


The Impact of Administrative Decisions on Implementation Fidelity of Direct Instruction and Student Achievement

Learning Disability Quarterly
2020, Vol. 43(1) 18–28
© Hammill Institute on Disabilities 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0731948719830346
journals.sagepub.com/home/ldq


Jean Stockard, PhD¹

Abstract

This article examined the extent to which administrative decisions that affected the implementation fidelity of Direct Instruction programs were related to student achievement. Data from three studies showed that administrative decisions that disregarded recommended protocols regarding teacher training, teacher preparation, and student schedules were related to lower levels of teacher fidelity, student progress at mastery, and student achievement. Most results were statistically significant and had large associated effect sizes. Implications for practice and policy are discussed.

Keywords

implementation fidelity, administrators, student achievement, Direct Instruction

Studies of implementation fidelity often focus on the actions of the implementer, the extent to which a teacher or clinician conforms to particular requirements of a program. Yet, teachers and clinicians work within an organizational setting. This setting and those who administer it can greatly influence the extent to which teachers have the training and support needed to properly implement programs. Administrators also control schedules and the extent to which students receive the recommended dosage or exposure. This article reports results from three different studies that examined the way in which administrative decisions related to program implementation influenced student achievement. All of the studies involved schools that implemented the highly technical and effective Direct Instruction (DI) programs under the guidance of skilled consultants. However, for some time periods or in some settings, administrators did not follow the guidelines. This produced natural experiments in which it was possible to examine the extent to which not following established guidelines was related to student outcomes. While none of the analyses focused explicitly on students diagnosed as having learning disabilities, all involved students or schools deemed “high risk” and thus, no doubt, included many participants who would be considered as having such a diagnosis.

DI Theory and Methodology

DI programs, developed by Siegfried Engelmann and his collaborators beginning in the 1960s, are often cited as an example of explicit and systematic instruction. The instructional approach is overwhelmingly cited as effective for both

general populations and those needing special help (e.g., National Reading Panel, 2000). Although the term direct instruction (lower case and sometimes referred to as “little di”) has been used to refer to a broad set of educational programs that incorporate elements of systematic or explicit instruction, this article focuses on schools using programs within the Direct Instruction (capitalized) Engelmann–Becker tradition (S. Engelmann & Colvin, 2006). DI programs incorporate all of the elements deemed essential to systematic and explicit instruction, but in an integrated manner (S. Engelmann, 2004). They are often used for instruction of students with learning disabilities and others who are perceived to be at high risk for academic failure.

The theoretical foundation of DI is complex and well developed. (See S. Engelmann, 1999; S. Engelmann & Carnine, 1991, 2011; S. Engelmann & Steely, 2004, for detailed theoretical discussions; Barbash, 2012, for an accessible summary; and National Institute for Direct Instruction [NIFDI], 2016, for citations to 45 experimental examinations of the tenets.) It is based on the assumption that students use their inherent logical abilities to interpret instruction they receive. All students can learn if they are given well-designed instruction with totally clear and unambiguous examples. DI instructional materials are carefully designed and field tested. Scripts are provided to

¹University of Oregon, Eugene, USA

Corresponding Author:

Jean Stockard, Department of Planning, Public Policy, and Management,
University of Oregon, Eugene, OR 97301, USA.
Email: jeans@uoregon.edu

the teachers to ensure that they provide examples that are clear and ordered in a manner that produces the most effective and efficient learning. Both effective and efficient learning are especially important for students thought to have learning disabilities, for learning more in a shorter period of time is needed to help them catch up with their peers (see Note 1).

DI programs incorporate mastery learning. The DI programs are based on the assumption that students learn most quickly when they have the prerequisite skills and knowledge. Each lesson builds on the previous lessons with only 10% to 15% new material introduced in each lesson. Teachers are instructed to check that students have mastered all prerequisite knowledge necessary for learning new things. These design elements make instruction more effective and more efficient. Teachers are also required to test students' skills to determine where to place them within the program and, throughout the year, to retest to make sure they are at the appropriate point in the curriculum for the greatest possible progress. Appropriate placement makes learning more efficient and more enjoyable for students as they are constantly learning and progressing without the material being too difficult or too easy (S. Engelmann, 2014).

Because DI lessons are so carefully sequenced, students' progress through the programs can be calculated as a continuous measure, commonly called "lesson progress at mastery." Benchmarks indicate where students should be at various points throughout the year to be at grade level or, for those who are behind, to eventually catch up with peers. Studies have documented the validity of these benchmarks, showing, for instance, that students who are at or near grade level in their DI programs are much more likely than other students to score at or above the national mean on standardized achievement tests and at the proficient and advanced level on state assessments (Stockard, 2014).

The Effectiveness of DI

When examining the impact of implementation fidelity of a program, it is important that the program involved has been found to be effective. If a program were highly effective yet implemented poorly, one would expect that outcomes would be less positive. However, if a program were not effective, poor implementation could actually result in better outcomes. In other words, a good program that is not done well would have poor outcomes. But a poor program that is not done well could actually have good outcomes simply because, to put it colloquially, "it couldn't get worse" (Stockard, 2010). A large literature indicates that DI programs are highly effective, thus supporting the decision to focus on those programs.

Coughlin (2014) summarized the results of six systematic reviews and seven meta-analyses of DI, all of

which found strong evidence of effectiveness. The positive results appeared with reviews of specific programs including mathematics, reading, and spelling. They appeared in studies of whole school reform projects, students in both general education and special education, and in studies with variety of research designs. A narrative review of studies specifically focused on students with learning disabilities or other special classifications also reported strong evidence of effectiveness (Wood & Stockard, 2012).

