ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/322332768

The Effectiveness of Direct Instruction Curricula: A Meta-Analysis of a Half Century of Research

Article *in* Review of Educational Research · January 2018

DOI: 10.3102/0034654317751919

CITATIONS
256

READS
5,653

All content following this page was uploaded by Cristy Coughlin on 01 August 2022.

The user has requested enhancement of the downloaded file.

Check for updates

Review of Educational Research
August 2018, Vol. 88, No. 4, pp. 479–507
DOI: 10.3102/0034654317751919
© 2018 AERA. http://rer.aera.net

The Effectiveness of Direct Instruction Curricula: A Meta-Analysis of a Half Century of Research

Jean Stockard and Timothy W. Wood
University of Oregon

Cristy Coughlin
Safe and Civil Schools

Caitlin Rasplica Khoury
The Children's Clinic, P.C.

Quantitative mixed models were used to examine literature published from 1966 through 2016 on the effectiveness of Direct Instruction. Analyses were based on 328 studies involving 413 study designs and almost 4,000 effects. Results are reported for the total set and subareas regarding reading, math, language, spelling, and multiple or other academic subjects; ability measures; affective outcomes; teacher and parent views; and single-subject designs. All of the estimated effects were positive and all were statistically significant except results from metaregressions involving affective outcomes. Characteristics of the publications, methodology, and sample were not systematically related to effect estimates. Effects showed little decline during maintenance, and effects for academic subjects were greater when students had more exposure to the programs. Estimated effects were educationally significant, moderate to large when using the traditional psychological benchmarks, and similar in magnitude to effect sizes that reflect performance gaps between more and less advantaged students.

Keywords: Direct Instruction, meta-analysis, academic achievement

The importance of explicit and systematic instruction has become a central element of discussions of effective instruction (e.g., National Reading Panel, 2000). Direct Instruction (DI), developed by Siegfried Engelmann and his collaborators beginning in the 1960s, is often cited as an example. Over the past half century the corpus of DI curricular materials has grown as has the literature evaluating its effectiveness. This article presents a quantitative analysis of this effectiveness literature. 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, our focus is only on Direct Instruction (capitalized) in the Engelmann–Becker tradition (Engelmann & Colvin, 2006).

Theoretical Base

Theoretical writings related to DI are numerous and complex, describing the logical basis of the approach and including empirical tests of subelements of each part of the theoretical development. (See Engelmann, 1999; Engelmann & Carnine, 1991, 2011; and Engelmann & Steely, 2004, for detailed theoretical discussions; Barbash, 2012 for an accessible summary; and National Institute for Direct Instruction, 2016, pp. 169–173, for citations to 45 experimental examinations of the tenets.) The discussion in this sub-section, and those that follow, provides only a brief overview.

Direct Instruction builds on the assumption that all students can learn with well-designed instruction. When a student does not learn, it does not mean that something is wrong with the student but, instead, that something is wrong with the instruction. Thus, the theory underlying DI lies in opposition to developmental approaches, constructivism, and theories of learning styles, which assume that students’ ability to learn depends on their developmental stage, their ability to construct or derive understandings, or their own unique approach to learning. Instead, DI assumes all students can learn new material when (a) they have mastered prerequisite knowledge and skills and (b) the instruction is unambiguous. In other words, DI assumes that students are inherently logical beings. Like the constructivist approach, DI assumes that students make inferences from examples that are presented to them. But, unlike constructivism, the theory underlying DI states that learning is most efficient when the examples are carefully chosen and designed. They must be as unambiguous as possible, sequenced to promote the correct inference for learning a new concept, and involve the fewest possible steps to induce learning.

Mastery learning is a key element of DI. DI theory posits that when students become fluent in a new task, fully grasping a new concept or skill, it becomes part of an existing repertoire. It is then easier to learn new things that build on that foundation. In addition, it is far easier to learn a new concept than to unlearn a faulty conceptualization. Two key elements of DI curricular programs derive from this theoretical point. First, it is important to ensure that students have mastered key concepts before moving forward. Second, proper placement in a curricular program is essential to make sure students have the prior knowledge needed to learn new concepts or skills and that they will not be wasting time on material that was already mastered.

The combination of these elements is seen as resulting in both effective and efficient instruction. Students should learn more in less time. The theoretical writings stress the importance of providing continuous positive reinforcement throughout the instructional process and celebrating students’ success at regular intervals. Thus, the learning process should be rewarding to students. In addition, the process should, theoretically, be rewarding to teachers as they see their students’ progress (Engelmann, 2014c).

