EDUCATION AND TREATMENT OF CHILDREN Vol. 11, No. 4, November, 1988

A META-ANALYSIS OF THE EFFECTS OF DIRECT INSTRUCTION IN SPECIAL EDUCATION

W.A.T. White

University of Oregon

ABSTRACT

Studies of the effectiveness of Direct Instruction programs with special education students were examined in a meta-analysis comparison. To be included, the outcomes had to be compared with outcomes for some other treatment to which students were assigned prior to any interventions. Not one of 25 studies showed results favoring the comparison groups. Fifty-three percent of the outcomes significantly favored DI with an average magnitude of effect of .84 standard deviation units. The effects were not restricted to a particular handicapping condition, age group, or skill area.

★ ★ ★

This analysis is based on studies that compared the effectiveness of Direct Instruction (DI) intervention with that of one or more comparison interventions. Only studies with students experiencing some form of learning handicap (e.g., learning disability, trainable mental retardation, reading disability) were included. Studies with students considered "at-risk" for learning problems did *not* qualify.

For a study to be included in the meta-analysis, the assignment of participants to experimental and comparison groups must have taken place prior to intervention. Studies with noncomparable experimental and comparison groups, established by statistically significant differences on pretest scores or by acknowledgment of an author in a report, were excluded.

A study was considered to contain a Direct Instruction treatment group if the author of the report considered one of the groups to be such. Studies were included if a treatment group was based on the Engelmann and Carnine (1982) model of Direct Instruction, or if a group utilized instructional materials developed by Engelmann and associates.

Literature Search

Studies were gathered from research previously known to the reviewer or to the reviewer's colleagues, from reports referenced in such research, and from research listed in a computer literature search conducted on April 30, 1986

For a more thorough examination of this research, refer to the author's December 1986 dissertation (The Effects of Direct Instruction in Special Education: A Meta-Analysis) at the University of Oregon.

Pages 364–374

using data compiled by the Educational Resources Information Center (ERIC). Descriptors used in the search were: direct instruction; direct teaching; directed instruction; directed teaching; DISTAR; direct verbal instruction; active teaching; and active-teaching.

The 25 studies in the meta-analysis for which treatment lasted for over a week are listed in Table 1. Twenty-one of the studies involved mildly handicapped students, one involved moderately handicapped students, and three a mixture of mildly and severely handicapped students.

Some of the studies require explanation about the manner in which they were analyzed for the meta-analysis. The intervention in the Branwhite (1983) study consisted of two phases, only the first of which was included in the meta-analysis. During the second phase, both the experimental and comparison groups received the same Direct Instruction treatment. Thus, data from the second phase clearly are of no relevance. In the Hursh (1979) report, comparisons involved both mildly handicapped students and nonhandicapped students. Only effect sizes based on the comparisons involving the mildly handicapped students were included in the meta-analysis. The Lloyd, Epstein, and Cullinan (1981) and Lloyd, Cullinan, Heins, and Epstein (1980) reports described the same study, but included different dependent measures. In the meta-analysis, the two reports were considered as one study. The Walker, McConnell, and Clarke (1983) report described two studies, but the first study had already been included in the meta-analysis from a separate report (Walker, McConnell, Walker, Clarke, Todis, Cohen, & Rankin, 1983). Thus, for the purposes of this meta-analysis, citations of Walker, McConnell, and Clarke (1983) refer only to the second study in that report. Finally, C. Walker's (1980) master's thesis was coded not from the complete original report, but rather from Lewis' (1982) description of the study, and from photocopies of tables from C. Walker's (1980) results section sent to the reviewer from England.

Coding Study Characteristics

Studies in the meta-analysis were coded on a number of study characteristics or potential moderator variables. *Treatment length* was coded as persisting either for one day, for two to five days, for six days to a month, for over a month to a year, or for over a year. *Fidelity of treatment* was coded as "high" if the research report mentioned that: (a) teachers using Direct Instruction methods were periodically observed for the quality with which they were implementing the treatment; and (b) observation indicated that the level of implementation was adequate. If either of these criteria was lacking, fidelity was coded as "low." The *teacher category* was coded according to whether students were taught during treatment by their regular teacher (i.e., an interventionist from the school district who worked with the students even after the study), or an experimental teacher (i.e., a professional brought in from outside the district for the duration of the study only).