The largest meta-analysis of DI's effectiveness examined 328 studies published from 1966 through 2016 and incorporated almost 4,000 effects (Stockard, Wood, Coughlin, & Khoury, 2018). Overall, estimates of effect size (Cohen's *d*) were large, ranging from 0.52 to 0.60. Effects for single-subject designs, studies that were more likely to focus on students with learning disabilities, ranged from 0.83 to 1.02. Twenty-nine percent of the studies involved students who had some type of at-risk status, such as a diagnosis of learning disability, and the multivariate analyses indicated no significant differences in the impact of DI by special status.

Fidelity of Implementation and DI

Even though the research literature indicates that DI programs are highly effective, some DI schools and classrooms are more successful than others. A number of studies have shown that a major reason for these differences is implementation fidelity, the extent to which teachers administer the programs as they were designed (Benner, Nelson, Stage, & Ralston, 2010; Carlson & Francis, 2002; Gersten, Carnine, & Williams, 1982; Gersten, Carnine, Zoref, & Cronin, 1986; Stockard, 2011). Students of teachers who follow the programs' protocols more faithfully have higher achievement scores and greater growth.

Kurt Engelmann (2014) outlined principles that govern the successful implementation of DI programs. He described the important role of teachers' instructional actions including the extent to which they follow program guidelines, ensure student mastery, and provide consistent and visible reinforcement. He also explicitly noted the necessity of organizational support:

Teachers' implementation of the DI approach (the instructional level) is highly successful only when the school's leadership team provides an environment in which effective and efficient instruction can take place. School and district leaders are responsible for establishing the structural components needed for a successful DI implementation, training staff to implement the program properly, monitoring instruction to ensure that the program is implemented with fidelity, and increasing the capacity of the school and district to support the model fully. (Engelmann, 2014, pp. 108–109)

Engelmann (2014) stressed the wide range of factors that can affect student performance, noting that they can encompass “everything in a school or district’s control, including the daily schedule, the assignment of personnel, the professional development of staff, the physical arrangement of classrooms and the public announcement system” (p. 101).

Summary

The analyses in this article are based on K. Engelmann’s formulation of the ways in which administrative actions and decisions can influence successful implementation of DI programs and student success. The logic can be conceived as a causal chain. A supportive and knowledgeable organizational environment promotes stronger implementation fidelity by teachers. In turn, teachers who are more skilled at implementing the program with fidelity have students who make greater progress through the curriculum at mastery, and this greater progress at mastery results in higher achievement scores. In other words, teachers in a more supportive and knowledgeable environment are more likely to become skilled DI teachers. Their students are then more likely to master the material and progress through the programs more quickly. As a result, they have higher achievement. Most important, strong organizational and administrative support is necessary for other elements of the chain to occur (K. Engelmann, 2014).

The following sections of this article present results from three different studies that address this logic. Each of the studies examined the extent to which administrative decisions affect student achievement. The first two involved settings in which some students were taught by teachers who had no training (Study 1) or less than optimal conditions for daily teacher preparation (Study 2). The other study involved an issue related to scheduling, providing instruction on a regular basis (Study 3). All schools received training and implementation support from highly regarded organizations. While none of the data sets provided indications of whether students had been labeled as learning disabled, all settings involved students deemed “high risk,” and one (Study 3) involved students receiving special education. Two of the studies involved reading programs and one involved math. As explained more fully in the Discussion section, it would no doubt be unethical to knowingly expose students to a less than optimal situation. The studies described below take advantage of data from real-life situations that approximate natural experiments, contrasting problematic implementations with better ones. To test the causal relationships posited above, multivariate analysis techniques are used in each study. Additional statistical results for Studies 2 and 3, all of which support the findings summarized below, are in a supplementary document.

Study 1: Teacher Training and Assignment

Given the highly technical nature of DI, training in the appropriate use of the programs is cited as a very important factor in promoting high fidelity. One of the most important decisions school principals can make is the assignment of teachers to grade levels and instructional groups for which they have been properly trained. As K. Engelmann (2014) put it, “teaching staff should receive a thorough preservice training before the start of the school year in the precise levels of the programs that correspond to their student’s mastery level” (p. 112). Study 1 examined the extent to which violating this recommendation can affect student learning.

Method

Setting, sample, and design. Study 1 used data from a high poverty school in the Southeastern United States. (Over 70% of the students received free or reduced lunch [FRL].) In response to concerns about the very low reading skills and achievement of its students, the school implemented the DI program *Reading Mastery Signature Edition* (RMSE) in some kindergarten and first-grade classrooms in the fall of 2013. Given the positive results in that year, they opted to use the program for reading instruction in all K–2 classrooms in 2014–2015. However, at the start of the school year, the principal unexpectedly re-assigned some teachers from upper grade to lower grade classrooms. This resulted in one kindergarten teacher and one second-grade teacher having no prior training in the DI program. It also resulted in a natural experiment in which achievement of students who had a trained teacher could be compared with the achievement of students who did not while controlling for levels of prior achievement. Thus, Study 1 used a pretest–posttest control group design with statistical controls.

The sample used for analysis was limited to students with data on all variables: 65 kindergarten students—37 with a trained teacher and 28 with an untrained teacher; and 43 second-grade students, 36 with trained teachers and seven with an untrained teacher. During the school year, all teachers, whether or not they had prior training, received regular coaching and implementation support. While the untrained kindergarten teacher continued using DI throughout the school year, the untrained second-grade teacher stopped the program at mid-year.