The Direct Instruction Approach

As implied in the previous section, developing unambiguous instruction for even very simple concepts is difficult. Research related to the theoretical discussions has shown how very small variations in the types of examples given to students can result in erroneous conclusions. DI curricular materials are designed to guard against this possibility by providing highly structured guidance to teachers in the wording, sequencing, and review of material presented to students. They incorporate a "tracked design," in which discrete skills and concepts are taught in isolation but are then brought together in increasingly more sophisticated and complex applications. Placement tests are included to ensure that students are taught material that is neither too challenging nor repetitive of material already mastered. Teachers and administrators are encouraged to regroup students at regular intervals when needed to promote the greatest learning.

Although the importance of following the scripted teaching material is stressed, the programs and associated writings also emphasize the importance of teachers using their own style and personality to animate the presentation, much as actors bring their own approach to a role. The instructional materials are designed to be fast-paced and to include consistent reinforcement for students, daily checking for learning, and regular testing of mastery. The aim is an instructional situation in which students are continuously learning and progressing through material. The careful design and fast pace of the material are thought to result in higher achievement, and the students' positive experience is expected to enhance their self-concepts and self-esteem (Barbash, 2012; Engelmann, 2014c).

Method of Curricular Development

DI programs are developed in a multistage, multiyear process. The development begins with a detailed logical analysis of the concept to be taught. Carefully worded examples and teaching scripts are developed and tested in-house and with small groups to help ensure that they are unambiguous. Materials are logically sequenced, with placement tests, systematic review of previously taught material, and regular testing of mastery. The programs are then field tested in schools, with teachers providing detailed feedback on any problems that students have with the programs. In response to this feedback the programs are revised and sent for a second round of field testing and potential revision. Only after this repeated process of field testing and revisions are the programs sent for publication (Collins & Carnine, 1988; Engelmann, 2014b; Huitt, Monetti, & Hummel, 2009).

History of Direct Instruction

The formal beginning of DI was a preschool program for children from very impoverished backgrounds at the University of Illinois in the mid-1960s. Siegfried Engelmann and colleagues, Carl Bereiter and Jean Osborn, incorporated short instructional periods into the program, focused on language as well as reading and math skills and using the careful instructional sequencing described above. Even though the sessions were quite short, from 20 to 30 minutes a day, the children showed marked improvements in their skills, encouraging further development of the approach (Bereiter & Engelmann, 1966). Bereiter left Illinois for another university and was replaced by the psychologist Wesley Becker. The team also expanded to include several undergraduate and graduate students, many of whom continued to work with DI through later decades. They applied the principles developed in the preschool to other groups and began developing formal instructional programs in language, reading, and math. The formal programs were termed *DISTAR*, for Direct Instruction System for Teaching Arithmetic and Reading (Engelmann, 2007; Wood, 2014).

In the late 1960s, DI was accepted as one of the programs to be part of Project Follow Through, a very large government-funded study that compared the outcomes of over 20 different educational interventions in high-poverty communities over a multiyear period. Communities throughout the nation selected programs to be implemented in their schools, and DI was chosen by 19 different sites, encompassing a broad range of demographic and geographic characteristics. External evaluators gathered and analyzed outcome data using a variety of comparison groups and analysis techniques. The final results indicated that DI was the only intervention that had significantly positive impacts on all of the outcome measures (Adams, 1996; Barbash, 2012; Bereiter & Kurland, 1996; Engelmann 2007; Engelmann, Becker, Carnine, & Gersten, 1988; Kennedy, 1978).

The developers of DI had hoped that the conclusions of the Project Follow Through evaluators would lead to widespread adoption of the programs, but a variety of political machinations seem to have resulted in the findings being known to only a few scholars and policy makers (Grossen, 1996; Watkins, 1996). However, Engelmann, Becker, and their colleagues, then based at the University of Oregon, continued with their work. The programs were expanded to the upper elementary grades, previously written programs were revised, new programs were developed, and attention was given to work with special groups of students such as those with severe disabilities and English language learners (ELLs). The new and revised programs were given different names, more specific to the subject matter, although much of the original content remained and the underlying structure was consistent with earlier versions. Throughout the past five decades, studies of the effectiveness of the DI programs have continued to appear, some as part of the work of the developers and their students but many more from scholars and students throughout the country and in other nations. This body of work, produced from the mid-1960s to the mid-2010s, is the focus of the present article.

Previous Reviews of the DI Effectiveness Literature

Several systematic reviews and meta-analyses of the DI effectiveness literature have appeared over the past quarter century (Coughlin, 2014). Systematic reviews have focused on different populations, including general education (American Institutes for Research, 1999) and special education students (Kinder, Kubina, & Marchand-Martella, 2005). Other reviews have examined literature related to specific programs including mathematics (Przychodzin, Marchand-Martella, Martella, & Azim, 2004), reading (Przychodzin-Havis et al., 2005; Schieffer, Marchand-Martella, Martella, Simonsen, & Waldron-Soler, 2002), and spelling (Simonsen & Gunter, 2001). All of these reviews reported strong, positive results for DI.