Degree of handicap was coded mild for students with labels of reading dis-

abled, learning disabled, emotionally disabled, or educably mentally retarded. Individuals with greater learning handicaps were considered moderately to severely handicapped. *Age range* was divided into categories of students' school grades determined as if they had progressed academically at the same rate as their age-group peers. *Grades* were coded as prekindergarten, kindergarten through third grade, fourth through sixth grade, junior to senior high school, and after high school. Whether a study was coded as experimental or quasi-experimental in design was determined by whether assignment of individual participants to groups was done randomly. Random assignment of entire classes to treatment groups was considered quasi-experimental in *design*, since all research reports used in the meta-analysis provided outcome means for individuals rather than for classes.

Outcome measures used in the studies fell into three categories: norm-referenced tests, criterion-referenced tests, and "other" tests (tests designed for use with a particular curriculum, observational measures, student data from school files, etc.). Tests that could not be positively determined to be criterion-referenced from a reading of the research report were placed in the category of norm-referenced measures.

Effect Sizes

An effect size was calculated for each dependent measure on which the experimental and treatment groups were compared. The effect size was computed by dividing the difference between the means of the experimental and comparison groups by the pooled standard deviation, as advocated by Wolf (1986). Effect sizes favoring Direct Instruction groups were assigned positive values; those favoring comparison groups were assigned negative values.

When the necessary figures for effect size computation (i.e.. means, standard deviations, and sample sizes) were not available in a report, estimates of the effect size were calculated from proportions, t values, or F values where possible. However, effect sizes were not estimated from p values for a number of reasons.

In synthesizing the effect sizes across the studies in the meta-analysis, the individual study rather than the individual outcome measure was used as the unit of analysis. Thus, for synthesis of overall effect size of Direct Instruction, the Maggs and Morath (1976) study contributed an effect size (ES) of 1.93, which was the mean of the six ESs from its individual measures. For synthesis of effect size of Direct Instruction on measures of *intellectual ability* (only), the Maggs and Morath (1976) study contributed an ES of 2.57, which was the ES of its only measure of intellectual ability.

A study-weighted "vote count" was also conducted for synthesizing the research results. For each study the proportion of measures for which there was a statistically significant difference favoring the experimental group, and the proportion significantly favoring the comparison group was computed. The

Table 1
Individual Study Effect Sizes for DI

Study	Target Skill	Research Design	Effect Sizes: Overall ES	Effect Sizes: Academic ES	Proportion of Significant Outcomes
Branwhite (1983)	reading	Q	+1.61	+1.61	1.00
Campbell (1983)	reading	Q	+1.08	+1.12	.83
Darch & Kameenui (1986)	reading	E	+1.59	+1.59	1.00
Gleason (1985)	math	E	+0.57	+0.76	.33
Gregory et al. (1982)	reading	E	+1.66	+1.71	1.00
Haring & Krug (1975)	reading	Q	+1.05	+1.05	1.00
Hursh (1979)	academics	Q	+0.71	+0.77	.25
Kelly et al. (in press)	math	E	+1.39	+1.39	.60
Leiss & Proger (1974)	language	Q	+0.40	–	.00
Lewis Study 1 (1982)	reading	E	+0.16	+0.16	.00
Lewis Study 2 (1982)	reading	E	−0.40	−0.40	.00
Lloyd et al. (1981)	reading	E	+0.84	+0.85	.17
Maggs (no date)	language	E	–	–	1.00
Maggs & Morath (1976)	language	E	+1.93	–	1.00
Moodie & Hoen (1972)	math	Q	−0.14	−0.14	.00
Proger & Leiss (1976)	language	Q	+1.30	–	.71
Richardson et al. (1978)	reading	E	+0.10	+0.10	.25
Stein & Goldman (1980)	reading	Q	+0.75	+0.75	1.00
Stephens & Hudson (1985)	spelling	E	+1.94	+1.94	1.00
Summerell & Brannigan (1977)	reading	Q	+0.54	+ .54	.50
C. Walker (1980)	reading	Q	+0.39	+0.04	.00
H. Walker et al. (1983)*	social	E	+0.29	–	.10
H. Walker et al. (1983)**	social	E	+0.99	–	.44
Weiherman (1984)	writing	E	+0.33	+0.33	.50
Woodward (1985)	health	E	+1.02	+1.22	.56
Mean ES			+0.84	+0.82	.53
SD			.64	.67	.40