Measures. For kindergarten students, achievement was measured at fall, winter, and spring with the i-Ready reading assessment (“i-Ready K-12,” 2018), which was used by the district in which Study 1 occurred (see Note 2). Scores on a standardized test were not available for second graders, so achievement was measured by students’ placement at

mastery in their DI reading program at fall and winter. As noted above, this measure has been found to be highly correlated with scores on standardized achievement tests and state assessments (Stockard, 2014). To examine the causal logic outlined above, the analysis also included, for kindergarten students, an indicator of their lesson progress at mastery in the spring. The achievement scores were gathered by school personnel using their regular procedures. The indicators of lesson mastery were made in conjunction with the trained implementation personnel, helping to ensure reliability and validity.

Analysis. Data were examined separately for each grade using bivariate and multivariate analyses. First, the average scores on the measures of achievement and lesson mastery of students with and without trained teachers were compared using *t* tests and effect sizes (Cohen's *d*). One-tail tests of significance were used in all analyses given the directional hypotheses noted above. Effect sizes were calculated, so that a positive value indicates support for the hypotheses (greater implementation fidelity associated with higher achievement). The psychological literature has traditionally interpreted an effect size of .20 as small, .50 as medium, and .80 and greater as large (Cohen, 1988). Educational researchers have traditionally used the threshold of .25 to indicate an educationally significant effect size (Tallmadge, 1977). Recently, however, Lipsey and associates (2012), after examining the distribution of effect sizes from studies of a wide range of educational interventions, concluded that an effect size of .25 should be considered large and that one of .50 would be "more like 'huge'" (p. 4).

Second, post-test measures of achievement were regressed on fall measures to examine the extent to which having a trained teacher was independent of, or provided "added value" to, the impact of prior achievement. For kindergarten students, a second model added the measure of students' progress through the reading program during the school year. Given the discussion above, it was expected that students whose teachers had received training prior to the start of school would have higher scores on the achievement measures even when prior achievement was controlled. It was also expected that, for kindergarten students, when the measure of lesson progress was added to the model, the impact of teacher training would decline markedly. In other words as posited earlier, the reason that students of trained teachers had higher achievement was that they progressed more quickly through the programs at mastery.

Results

Table 1 shows the descriptive statistics on key variables for students with trained and untrained teachers and the

associated *t* ratios and effect sizes. As expected, the results indicate significantly stronger growth for students who had a trained teacher, and the associated effect sizes (.46 for kindergarten and .81 for second graders) would be considered moderate to large using the criteria from the psychological literature and "huge" by educational research standards (Lipsey et al., 2012, p. 4). Kindergarten students with the trained teachers began the year with lower i-Ready scores than those with the untrained teacher, but by spring, they had higher scores and had completed more lessons at mastery. Second-grade students with the trained teacher began the year with higher placements in RMSE than students with the untrained teachers, and the gap between the two groups widened over time as the students with the trained teacher had significantly greater growth.

Regression results for kindergarten students are in the top panel of Table 2 and those for second graders are in the bottom panel. As expected, in both grades, students with a trained teacher had significantly higher post-test achievement scores when pretest (fall) scores were controlled. In addition, results with Model 2 for kindergarten students confirm the expectation that much of the impact of having a trained teacher involves the rate of student progress through the programs. The coefficient associated with teacher training declined markedly when student progress at mastery was entered into the equation while the coefficient associated with lesson progress was highly significant.

Summary

Results from Study 1 support the expectations implied by the causal chain described above, showing that students of teachers who have been trained were more likely to master material and score higher on achievement tests. These results persisted when controls were introduced for prior levels of achievement. The administrative decision to put untrained teachers in the classroom had a negative effect on student achievement, and part of this effect was due to students' lower rate of progress through the curriculum when taught by an untrained teacher.

Study 2: Providing Time for Teacher Development

Teaching DI programs is technical and involved. It requires not only training, but practice and careful preparation for each lesson. One of the most important elements is regular rehearsal of lessons. Such practice helps teachers learn to present the material easily and fluently so that they can give full attention to their students during the lessons. Thus, one of the key elements of good implementations is providing time for teachers to practice their teaching formats. Study 2

Table 1. Average Values, *t* Tests and Effect Sizes, Reading Achievement, Lesson Progress, and Previous DI Exposure, by Training of Teacher and Grade, Study 1.

Variable	Untrained teachers		Trained teachers		<i>t</i>	Cohen's <i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Kindergarten						
Fall i-Ready	337.9	29.1	334.8	25.9	0.44	-0.11
Winter i-Ready	363.6	29.1	357.9	25.9	0.83	-0.21
Spring i-Ready	394.0	34.0	404.4	32.7	1.25	0.31
Growth, Fall-Spring	56.1	22.5	69.5	32.3	1.88*	0.46
Spring Placement, RMSE	129.2	63.3	154.9	37.4	2.05*	0.50
Second grade						
Fall Placement, RMSE	215.0	0.0	245.0	65.0	1.21	0.50
Winter Placement, RMSE	274.1	57.1	353.5	24.8	6.08***	1.84
Growth	59.1	57.1	108.6	59.1	2.03*	0.81

Note. Probability levels are one-tailed to reflect the directional hypotheses. Degrees of freedom for the analysis of kindergarten students was 63 and for second-grade students was 41. DI = Direct Instruction; RMSE = Reading Mastery Signature Edition.

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

Table 2. Regression Analyses, by Grade Level, Study 1.