Quantitative meta-analyses have also focused on a range of populations and subjects, usually limiting the studies included with methodological criteria or sample characteristics. They have examined the use of DI in whole school reform (Borman, Hewes, Overman, & Brown, 2003), special education populations (White, 1988), and results with the *Reading Mastery* program (Stockard, 2013; Stockard & Wood, 2017; Vitale & Kaniuka, 2012). Two meta-analyses examined results with both general education and special education students and a variety of subjects, restricting the analysis to experimental and quasi-experimental designs (Adams & Engelmann, 1996) or randomized control trials (Coughlin, 2011).

Like the systematic reviews, the results of the meta-analyses were consistently positive. The magnitude of the estimated overall effects ranged from a low of 0.15 in the study of whole school reform models to values of 0.87 and higher in the broader samples (e.g., Adams & Engelmann, 1996; Coughlin, 2011; Hattie, 2009). Hattie (2009) stressed the similarity of results across subjects, but others have reported a range of effects. Adams and Engelmann (1996) found effects that ranged from a low of 0.49 for language to a high of 1.33 for spelling. In contrast, Coughlin (2011) reported values of 0.53 and 0.54 for reading and miscellaneous subjects and 0.81 for language and 1.03 for math. Few of the meta-analyses used metaregression to examine the possible role of moderating variables. None of the analyses included measures other than achievement in academic subjects, and none included results from all available years, subjects, and designs.

The Current Review

The current review was designed to, as much as possible, counter these limitations by applying quantitative meta-analysis techniques to the full range of literature on the effectiveness of DI that appeared over the past half century. Like Adams and Engelmann (1996) and Coughlin (2011), we wanted to examine the effectiveness of the programs in multiple subjects and with different student populations. At the same time, we expanded our focus, including studies with a broader range of research designs and dependent variables than in previous works. Thus, we wanted, as much as possible, to develop a comprehensive view of estimates of the effectiveness of DI. Using multivariate analyses, we also wanted to examine the extent to which these estimates varied across subjects, different types of publications, methodological approaches, sample characteristics, and intervention procedures.

The previous literature provided a few expectations to guide our work. Given the consistent findings of previous reviews, we expected that estimates of effects would be positive across all of the academic areas taught in the programs. It was more challenging to predict results in other areas examined, specifically measures of ability (IQ), student behavior and attitudes, and teacher and parent attitudes. However, we had no reason, given isolated findings as well as the results of Project Follow Through, to expect that effects would be negative. Analyses of other areas within education have concluded that reported effects are stronger in published than in unpublished sources (Polanin, Tanner-Smith, & Hennessy, 2016), but we knew of no studies of this phenomenon within the DI literature. Similarly, previous work offered limited guidance regarding variations in effects related to characteristics of methodologies and student samples. In contrast, previous research did prompt expectations regarding the impact of several variables related to the nature of the intervention. For example, stronger effects were expected when students had greater exposure to the material (e.g., starting in kindergarten, longer periods of intervention, and greater daily exposure) and when teachers had been well trained (Carlson & Francis, 2002; MacIver & Kemper, 2002; O'Brien & Ware, 2002; Stockard, 2011; Vitale & Joseph, 2008). Previous research also suggested that positive effects would be maintained after discontinuation of intervention (Becker & Gersten, 1982; Meyer, 1984; Stockard, 2010). However, it could be reasonable to expect that the maintenance effects would be smaller than those from immediate postintervention and that the impact of the programs could decline as the maintenance period lengthened. Finally, previous research provided little guidance regarding the relationship of sample size in a study and reported effects, although it could be logical to expect that effects would be stronger in studies with smaller samples that could have greater control over the fidelity of implementation.

Method

Procedures

Subsections below describe the inclusion and exclusion criteria that were used, the method of searching the literature, and coding procedures.

Inclusion and Exclusion Criteria

Throughout this discussion the term *report* is used to refer to an individual publication, such as an article or dissertation, and the term *study* to refer to the data-gathering effort, such as an experiment or other type of intervention, on which a report was based. (As explained in greater detail in the results section, a report could involve more than one study and a study could result in more than one report.) We limited our analysis to studies that examined DI in the Engelmann–Becker tradition and omitted studies that used only elements of the approach. We omitted studies that combined interventions, presenting data for students exposed to both DI and other programs, and also omitted studies that reported only the impact of more or less exposure to the program (i.e., with no data regarding results with no exposure). Within studies we omitted comparisons that were aggregates of others. The coders noted any issues regarding the quality of the research procedures and reporting. Studies that were regarded as having moderate or serious quality issues were omitted. Finally, we omitted outliers, any effects with an absolute value greater than 3.0. (The impact of the exclusions related to quality and outliers were examined in the sensitivity analysis explained below. Appendix B, in the online version of the journal, includes a list of all reports and studies examined, both included and excluded, as well as their associated effect sizes.)