* refers to Walker, McConnell, & Clark
** refers to Walker, McConnell, Walker, Clarke, Todis, Cohen, & Rankin
E = Experimental Design (random assignment to groups).
Q = Quasi-experimental design.

mean of the individual studies' proportions represents a study-weighted proportion of significant outcomes.

Results

The effect sizes for 25 studies that compare the outcomes for DI groups of handicapped students with the outcomes for comparison groups are listed in Table 1. Not a single outcome measure in any of the 25 studies significantly favored the comparison treatment. The means show that on the average, 53 percent of outcome measures significantly favor DI. This value far exceeds the 5 percent that would be expected by chance if there were actually no differential effects between the DI and the comparison treatments. The average ad-

vantage of .84 standard deviation units that DI treatment maintains over comparison treatments is well above the standard of .25 to .33 that has been typically used to determine educational significance of an educational treatment effect (Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977).

A quasi-experiment that produced a non-significant negative effect for DI (Moodie & Hoen, 1972) compared DISTAR Arithmetic with traditional math instruction in Canadian learning assistance classes for one school year. Post-intervention interviews and questionnaires indicated that teachers liked DISTAR and were pleased with the apparent progress of their students. The questionnaires also indicated, however, a low level of implementation of the DISTAR system. No information was provided on the comparability of experimental (N = 14) and comparison (N = 24) students at the start of the study. The authors (Moodie & Hoen, 1972) emphasized that their study had severe limitations and offered strong subjective support for DISTAR.

One of Lewis' (1982) experiments investigated the effects of DI for 11- and 12-year-olds with reading disorders. Students who were taught with traditional remedial programs and other model programs *scored higher* than DI students on *posttests* of word attack skills and reading comprehension. The respective ESs were −.47 and −.32. However, DI students averaged *gains* of 8.4 months in spelling, compared to 5.4 months and 3.0 months for the two comparison groups. Adequate information was not available for the calculation of the spelling effect size. None of the measures in Lewis' study produced statistically significant differences.

Reading and Mathematics

The DI studies that investigated academic outcomes have been divided according to specific skill areas. The study-weighted mean ESs for measures of reading and mathematics achievement are listed in Table 2. The mean ES in reading of .85 is consistent with the mean ESs for DI, both overall and in achievement measures. A further subdivision of reading measures (into comprehension, word-attack, and total reading measures) does not support the arguments of those educators who contend that DI teaches basic academic skills of a lower-order cognitive level (such as word-attack skills) at the ex-

Table 2
Study-Weighted DI Effect Sizes in Reading and Mathematics

	Measures	
	Reading	Math
Mean	+0.85	+0.50
Median	+0.80	+0.38
SD	.78	.71
(N)	(13)	(4)

pense of higher-level skills (such as comprehension). The study-weighted mean for DI on word-attack measures across 10 studies was .64. The corresponding mean for measures of reading comprehension across eight studies was .54. Using a difference of .33 standard deviation units as the criterion for an educationally significant difference, there is no important difference between ESs for DI in the "low level" word-attack skills and the "high level" reading comprehension skills.

The study-weighted mean ES of .50 for math was lower than the corresponding mean for reading. However, less confidence can be placed in a mean ES resulting from only four studies.