Kindergarten students, dependent variable is spring i-Ready scores

Independent variables	Model 1			Model 2		
	<i>b</i>	<i>t</i>	Probability	<i>b</i>	<i>t</i>	Probability
Trained teacher	12.51	1.82	0.037	3.43	0.53	0.30
Fall i-Ready scores	0.70	5.50	<.001	0.36	2.60	0.006
Lesson progress	—	—	—	0.31	4.13	<.001
Constant	158.7	3.68	<.001	231.6	5.48	<.001
R^2	.34, $F(2, 62) = 16.26, p < .001$.49, $F(3, 61) = 19.35, p < .001$		

Second graders, dependent variable is winter lesson placement

Independent variables	<i>b</i>	<i>t</i>	Probability
Trained teacher	74.59	5.82	<.001
Fall placement	0.16	2.02	0.026
Constant	239.78	11.65	<.001
R^2	.52, $F(2, 40) = 21.87, p < .001$		

Note. Probability levels for regression coefficients are one-tailed to reflect the directional hypotheses.

examined the extent to which student achievement was related to providing teachers this practice time.

Method

Setting, sample, and procedures. Study 2 used data provided by a school district located in the rural Midwest. For the first 2 years that the schools used the DI curriculum they did not provide time for teachers to practice their teaching formats but then, in later years, scheduled regular time for practice and preparation as recommended by the developer.

Thus, Study 2 used a cohort comparison design, comparing data for three cohorts: (a) those who began kindergarten before the schools were using DI ($n = 166$), (b) those who began kindergarten when their teachers were using DI but were not given practice time ($n = 299$), and (c) those who began kindergarten when the teachers were given time to practice and become more fluent in their presentations ($n = 142$). Cohort comparison designs are recommended by the classic methodological literature as especially appropriate for institutional settings such as schools (Shadish, Cook, & Campbell, 2002, see also Stockard, 2013).

Table 3. Best Fitting Linear Growth Models, Nonsense Word Fluency, Mid-K to Fall Grade 2 Regressed on Time, At-Risk Status, Implementation Fidelity, and Interactions.

Fixed effect coefficients and model fit statistics	Model 1		Model 2	
	<i>b</i>	SE	<i>b</i>	SE
Fixed effect coefficients				
Time	6.50***	0.25	6.91***	0.26
At-risk status	-5.74*	2.52	-7.53**	2.72
DI, Teacher Practice	12.09***	2.76	11.37***	2.87
DI, No Teacher Practice	9.39***	2.91	5.22 [†]	3.21
DI, Teacher Practice × Time	1.71***	0.50	1.81***	0.52
At Risk × DI, No Teacher Practice	-7.40*	3.54	-4.03	3.82
Teacher 2—K	—	—	-5.53*	2.57
Teacher 3—K	—	—	-6.99**	2.57
Teacher 8—K	—	—	-5.76*	2.58
Constant	31.54***	2.35	37.69***	2.81
Model fit statistics				
LL	20,987.3		18,257.5	
Change in -2LL	895.2***		2,729.8***	
<i>df</i>	6		3	

Note. Analysis based on 2,282 observations, 607 students, one to five observations per student, average 3.8. The -2 LL value for the baseline model was 21,882.5. The -2 LL value for Model 1 is compared to baseline; the -2LL value for Model 2 is compared to Model 1. Time is coded with the first period = 0. Further details on the analysis including results for other models are in the supplemental material, Tables S1 to S3. DI = Direct Instruction; LL = log likelihood.

[†]*p* < .10. **p* < .05. ***p* < .01. ****p* < .001.

Measures. Student achievement was measured with the *Dynamic Indicators of Basic Early Literacy* (DIBELS) measure of *Nonsense Word Fluency* (NWF; DIBELS, 2008). A substantial literature has shown that this measure is a valid predictor of later reading skills (e.g., Burke, Hagan-Burke, Kwok, & Parker, 2008; Fien et al., 2008; Vanderwood, Linklater, & Healy, 2008). The data were gathered each year using the district's regularly established procedures, and there was no reason to believe that the data gathering or test administration procedures varied for the cohorts included in the analysis.

Students' at-risk status was measured by demographic variables: receipt of FRL and minority status. Almost half (47%) of the students received FRL and close to a third (31%) were minorities, primarily Hispanic. Students with a minority status were significantly more likely to also receive FRL ($\chi^2 = 168.04$, $df = 2$, $p < .001$). Because this association was so strong, the demographic indicators were combined into a single dummy variable with a code of 1 if the student was a member of a minority group and/or received FRL (52% of the sample). There was no difference between the cohorts in the percentage classified as "at risk" based on these demographic characteristics ($\chi^2 = .48$, $df = 1$, $p = .79$).

Information was available on the students' teachers at each grade. Dummy variables were constructed for each teacher, with the first teacher in the alphabetic list as the reference category. Dummy variables were also used to denote which cohort a student was in, with no exposure to DI as the reference category.

Analysis. Linear growth models were used to examine variations in changes over time in NWF scores of students in these three groups from the spring of kindergarten through fall of second grade. Main and interaction effects for at-risk status and teacher were included in the models. Based on the theoretical discussion above, it was expected that students in cohorts whose teachers had been given time to practice their presentations would have the greatest growth over time and that this relationship would persist when strong controls were entered for teacher effects. (Descriptive statistics, ANOVA results, effect sizes, model fit statistics, and fixed effect coefficients for all models examined are in the supplementary material.)

Results

Table 3 reports results with the two best-fitting growth models. Model 1 includes time, at-risk status, cohort, and significant interaction effects. Model 2 adds dummy variables associated with the three teacher effects that were significant. Results were as hypothesized. The largest changes over time occurred for DI students whose teachers had time to practice, and the smallest changes occurred for the students who did not have DI. Growth over time was significantly greater for DI students whose teachers had time to practice. Growth was significantly smaller for DI students whose teachers did not practice, but this interaction effect declined slightly when the dummy variables for teachers were added in Model 2.