Beyond these limitations, we purposely took an inclusive approach to developing the sample. Articles published in peer-reviewed journals, dissertations, masters' theses, and technical reports and other nonpublished material (so-called gray literature) were included in our initial review. Although most of the reports examined student academic achievement, we also found and included reports that considered ability (IQ) measures; student affective outcomes, such as attitudes, self-esteem or behavior; teachers' perceptions of effectiveness; and teacher or parent views of the programs. We had no limits on date of publication, beginning our analysis with the first published reports on DI in the mid-1960s. We also had no limits on the location or site of a study and included all research designs for which we could compute a valid effect size. Only reports published in English were included.

Literature Search

To identify studies, we began with an extensive bibliography compiled by the National Institute for Direct Instruction (2016). We then examined bibliographies of the meta-analyses and reviews cited above. We also searched computerized databases, including Google Scholar, Dissertation Abstracts, and ERIC. In these searches we used "Direct Instruction" as a key word as well as the names of each of the programs and authors who were known for writing about them. As we reviewed reports we examined the reference lists to see if there were additional items that should be added. We also examined the vita of authors known to have published extensively on DI, wrote to scholars with wide knowledge of the field, and asked members of an email list of DI researchers to send us any needed additions.[1] Finally, we went through each issue of four publications that specialized in reporting on effectiveness studies of DI: *ADI News*, published from 1981 to 1992; *Effective School Practices*, published from 1993 to 2000; *Direct Instruction News*, published from 2001 to 2013; and the *Journal of Direct Instruction*, published from 2001 to 2012. We ceased our search for materials in January 2017.

Coding Procedures

Some coding was done by a team of advanced doctoral students at the University of Oregon. This team was trained and supervised by one of the authors. The rest of the coding was done by the authors. All of the coding by the graduate student team was checked by at least one of the authors and often by another coding team member. In addition, when coding was done by only one author, that author reviewed and rechecked the coding several months after the original examination to resolve any discrepancies. (Interrater reliability exceeded 90%.) All coding and calculation of effect sizes were completed in Excel and the values were then transferred to STATA (StataCorp, 2011) for statistical analysis. After transferring the data, extensive additional checks on each code, involving both possible range and logical relationships, were conducted to help ensure accuracy.

Measures

The first subsection below describes our measure of effect size. Subsequent subsections describe variables used in metaregressions. For the multivariate analyses categorical variables were converted to dummy (0, 1) codes. To maintain adequate degrees of freedom, when data were not available for a given variable, the missing cases were included in the reference (0) category. As noted below, some indicators used in the multivariate analysis were measured at the comparison, or effect size, level of analysis, and others were measured at the study or design level of analysis. Additional details are in Appendix A (in the online version of the journal), and a full codebook is available on request.

Outcome Measure (Effect Size)

Cohen's *d* was our measure of effect, defined as the difference between the means divided by the common standard deviation:

$$d = \left(M_1 - M_2\right) / SD_c, \tag{1}$$

where M_1 and M_2 are the means (averages) of Groups 1 and 2 and SD_c is the pooled or common standard deviation, calculated as the weighted average of the *SD* of the two groups. In our calculations Group 1 was always the DI group. Thus positive effect sizes indicate an advantage for that group.

The psychological literature has often used a criterion of 0.20 to designate small effects, 0.50 medium effects, and 0.80 and greater as large (Cohen, 1977, 1988). Education researchers have traditionally used the threshold of 0.25 to indicate an educationally significant outcome (Tallmadge, 1977). A more recent extensive discussion of effect sizes in education research suggested another comparison benchmark. Based on a review of a "wide range" of educational interventions, Lipsey et al. (2012) concluded that effect sizes in the field

> are rarely as large as .30. By appropriate norms—that is norms based on empirical distributions of effect sizes from comparable studies—an effect size of .25 on such outcome measures is large and an effect size of .50, which would be only "medium" on Cohen's all encompassing distribution, would be more like "huge." (p. 4)

We reference all three benchmark comparisons—Cohen's (1977, 1988), Tallmadge's (1977), and Lipsey et al.'s (2012)—in our discussion below.

When studies reported percentages rather than means, we calculated the effect size from a standard difference of proportion test. Percentiles were translated to normal curve equivalent scores before calculating the effect size. When possible we also computed Hedges's *g*, which differs from Equation 1 only when samples are quite small. As would be expected, the values of *d* and *g* were virtually identical, differing, on average, by only 0.008. When only inferential statistics or odds ratios were available we converted the values to *d*. Results regarding single-subject designs were aggregated across all subjects within an analysis. Two types of adjustments were used, one that adjusted for extensive differences between intervention and control groups at pretest and another that adjusted for regression to the mean.[2]