Intelligence and Readiness

Effects of Direct Instruction on measures of intellectual ability and readiness skills are indicated in Table 3. Because of the low number of studies (4, 5, and 2) in these effect sizes, they should be treated with caution. Typically, standardized measures of intelligence are not the most responsive measures of educational intervention. However, since the earliest research on the DI model in the mid-sixties with "at risk" nonhandicapped preschoolers (Engelmann, 1971), DI has produced appreciable gains in IQ. All studies that measured IQ in this meta-analysis (Leiss & Proger, 1974; Lloyd, Cullinan, Heins, & Epstein, 1980; Maggs, n.d.; Maggs & Morath, 1976; Proger & Leiss, 1976) made use of the DISTAR Language curriculum, which was quite similar to that used in the preschool studies. Apparently, the same curriculum and approach that were beneficial for young students who are at risk for developing learning handicaps are also effective with older students with demonstrable learning handicaps. DISTAR Language presents basic language concepts in a controlled, systematic manner, and teaches some of the language abilities (e.g., analogy, deduction) measured by most intelligence tests.

Measures of academic preskills, basic concept learning, language development, psycholinguistic abilities, and Piagetian cognitive development were pooled together and called "readiness" measures. Except for learning basic language concepts (e.g., under/over, singular/plural, past tense/present tense),

Table 3
Study-Weighted DI Effect Sizes in Intelligence and Readiness

	Measures	
	Intellectual Ability	Overall Readiness
Mean	+1.32	+1.13
Median	+1.13	+0.89
SD	.93	.88
(N)	(4)	(5)

Direct Instruction programs usually skip over so-called readiness activities in favor of academic skills. However, six studies (Campbell, 1983; Hursh, 1979; Leiss & Proger, 1974; Maggs, n.d.; Maggs & Morath, 1976; Proger & Leiss, 1976) suggest that Direct Instruction students more than hold their own in readiness skills. Table 3 shows a study-weighted mean ES for the pooled readiness measures of 1.13 for 5 of the 6 studies that could be included in the effect-size analyses. Positive effects were found in all five subcategories of readiness measures.

Degrees of Handicap and Type of Comparison Group

The research results show that Direct Instruction can be equally effective for mildly and moderately/severely handicapped students. The mean ES for mildly handicapped students was .80, and the mean for the more severely handicapped students was 1.01. This difference of .21 standard deviation units between the two figures does not meet the standard of .33 for educationally meaningful differences. It is difficult to compare the two groups of studies, because 18 of the 20 studies involving mildly handicapped students measured academic achievement (mean effect size of .85), whereas only 1 study of 4 involving moderately handicapped students did so.

One variable that did have a significant effect on effect size was the type of comparison group(s) used in a study. None of the studies utilized a pure control (i.e., no treatment) group, but three of them (Campbell, 1983; Walker, McConnell, & Clarke, 1983; Walker et al., 1983) utilized a comparison group that was involved in activities unrelated to the final outcome measures. These studies produced an average effect size of .79. The mean ES for these studies was probably held down to some extent by the rigorous tests of generalization in social skills used in the Walker, McConnell, and Clarke (1983) and the Walker et al. (1983) studies.

Grade Level and Other Study Characteristics

Most of the studies in the meta-analysis were conducted with students in the age ranges of the intermediate grades (4th to 6th) and secondary level (7th to 12th). The 13 studies in the intermediate grades were actually composed of 6 studies that fit neatly into the category, and 7 studies that included students in a wide range of grades (e.g., grades 1–6, grades 1–9), which were judged closer to the intermediate category than to any of the other categories. The mean effect size for the 6 studies that fit the category neatly was .65, which is similar to the mean of .69 for all 13 studies categorized as intermediate. The mean effect size for 7 secondary studies was 1.15. The data show that DI is quite effective for both age groups.

Type of Posttest

Another variable that had a significant impact on the magnitude of effect sizes was the type of the posttest. Table 4 shows that criterion-referenced measures generated significantly greater effect sizes than did norm-referenced

Table 4
Study-Weighted DI Effect Sizes for Different Forms of Measures

	Form of Measure		
	Criterion-Referenced	Norm Referenced	Other
Mean	+1.67	+0.77	+0.70
Median	+1.13	+0.71	+0.71
SD	.94	.72	.50
(N)	(8)	(17)	(8)

measures or "other" measures (i.e., observation, official records, and self-report). This result is consistent with the premise that criterion-referenced tests can be more sensitive to the effects of instruction than are standardized tests or non-academic measures. Also, criterion-referenced tests that were *closely aligned* with the tasks assigned to students during intervention yielded higher effect sizes (mean of 1.76 across eight studies) than did criterion-referenced tests of *low alignment* (mean of 1.06 across six studies). These figures support Cohen and Hyman's (1982) contention that congruence between intervention task and outcome measure is a major determinant of effect size. The effect of 1.06 for measures of low alignment also suggests that Direct Instruction students transfer what they have learned to somewhat different kinds of tasks.