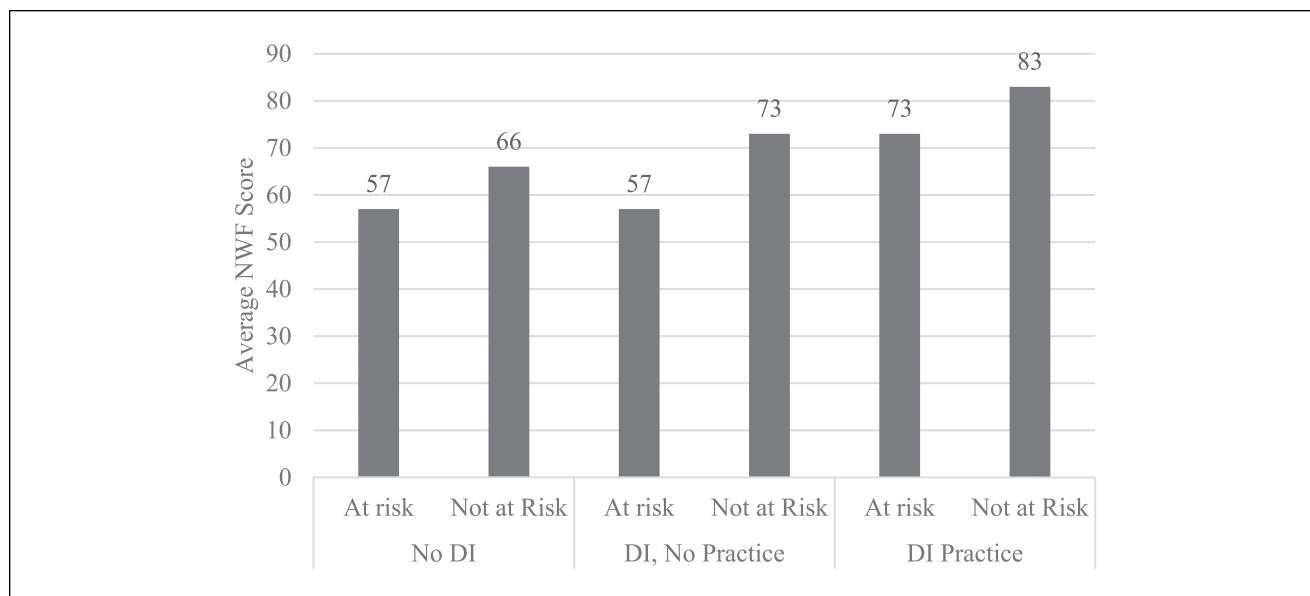


Figure 1. Average NWF scores, fall, Grade 2, by cohort and at-risk status.

Note. NWF = Nonsense Word Fluency.

Inspection of average scores for students in each cohort and risk category illustrate the results. Figure 1 shows the average NWF values of students by cohort and at-risk status at the start of second grade, the last point in the growth analysis. The at-risk students in the non-DI group and the DI group whose teachers did not practice had the lowest average scores. The students in the DI group whose teachers practiced presentations had the highest scores. The average scores of the at-risk DI students whose teachers had time to practice were next highest and equal to those of the not-at-risk students of DI teachers who did not practice. (Tables S1 and S2 in the supplemental material give additional details on average values at all time points.)

Summary

Results of Study 2 also supported the expectations outlined in the introduction. While both cohorts that were instructed with DI had higher achievement scores than those who did not, significantly higher scores and greater rates of growth were obtained when administrators provided time for teachers to practice their teaching presentations so they could be more fluent and attentive to student needs. Results persisted with strong statistical controls for teacher effects and indicated especially strong impacts for students deemed at risk based on demographic characteristics.

Study 3: Scheduling Instructional Time

In addition to determining teacher assignments and the time teachers have for preparation, administrators control

school schedules and the time that can be devoted to teaching. In other words, administrative decisions can influence the extent to which students are exposed to instruction. The implementation guidelines for DI programs stress the importance of maintaining a regular schedule of teaching. Students learn the most when they are regularly exposed to the material for the recommended amount of time each day. In addition, it would be expected that teachers who regularly implement the program would, assuming proper training and coaching, become more skilled in their implementation skills and exhibit higher fidelity in teaching practices (K. Engelmann, 2014). Study 3 examined the relationship of school schedules to teacher fidelity and student achievement.

Method

Sample, setting, and procedures. Study 3 used data from students receiving special education services in 13 schools in the upper Midwest. The teachers began using the DI mathematics program *Connecting Math Concepts: Comprehensive Edition* (CMCCE) in the fall of 2014. All of the teachers involved had the same initial training and support throughout the school year. Some schools fully implemented the program, regularly using it each day for the specified amount of time. Others, however, partially implemented the program, using it only some days of the week and with a variable schedule. Students were in kindergarten to Grade 6, although the majority were in Grades 2 to 4. There were 83 students in the schools with a regular schedule and 40 students in the schools with only intermittent exposure.

Achievement data were gathered in the fall and spring for each student. Thus, a pretest–posttest control group design with statistical controls was used, comparing the achievement growth of students with and without regular exposure to *CMCCE*. Assessment data were gathered through the regular procedures in place at each school. Measures of teacher fidelity were gathered by trained supervisors from a regional consortium. Data on fidelity were available for teachers of 91 students, 65 at schools with a regular schedule and 18 at schools with intermittent exposure.