Nature of Publication

We included five sets of measures related to the nature of the research reports and studies, all measured at the study unit of analysis: (a) the year of publication, with four dummy variables distinguishing those published in 1977 to 1986 and each subsequent decade to those published from 1966 to 1976; (b) the nature of the publications, with two dummy variables contrasting (i) articles and (ii) dissertations and theses from those that were only disseminated as gray literature, such as technical or newsletter reports; (c) the source of the publication, with two dummy variables distinguishing (i) material disseminated by the publisher of most of the DI programs and (ii) those by the authors of this article from other works; (d) a dummy variable indicating if the study was based on data from Project Follow Through; and (e) a series of five dummy variables that were used to denote studies where the data had been gathered from the same community.[3]

Methodology-Related Variables

We included six sets of dummy variables related to methodology. One set, regarding the design used to gather the data, varied at the study design level of analysis. It included dummy variables distinguishing (a) randomized assignment of students to groups (either pretest–posttest or posttest only), (b) norm or goal comparisons, (c) cohort control groups, (d) statistical controls (either pretest–posttest or posttest only), (e) other pretest-posttest control group designs, and (f) other posttest only control group designs from single subject designs.[4] The other five sets of methodological variables were measured at the comparison or effect size level of analysis: (a) the type of assessment used, distinguishing (i) norm-referenced and other published assessments, (ii) curriculum-based measures, and (iii) state assessments from researcher-designed and other measures; (b) the type of data used to calculate effects, with dummy variables distinguishing effects based on (i) percentages or counts, (ii) percentiles converted to normal curve equivalent scores, (iii) statistical results such as regression coefficients, and (iv) other alternative data types from those based on means and standard deviations; (c) whether the data used to calculate effects had been adjusted by the original authors (as in reports of adjusted means); (d) the method used to calculate the effect, with dummy variables distinguishing calculations involving (i) the difference of the pre- and posttest effect size, (ii) transformations of inferential or other statistics, (iii) transformations from odds ratios, and (iv) effects supplied by the study author, with effects calculated with the formula in Equation 1 as the reference category; and (e) whether any additional calculations (e.g., of means from raw data) or estimations (e.g., of sample size) were used in calculating effects.

Sample-Related Variables

Four sample-related variables varied at the study design unit of analysis: (a) the breadth of the sample, with dummy variables distinguishing the inclusion of (i) more than one classroom, (ii) more than one school, and (iii) more than one district in both the intervention and control group from samples that included only one classroom in the intervention or control group; (b) the location of the study, distinguishing (i) urban U.S. locales, (ii) suburban U.S. locales, and (iii) rural U.S. locales from those with multiple locations or in other countries; (c) the region in which the study occurred, distinguishing four broad areas of the United States from international and multiple locations; and (d) a dummy variable indicating high rates of student poverty (either greater than 75% receiving free or reduced-price lunch or a statement of high poverty given by the researcher). Four sets of sample-related measures varied at the effect size or comparison level: (a) students' at-risk status, with a single dummy variable distinguishing results involving students with any type of at-risk status (e.g., receiving special education services, “remedial,” “low-achieving”) from other students; (b) ELL status, again measured with a single dummy variable; (c) race-ethnicity, with dummy variables indicating that the authors had indicated results involved large proportions of (i) African American, (ii) Latino/a, (iii) American Indian, or (iv) Caucasian students; and (d) grade level, which was reduced to a set of dummy variables distinguishing those in (i) preschool and kindergarten, (ii) Grade 1, (iii) Grade 2, (iv) Grade 3, and (v) Grade 4 from effects involving students in higher grades.

Intervention-Related Variables

Four intervention-related variables were measured at the study design level of analysis: (a) how the program was delivered, distinguishing interventions delivered by the classroom teacher from other approaches; (b) the nature of the comparison program, with a dummy variable indicating if the comparison was the curriculum usually or already used at a school or some type of experimental curriculum; (c) whether teachers had been trained or coached in proper use of the programs; and (d) the specific DI program that was used, with dummy variables distinguishing (i) *DISTAR*, (ii) *Reading Mastery* or *Horizons*, (iii) *Corrective Reading*, (iv) *Connecting Math Concepts* and other math programs, (v) *Language for Learning*, (vi) spelling programs, and (vii) multiple programs with other DI curricula as the reference category.

There were five intervention-related sets of variables measured at the effect size or comparison level: (a) the length of the intervention, measured continuously in years; (b) the amount of exposure, with dummy variables distinguishing daily exposure of (i) 60 minutes or more and (ii) 30 to 59 minutes from lesser amounts; (c) when students began their work with DI with a dummy variable distinguishing those who started in preschool or kindergarten from those who began at Grade 1 or later; (d) for comparisons that involved follow-up or maintenance data, the amount of time (measured in months) that had elapsed between the end of the intervention and the assessment; and (e) subject matter, with a categorical variable distinguishing reading, math, language, spelling, multiple and other academic subjects, ability (IQ) measures, affective measures (e.g., self-esteem, behavior), and teachers’ and parents’ attitudes or views.