Teachers and Treatment Length

The teacher variable played a significant part in *academic* effect sizes, but not in *overall* effect sizes. Study-weighted mean ESs, across all outcome measures, were .81 for 18 studies that used regular (usually classroom) teachers, and a slightly higher .94 for six studies that used experimental teachers. Across measures of academic achievement, however, the difference between effect sizes exceeded the .33 standard for educational significance. For 15 studies with regular teachers, the mean was .79; for four studies with experimental teachers, the mean was 1.13. It makes sense that experimental teachers, having been specially trained in the experimental curricula, might implement the Direct Instruction intervention more faithfully, which in turn might bring about a greater effect.

Another variable, length of treatment, also had a greater differential impact on academic than on other measures, but the effect is confounded. When all measures within a study were averaged, the study-weighted mean effect for Direct Instruction in interventions lasting from six school days to one month ($N = 5$) was .98 standard deviation units. The corresponding overall mean ES for interventions ranging from over a month to a year ($N = 17$) was .77. When only academic measures within a study were averaged together, the study-weighted mean for the 5 comparatively shorter studies was 1.06, which is greater than the corresponding mean effect size of .77 for the 13 comparatively longer studies. The confound comes from the fact that the 5 shorter studies (Darch

& Kameenui, 1986; Gleason, 1985; Kelly, Carnine, Gersten, & Grossen, in press; Weiherman, 1984: Woodward, 1985) were all university-generated projects using experimental teachers. The implementation levels of these short university projects very likely exceed those managed by school district personnel.

Level of Implementation and Research Design

Whether or not fidelity of treatment was assessed had no systematic impact on effect sizes. Studies (N = 11) that referred to high levels of Direct Instruction implementation produced a mean ES of .86; those (N = 13) that *failed to mention* level of implementation or that certified that implementation was poor produced a mean ES of .82. These numbers seem to conflict with other research (Gersten, Carnine, & Williams, 1982: Siegal & Rosenshine, 1973) that indicate that fidelity of treatment plays a significant role in the impact of Direct Instruction. However, some authors and observers may have judged level of implementation to be adequate, when in fact it was not. Also, in some other studies, the Direct Instruction teachers probably followed the experimental programs carefully, yet the reports make no mention of fidelity of treatment. Type of research design (experimental versus quasi-experimental) also had no systematic effect.

Summary

It seems that there are occasional circumstances under which Direct Instruction can produce a negative effect. However, only 14 percent of the comparisons showed a negative effect for Direct Instruction. (One percent of the comparisons produced a neutral, or .00, effect. *None of the negative effects were statistically significant.*)

The 25 studies on Direct Instruction treatments of over a week in length found a strong, consistent effect for the treatment. The strength is not limited to a particular age range, or handicapping condition, or skill area. The meta-analysis indicates that, based on 25 studies, instruction grounded in Direct Instruction theory (Engelmann & Carnine, 1982) is efficacious for both mildly and moderately/severely handicapped learners, and in all skill areas on which research has been conducted.

References

Branwhite, A. B. (1983). Boosting reading skills by direct instruction. *British Journal of Educational Psychology, 53*, 291–298.

Campbell, M. L. (1983, May). *A study of Corrective Reading as an effective and appropriate program for reading disabled, learning handicapped secondary students.* Report presented to the Faculty of the School of Education, San Diego State University, San Diego.

Cohen, S. A., & Hyman, J. S. (1982, March). *Two components of quality instruction.* Paper presented at the meeting of the American Educational Research Association, New York.

Darch, C., & Kameenui, E. J. (1986). *Teaching learning disabled students critical reading skills: A systematic replication.* Unpublished manuscript, Auburn University.