Measures. Mathematics achievement was assessed with the nationally normed *Measures of Academic Progress* (MAP) from the *Northwest Evaluation Association* (NWEA; 2011) using Rasch Unit Scale (RIT) scores. These scores are measured on an interval scale and cumulative in nature. *Z* scores were calculated for each student by comparing their RIT scores with the average values for students in the norming population for their grade and testing period. Scores for fall and spring, as well as change in these scores over the school year, were examined. Positive change scores indicate improvement over time relative to the national norm while negative scores indicate a decline. (See supplemental material for additional details.)

The measure of teacher fidelity was the average value obtained from multiple observations during the school year, given as a percentage, with a score of 100 indicating full fidelity to the areas examined. The scale focused on instructional techniques and did not include measures related to organizational elements, such as school scheduling and exposure. Teachers were observed from 1 to 4 times (average 2.3) during the year.

Analysis. The analysis first focused on differences in achievement growth and teacher fidelity between the two groups of schools using *t* tests and effect sizes. It was expected that students with irregular exposure to the program would have less growth in mathematics achievement and that their teachers would have lower fidelity. Second, correlation coefficients and regression techniques were used to examine the relationship of changes in student achievement to scheduling and teacher fidelity. It was expected that the relationship of teacher fidelity with achievement would be greatly diminished when exposure through regular scheduling was controlled.

Results

As shown in Table 4, students in the schools using the recommended schedules had slightly lower average fall achievement scores than the other students (Cohen's $d = -0.20$). But, as expected, the students regularly exposed to the program had significantly higher levels of growth during the year ($d = 0.38$) and significantly higher achievement scores in the spring ($d = 0.54$). In addition, as

Table 4. Average Values Achievement Scores, Fall, Spring, and Growth; Teacher Fidelity, by Scheduling, *t* Tests and Effect Sizes, Study 3.

Variable	Full schedule	Intermittent schedule
Fall Z score		
<i>M</i> (<i>SD</i>)	-0.43 (1.33)	-0.18 (1.10)
<i>n</i>	83	40
<i>t</i> ratio = -1.04, probability = .15, Cohen's $d = -0.20$		
Spring Z Score		
<i>M</i> (<i>SD</i>)	-0.40 (1.19)	-0.88 (1.32)
<i>n</i>	83	40
<i>t</i> ratio = 2.03, probability = .02, Cohen's $d = 0.38$		
Growth fall to spring		
<i>M</i> (<i>SD</i>)	10.93 (12.35)	4.45 (9.72)
<i>n</i>	83	40
<i>t</i> ratio = 2.91, probability = .002, Cohen's $d = 0.54$		
Teacher fidelity score		
<i>M</i> (<i>SD</i>)	85.51 (5.08)	76.19 (6.85)
<i>n</i>	65	26
<i>t</i> ratio = 7.13, probability < .0001, Cohen's $d = 1.33$		

expected, their teachers exhibited significantly higher fidelity ($d = 1.33$). There were no significant differences in growth scores for students with and without teacher fidelity measures ($t = .26$, $df = 121$, $p = .79$) nor any association between the presence of fidelity measures and representation in schools with or without continuous exposure ($\chi^2 = 2.48$, $df = 1$, $p = .11$).

Table 5 reports the results of the correlation and regression analyses. The first column reports the unstandardized regression coefficients (*b*); the second column gives the standardized coefficients (beta); the third column reports the standard errors, used to calculate the level of significance; and the last column gives the zero-order correlations (*r*) between achievement growth and scheduling and fidelity. As expected, both of the zero-order correlations were positive and statistically significant, indicating that students had higher achievement when their teachers adhered more fully to the instructional model and when they were regularly exposed. However, the regression results indicate that, once scheduling was controlled, teachers' fidelity had no significant relationship with student growth. In other words, whether or not teachers exhibited better teaching skills, being in a setting that promoted regular and consistent exposure to the program was a far more important influence on student achievement.

Correlations and standardized regression coefficients are both in standard deviation terms and can be directly compared and interpreted as effect sizes. It can be seen that the effect of having a teacher with high fidelity dropped by more than half when in an environment that did not have a regular schedule (from .34 to .13). The R^2 value indicates that the measures of exposure and fidelity accounted for almost 20% of the variance in student achievement growth.

Table 5. Regression of Change in Achievement Score Fall to Spring on Teacher Fidelity and Regular Versus Intermittent Schedule, Study 3.

Independent variables	<i>b</i>	Beta	<i>SE</i>	<i>r</i>
Regular schedule	1.35**	0.34	0.48	.32***
Teacher fidelity	0.03	0.13	0.03	.34***
Constant	-2.77		2.65	—
<i>R</i> ²	.19***			

Note. Correlation of fidelity and having a regular schedule was .60, $p < .001$.

** $p < .01$. *** $p < .001$.

Summary

The results of Study 3 supported expectations. Students in settings with regular and consistent exposure had significantly larger gains in achievement than students with inconsistent schedules. On average, the achievement scores of students in schools with inconsistent schedules fell relative to the national norm over time. While zero-order correlations showed a significant relationship between teacher fidelity and student growth, this relationship was insignificant when scheduling was controlled. Individual teacher fidelity only promoted student achievement when the organizational environment allowed regular scheduling and exposure, again supporting the causal chain posited above.