Control Variables

Two control variables were included, both measured at the effect, or comparison, level of analysis: (a) a dummy variable indicating if the comparison involved postintervention (maintenance) data and (b) the natural log of the number of cases used to calculate the effect. (The value was transformed because the raw data had a large amount of positive skew.)

Analysis

Estimates of effects for both the total sample and within nine mutually exclusive subareas were examined. Eight of the subareas involved the subject matter of the dependent measure as described above, and the ninth included all of the single-subject designs regardless of subject matter. We chose to examine single-subject designs separately because initial analyses indicated that they had average effect sizes that were substantially larger than other designs and they also differed systematically, as would be expected, in methodological and sample characteristics. Thus, examining them separately resulted in smaller standard errors and more precise estimates.

We began our analysis by examining descriptive statistics, both to understand the characteristics of the sample and to determine if there were variables that did not have sufficient variation to be included in the metaregressions. We examined correlation matrices to explore the presence of any collinear relations. We then used multivariate mixed linear models with random effects, applying the analysis to both the total group and each of the subareas. This approach is considered especially appropriate for meta-analyses that include studies with "nested" results, that is, with potentially multiple effects within a given study (Kalaian & Raudenbush, 1996; Raudenbush, 2009; see also Dedrick et al., 2009). All of our analyses used the xtmixed procedure in STATA.

As explained in the Results section, some studies were described in more than one research report. In addition, some studies involved more than one design. To ensure independence of observations and to minimize variability and increase precision, the Level 2 (group) variable in most of the analyses was the study design. When there were two or more reports of a given study we examined the data from those reports together, and when data were given within a research study employing more than one research design, we examined results separately for each of those designs. However, as noted below, we examined consequences of this decision in the sensitivity analysis.

We started our multivariate analysis with a baseline, intercept only model, with study design as the Level 2 variable. The variables described above were then added as predictors.[5] To preserve degrees of freedom each set of variables (e.g., dummy variables pertaining to assessment or those pertaining to grade level) was added as a group in a single analysis of that set of variables. Results from the analysis of descriptive data were used to limit the variables from each set to ensure that there was adequate variation in the predictor variables, with a minimum n of 20 effects required in each category for inclusion as a predictor. Each of these exploratory metaregressions also included, as control variables, the indicator of maintenance data and log of the sample size. When the indicator related to having maintenance data was significant the length of maintenance was substituted as a predictor.

All variables that were significant within these initial analyses were then included in a joint analysis. Variables that were not significant in this joint analysis were then eliminated, resulting in a final, joint reduced model for the total sample and for each of the subareas. We focused most of our examination of results on the magnitude of the intercept, which is equivalent to the average weighted effect size. We looked at how the estimate of effect changed from the baseline, intercept-only models to the joint reduced models and variations from the total sample to the various subareas. We also examined the way in which the various independent variables were related to estimates of effect size. Finally, we conducted a sensitivity analysis, looking at how effect estimates varied across a variety of conditions including the use of study, rather than design, as the Level 2 variable; when alternative methods of estimation were employed; when controls were introduced for data from Project Follow Through or from sites with multiple studies; and when outliers and studies with questionable quality were included.

For the baseline model we calculated the intraclass correlation (ICC), also sometimes called I^2, which reports the percentage of total variation in effects that was between studies (Borenstein et al., p. 117). For the joint reduced models we calculated two measures of model fit: (a) changes in the −2 log likelihood ratio from baseline to the reduced model, which has a chi-square distribution, and (b) the proportion of variation in effect size that was explained by the model (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 200; Raudenbush, 2009, p. 304).

Results

Results are summarized below regarding the search and screening of the literature, characteristics of the identified studies, and then results of the preliminary analyses, the joint models, and the sensitivity analyses.

Search and Screening Results

A total of 549 research reports regarding the effectiveness of DI programs were identified. The University of Oregon library provided extraordinary assistance in finding material but was not able to locate 16 reports (3% of the total), including unpublished manuscripts cited in other publications, a few master's theses, and a dissertation from another country. Twenty reports gave only data that combined the DI program with another intervention, and 10 reports did not provide data for a group with no exposure to the programs (i.e., giving only data on the impact of greater or less exposure). An additional 58 reports did not provide sufficient information to calculate effects. Of the remaining 445 reports, 15 were excluded because they were judged to have serious methodological problems and 34 were excluded because they had more moderate methodological issues. The majority of reports judged to have quality issues were master's theses or doctoral dissertations. Serious issues involved very unclear or inconsistent labeling and reporting of results; questionable implementation procedures, such as unequal implementation support or using programs that were highly inappropriate for a given grade level; or, in one case, extensive criticism within the published literature. More moderate issues generally involved cases where substantial information was missing or an inappropriate analysis technique appeared to have been used. Finally, three reports were excluded because all of the calculated effects were outliers. In some cases, only some of the elements of a report were deemed as having quality issues or had results that were outliers. Figure 1 gives numbers omitted at each stage, and Appendix B (in the online version of the journal) has citations of all studies examined.

