results in a z-value of -.40; for year 10, the district proportion of .7 compared to a state proportion still at .5 results in a z-value of +.40. In other words, in year 1 the district proportion was .4 of a standard deviation below the state value, but in year 10 the value was .4 of a standard deviation above the state value. Using formula (7) above the effect size for the district change, while controlling for the state values $d_z = .40 - (-.40) = .80$. This value is quite close to the unadjusted value of .87, which would

be expected given the lack of change in the state. (See the last column of Table B-2, which simply reports the difference of dz and the value of d without controlling for changes in the state.)

Table B-2

Example of Calculating d with and without controlling for changes in the state

District Values: Year 1, p = .3; Year 10, p = .70; d without controlling for state changes = .87

State Year 1	State Year 10	s.d. yr. 1	s.d. yr. 10	z year 1	z year 10	d_z	Bias
0.50	0.50	0.50	0.50	-0.40	0.40	0.80	0.07
0.40	0.60	0.49	0.49	-0.20	0.20	0.41	0.46
0.30	0.70	0.46	0.46	0.00	0.00	0.00	0.87
0.20	0.80	0.40	0.40	0.25	-0.25	-0.50	1.37
0.60	0.40	0.49	0.49	-0.61	0.61	1.22	-0.35
0.70	0.30	0.46	0.46	-0.87	0.87	1.75	-0.88
0.80	0.20	0.40	0.40	-1.25	1.25	2.50	-1.63

The second, third and fourth lines of data depict a situation where there were positive changes within the state. The second line shows a change from .40 to .60 meeting criteria, slightly less than in the district; the third line has a change that exactly matches that within the district, from .30 to .70; and the fourth line depicts a situation with a larger change, .20 to .80. In other words, these are situations where the district improved over time, but so did all students within the state. In one situation the change was slightly greater in the district than in the state (line 2); in the next situation (line 3) the change was exactly the same as in the state; and in the third situation (line 4) the change in the district was actually less than in the state. Clearly the effect size of .87 does not accurately portray what really happened in the district *relative to what was happening in the state as a whole*. However, the effect size calculated with equation (7) (in the next to last column of the table) accurately reflects these changes. The d_z in line 2, with district changes that are slightly greater than in the state is .41, still positive but smaller than with the unadjusted value. The d_z in line three with changes equal to the state is, as one would expect, equal to zero. Finally, the d_z in line four, where the state had greater positive changes than the district, is negative. This is appropriate because the changes in the district, although positive, were less than in the state as a whole. Relative to the state as a whole the students in our hypothetical district lost ground over time.

The final three lines in Panel A depict a situation where the students in the state became less likely to meet criteria over time. In these situations the effect size that only considers district data (d = .87) underestimates the actual magnitude of change. For instance, the fifth line of data simulates a change in the state from .60 meeting criteria in Year 1 to .40 in year 10. The effect size when these state level changes are considered is 1.22, substantially larger than the value of .87 calculated without this control.

Similar results occur with other patterns of change. The essential point is that, if one wants to examine the amount of change relative to some type of larger comparison group, data for the comparison group need to be considered. Examining changes in the z-scores relative to this larger population provides a simple, accurate way to describe these changes.

Inferential Test to Examine Changes over Time in a District Controlling for Changes in the State

While the effect size computed above (d_z) provides a descriptive measure of the magnitude of change, practitioners, policy makers, and researchers are often interested in whether or not the changes might have occurred by chance. To answer this question researchers use simple hypothesis tests. For the data presented in this report we can use simple t-tests.[1]

The null hypothesis tested is that there was no change in a district's results, relative to the state, over time. In other words, the null hypothesis is that the z-scores in year a equal the z-scores in year b. If a curriculum had no effect we would expect that the z-scores relative to the state would be the same in both years. This is our null hypothesis: the difference between the two z-scores is zero.

H_o: $z_b - z_a = 0$. (9)

Alternatively, if there were an effect, we would expect there to be fewer students at risk or with a deficit and more students at low risk in the later years. (This is our alternative hypothesis.)

H_1: $Z_b - Z_a \neq 0$. (10)

To test this hypothesis we can do a simple comparison of means test using the t-distribution, treating the z-scores for each year as the means and using the standard formula for a t-test

$t = (M_2 - M_1)/s.e._{2\text{-}1}$ (11)

where s.e. = the standard error and the z-scores are treated as the means.

The standard error is simply a function of the standard error and the sample size.

$s.e. = \sqrt{[(s.d.1/n1) + (s.d.2/n2)]}$. (12)

By definition, the z-scores have standard deviations of 1.0.

Thus, $s.e. = \sqrt{[(1/n1) + (1/n2)]}$. (13)

To illustrate these calculations, consider the data for 2004-05 and 2008-09 from District B reported in Table 2 and reproduced in Table A-1 above. For 2004-05, n = 117 and the z-score comparing the value to the state = -.18; for 2008-09, n = 144, and the z-score = +.14.

$s.e. = \sqrt{(1/117) + (1/144)} = .12$. (14)

[1] The data presented in comparisons of fourth graders from one year to another, as in Table 4 in this paper, may be seen as independent. The vast majority of children in fourth grade in one year would not be in the grade in the next year.

$$t = (Z2 – Z1)/s.e. = (.14-(-.18)/.12 = 2.67. \quad (15)$$

The degrees of freedom associated with this test are

$$df = n1 + n2 – 2 = 117 + 144 – 2 = 259. \quad (16)$$

Using a standard t-table, it can be found that the probability of getting a t-value of 2.52 by chance with samples of this size is .01 (two-tail).