Discussion

A central goal of educators working with students with learning disabilities is promoting higher academic achievement so they can catch up with their peers. However, there is variability in student achievement growth, even among those using the same instructional programs, and poor implementation fidelity by instructors is often cited as a reason for these differences. Yet, teachers work within organizational settings. The constraints of these organizations and, especially, decisions by their administrators can influence the extent to which teachers exhibit implementation fidelity and, consequently, their students' achievement. This article examined this relationship with three studies involving the highly effective and technically demanding DI programs. It was hypothesized that students would have higher achievement when they were in schools where administrators followed recommended guidelines regarding teacher support and student scheduling because such support promotes teacher implementation fidelity and students' progress at mastery through the programs.

Results of the three studies provided consistent support for these expectations. Students had higher achievement when they had teachers who had received appropriate training (Study 1) and had been given the recommended preparation time (Study 2), and when they received instruction for the recommended amount of time (Study 3). This organizational and administrative support enhanced teachers'

fidelity to the instructional program (Study 3) and students' progress at mastery through the programs (Study 1) even when the effects of individual teachers were controlled (Study 2), thus supporting the hypothesized causal chain. In addition, multivariate analyses showed that the potential impact of teachers' instructional fidelity to the programs could be negated by a non-supportive environment (Study 3). The results appeared with different study designs, measures, and community settings. Most results were statistically significant and had large associated effect sizes.

Limitations and Implications for Research

Each of the studies reported above had methodological limitations. Sample sizes could have been larger, measures more extensive, and the time periods longer. Given the large effect sizes for DI programs discussed in the introduction, no ethical researcher would purposely design research such as that reported here by assigning students to less than optimal circumstances, nor try to extend unfavorable circumstances just to enhance adherence to various methodological criteria. Instead, by necessity, each of the studies occurred within natural settings, using the designs and measures that were available. That said, each study used designs that are recommended as internally valid and especially appropriate for natural, field settings (Stockard, 2013; Stockard et al., 2018) as well as measures commonly used in schools and research. For example, cohort comparison designs benefit from examining the same students, teachers, administrators, and school environments, something that often does not occur with other comparative studies. Moreover, one could argue that the studies embody a great deal of external validity, involving settings similar to those found in other schools and districts and measures that are commonly used and administered within schools.

Even though ethical researchers would not purposely design replications of the work reported here, they could replicate the approach used in this article to help understand why results within a given setting were not as strong as the research literature would suggest. As noted above, improper implementations would result in lowered effectiveness only with programs that are actually effective (Stockard, 2010). For programs such as DI, where there is substantial documentation of their effectiveness, researchers and schools would be well served by examining why changes in effectiveness were less than expected by the literature. It would be reasonable to hypothesize that a lack of achievement gains was related to poor instructional fidelity and failing to provide full and appropriate organizational support and conditions. Meta-analyses of the DI effectiveness literature have also found significant positive impacts on student and teacher views (Stockard et al., 2018). When implementations do not result in such positive attitudes, the role of implementation fidelity and associated lower achievement should also be examined. It could be expected that implementations with greater fidelity would result in both higher

student achievement and more positive student and teacher views. In general, negative results are important results, especially in the context of a large body of research, and deserve further examination to determine why they occurred.

Implications for Practitioners

There are a number of implications of this work for practitioners. The first parallels suggestions for researchers. Practitioners are often urged to adopt evidence-based curricula, and it seems reasonable to assume that those responsible for the choice have some familiarity with the associated effectiveness literature. Schools also obtain measures of their students' progress, whether through curriculum-based measures, standardized achievement tests, or state assessments. Thus, it is also reasonable to assume that administrators and teachers examine the extent to which their students' skills are improving. User-friendly procedures are available for determining the effect sizes associated with curriculum changes (e.g., NIFDI, n.d.; Stockard, 2013), and these should be compared with the literature to determine if the changes found within a given school equal those found in the research literature. If they do not, practitioners should examine why the discrepancies occurred and especially focus on the role of suboptimal instructional fidelity and organizational support.

Second, practitioners should avoid the peril of "satisficing," or choosing an acceptable solution even though substantially better outcomes are possible (Simon, 1947, 1956). One of the studies described above (Study 2) included comparisons to students in non-DI programs and found that students with suboptimal implementations of DI had higher achievement than students using other programs, but lower achievement than implementations with higher fidelity. It is possible that schools could simply settle for the suboptimal DI implementation because it was better than the non-DI alternative. However, if they had compared their results with the research literature, they would understand that it was possible to do even better.

Finally, it is important to stress that this article only looked at three elements of DI implementations. In reality, successful implementations of the program involve a complex interplay of many different elements for teachers in their classrooms as well as administrators at both the building and district levels. Each of the separate elements, such as those examined in this article, may seem small in isolation, but each one is important (K. Engelmann, 2014; S. Engelmann, 2014; S. E. Engelmann & Engelmann, 2004). While developing school practices and norms that support effective implementations can take time and effort, the potential for significantly helping all children succeed is undoubtedly worth the investment.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this

article: Preliminary analysis of the data examined in this paper was conducted while the author was employed, on a part-time basis, as a researcher by the National Institute for Direct Instruction (NIFDI), which provides technical support to schools implementing DI programs.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preliminary work on the analysis reported in this paper was conducted while the author was employed, on a part-time basis by the National Institute for Direct Instruction (NIFDI).

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The Direct Instruction (DI) tradition would implicitly reject applying a label of "learning disabled" to a student who has not learned, arguing that, instead, the instruction has been faulty or ineffective.
2. It should be noted that the i-Ready assessment has questionable validity for beginning readers taught with DI. The Reading Mastery program does not teach letter names until well into the kindergarten-level program and after children have already developed many reading skills. Yet, the computer algorithms within the i-Ready assessment do not allow students who do not know letter names to be tested on their reading skills, simply stopping the assessment and giving the student a correspondingly low score. Thus, while the fall assessment, taken before instruction began, can be seen as a potentially valid measure of equality of beginning skills, scores at later points in the year should be seen as conservative estimates of DI students' achievement. In studying a different curriculum, Smolkowski and Cummings (2016) made a similar point, concluding, "Research [on measures] should also account for the scope and sequence of curricula. . . . The validity of academic screeners depends in part on their alignment with instruction" (p. 115).