Several of the reports that met the inclusion criteria (n = 12 or 3%) included results of more than one study. About a third of the reports (n = 134 or 34%) included information given in another report, such as summarizing results of a dissertation in a journal article or findings from a journal article in a newsletter or other gray literature. In other cases one report might focus on results immediately after an intervention, while subsequent reports gave data regarding follow-up periods. In addition, individual studies could incorporate more than one design.

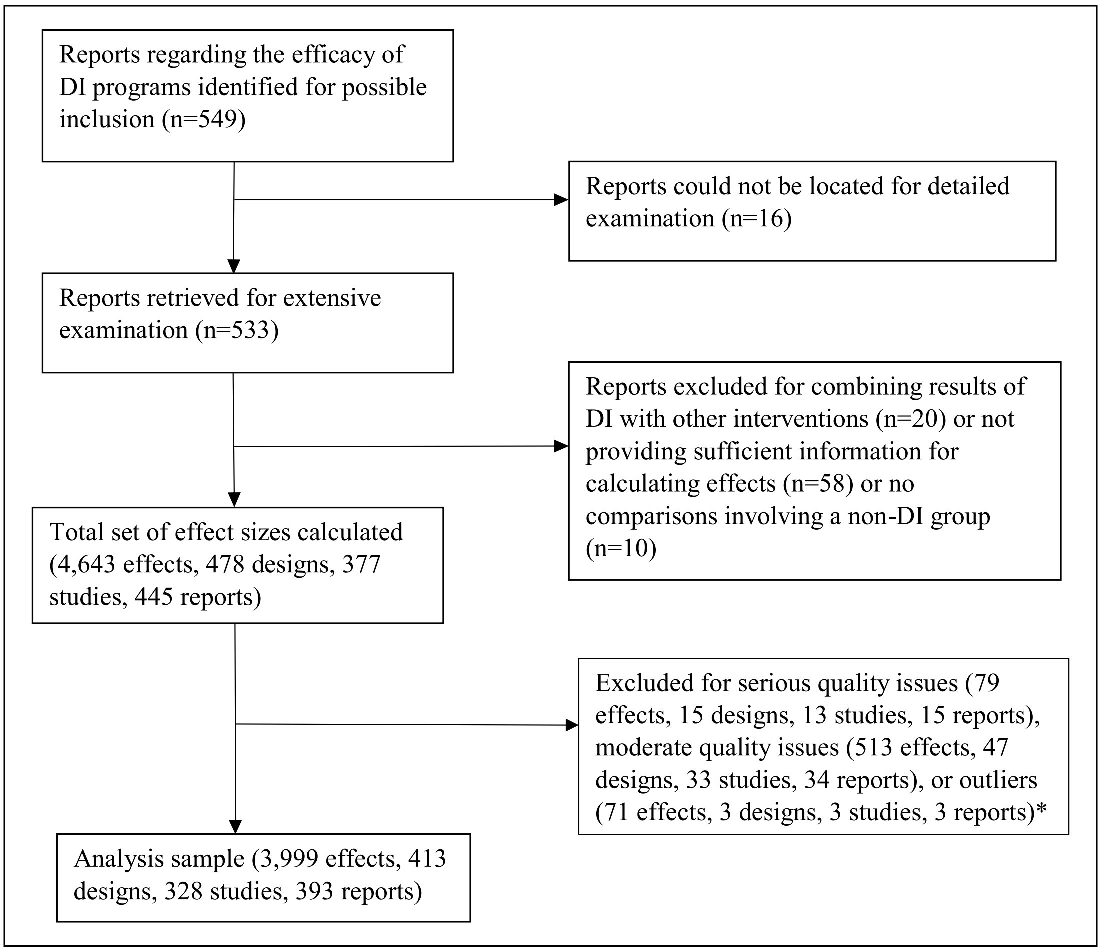
Reports regarding the efficacy of DI programs identified for possible inclusion (n=549)
Reports could not be located for detailed examination (n=16)
Reports retrieved for extensive examination (n=533)
Reports excluded for combining results of DI with other interventions (n=20) or not providing sufficient information for calculating effects (n=58) or no comparisons involving a non-DI group (n=10)
Total set of effect sizes calculated (4,643 effects, 478 designs, 377 studies, 445 reports)
Excluded for serious quality issues (79 effects, 15 designs, 13 studies, 15 reports), moderate quality issues (513 effects, 47 designs, 33 studies, 34 reports), or outliers (71 effects, 3 designs, 3 studies, 3 reports)*
Analysis sample (3,999 effects, 413 designs, 328 studies, 393 reports)