References

- Barbash, S. (2012). *Clear teaching: With Direct Instruction, Siegfried Engelmann discovered a better way of teaching*. Arlington, VA: Education Consumers Foundation.
- 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, 32*, 79–88.
- Burke, M. D., Hagan-Burke, S., Kwok, O., & Parker, R. (2008). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *The Journal of Special Education, 42*, 209–226.
- 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*, 141–166.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

- Coughlin, C. (2014). Outcomes of Engelmann's Direct Instruction: Research syntheses. In J. Stockard (Ed.), *The science and success of Engelmann's Direct Instruction* (pp. 25–54). Eugene, OR: NIFDI Press.
- DIBELS. (2008). *DIBELS data system: Using data to improve achievement for each and all*. Retrieved from <https://dibels.uoregon.edu/>
- Engelmann, K. (2014). Creating effective schools with Direct Instruction. In J. Stockard (Ed.), *The science and success of Engelmann's Direct Instruction* (pp. 99–122). Eugene, OR: NIFDI Press.
- Engelmann, S. (1999). Theory of mastery and acceleration. In J. W. Lloyd, E. J. Kame'enui, & D. Chard (Eds.), *Issues in educating students and disabilities* (pp. 177–195). Mahwah, NJ: Lawrence Erlbaum.
- Engelmann, S. (2004). The Dalmatian and its spots: Why research-based recommendations fail logic 101. *Education Week*, 23, 34–35, 48.
- Engelmann, S. (2014). *Successful and confident students with Direct Instruction*. Eugene, OR: NIFDI Press.
- Engelmann, S., & Carnine, D. (1991). *Theory of instruction: Principles and applications* (Rev. ed.). Eugene, OR: ADI Press. (Original work published 1982)
- Engelmann, S., & Carnine, D. (2011). *Could John Stuart Mill have saved our schools?* Verona, WI: Full Court Press.
- Engelmann, S., & Colvin, G. (2006). *Rubric for identifying authentic Direct Instruction programs*. Eugene, OR: Engelmann Foundation.
- Engelmann, S., & Steely, D. (2004). *Inferred functions of performance and learning*. Mahwah, NJ: Lawrence Erlbaum.
- 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.
- Fien, H., Baker, S. K., Smolkowski, K., Smith, J. L. M., Kame'enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, 37, 391–408.
- 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, 67–79.
- Gersten, R. M., Carnine, D. W., Zoref, L., & Cronin, D. (1986). A multifaceted study of change in seven inner-city schools. *The Elementary School Journal*, 86, 257–276.
- i-Ready K-12. (2018). *i-Ready K-12 Adaptive Diagnostic/K-8 Instruction*. Retrieved from <https://www.curriculumassessates.com/products/iready/diagnostic-instruction.aspx>
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.
- National Institute for Direct Instruction. (n.d.). *Educational impact calculator: A tool for education consumers*. Retrieved from <https://www.nifdi.org/research/educational-impact-calculator>
- National Institute for Direct Instruction. (2016). *Writings on Direct Instruction: A bibliography*. Eugene, OR: NIFDI. Retrieved from <https://www.nifdi.org/docman/research/bibliography/205-di-bibliography-reference-list/file>
- National Reading Panel. (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*. Washington, DC: U.S. Department of Education.
- Northwest Evaluation Association. (2011). RIT scale norms: For use with Measures of Academic Progress (MAP®) and MAP® for primary grades. Portland, OR: Author.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Simon, H. A. (1947). *Administrative behavior: A study of decision-making processes in administration organization*. New York, NY: Macmillan.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129–138.
- Smolkowski, K., & Cummings, K. (2016). Evaluation of the DIBELS (sixth edition) diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. *Journal of Psychoeducational Assessment*, 34, 103–118.
- Stockard, J. (2010). An analysis of the fidelity implementation policies of the what works clearinghouse. *Current Issues in Education*, 13(4). Retrieved from <https://cie.asu.edu/ojs/index.php/cieatasu/article/view/398>
- Stockard, J. (2011). Direct Instruction and first grade reading achievement: The role of technical support and time of implementation. *Journal of Direct Instruction*, 11(1), 31–50.
- Stockard, J. (2013). Merging the accountability and scientific research requirements of the No Child Left Behind Act: Using cohort control groups. *Quality & Quantity: International Journal of Methodology*, 47, 2225–2257.
- Stockard, J. (2014). *The relationship between lesson progress in Direct Instruction programs and student test performance* (NIFDI Technical Report 2014-1). Eugene, OR: National Institute for Direct Instruction.
- Stockard, J., Wood, T. W., Coughlin, C., & Khoury, C. R. (2018). The effectiveness of Direct Instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, 88, 479–507. doi:10.3102/0034654317751919
- Tallmadge, G. K. (1977). *The joint dissemination review panel idea book*. Washington, DC: National Institute of Education.
- Vanderwood, M. L., Linklater, D., & Healy, K. (2008). Predictive accuracy of nonsense word fluency for English language learners. *School Psychology Review*, 37, 5–17.
- Wood, T. W., & Stockard, J. (2012). *Reading mastery and learning disabled students: A bibliography*. Eugene, OR: National Institute for Direct Instruction.