FIGURE 1. *Diagram of search and screening steps to identify studies of the efficacy of Direct Instruction (DI) programs. *In some cases only part of the analysis within a study (i.e., some effects or some designs) were marked as having questionable quality; and in one case two reports involving the same study were marked as questionable. Nineteen of the 71 effects that were outliers were also tagged for quality issues. Four of the outliers were negative.*

From the 393 reports that met the inclusion criteria 328 studies were identified.[6] These studies involved 413 designs and had a total of 3,999 effects. Many of the studies included data regarding more than one subject area. The bottom lines in each panel of Table 1 report the number of effects, studies, and designs within the total sample and each of the subareas. The largest number (226 studies with 269 designs) involved reading, followed by math, language, and spelling. The smallest number of studies (n = 17) involved views of teachers and parents. Appendix B (in the online version of the journal) has citations for all reports on the initial list, any reasons for exclusion, and for those for which effects could be calculated, information on sample size, study design, number of calculated effects, and the minimum, maximum, and average effect. This listing is given for the total sample (Appendix Table B1) and for each of the subject areas (Appendix Tables B2 to B10).

TABLE 1

Initial estimates of effects, total group and subareas

Total group, reading, math, language, and spelling

Measure	Total group	Reading	Math	Language	Spelling
Estimate	0.54***	0.51***	0.55***	0.54***	0.66***
SE	0.03	0.03	0.05	0.07	0.08
95% Confidence interval	[0.49, 0.59]	[0.44, 0.57]	[0.46, 0.65]	[0.40, 0.69]	[0.50, 0.82]
Intraclass correlation	.50	.58	.49	.41	.53
Log likelihood	−3000.72	−1246.50	−475.92	−307.45	−258.69
No. of observations	3,999	1,896	685	299	299
No. of studies	328	226	70	56	52
No. of designs	413	269	91	67	60
Minimum observations/ designs	1	1	1	1	1
Maximum observations/ designs	195	79	65	33	51
M observations/designs	9.7	7	7.5	4.5	5

Multiple subjects, single-subject designs, and nonacademic areas

Measure	Other and multiple	Single-subject	Ability	Affective	Teacher/ parent
Estimate	0.41***	0.83***	0.34***	0.33***	0.40***
SE	0.08	0.07	0.09	0.07	0.11
95% Confidence interval	[0.25, 0.57]	[0.70, 0.97	0.16, 0.52]	[0.19, 0.47]	[0.18, 0.63]
Intraclass correlation	.67	.33	.67	.43	.37
Log likelihood	−144.88	−123.39	−96.23	−82.56	−88.81
No. of observations	224	168	161	167	100
No. of studies	40	38	34	26	17
No. of designs	47	38	39	26	18
Minimum observations/ designs	1	1	1	1	1
Maximum observations/ designs	24	27	29	55	12
M observations/designs	4.8	4.4	4.1	6.4	5.6

Note. Estimates are taken from models that had designs as the Level 2 unit of analysis.
****p* < .001.

Study Characteristics

There was substantial variability on the measures described above. Descriptive statistics for the total group and subareas are in Appendix A (in the online version of the journal). Over half of the studies were published in the past two decades. More than half were in articles and slightly less than a fifth in dissertation or thesis projects. Fifteen percent were disseminated by the major publisher of the DI programs, 12% came from sites associated with Project Follow Through, and 5% or less involved studies written by the authors of this article or involved sites that were the focus of multiple studies.

There was substantial variability in methodological characteristics. The effects were produced from a variety of research designs: almost one fourth from norm or goal control group designs, one fifth from studies with random selection or matching of individual students, and slightly fewer using statistical controls or cohort control groups. Almost three fourths of the effects were calculated from data obtained from normed or other published assessments, with the remainder about equally divided between curriculum-based measures, state assessments, and researcher-designed instruments. The majority of effects were calculated from continuous data, and close to half of the effects adjusted for pretest differences between groups.

Sample-related variables also indicated substantial variability. The majority of studies involved more than one classroom, and about one tenth included more than one district. About half of the studies occurred in urban regions. Studies were based in all parts of the United States as well as in other countries. A quarter of the studies noted that students were from high-poverty backgrounds. About a third of the effects involved students with some type of special need, most often in remedial or special education or having low skills. Only 5% of the effects specifically focused on ELL students, and close to a fifth focused on African American students. Effects were calculated for students from preschool through college age and adult, but the majority involved students at Grade 3 and younger.

The classroom teacher delivered the intervention in about two thirds of the studies, and the comparison program was a "usual" practice in three fifths of the studies. Teachers were reported to have been trained in almost two thirds of the studies and coached in almost half. A wide range of programs were used, with *DISTAR* and *Reading Mastery*, the current K–5 basal reading program, being most common, followed by *Corrective Reading*, a program for older students with low reading skills, and multiple programs. There was substantial variability in the amount of exposure students had to the program, averaging 1.9 years but ranging from a few days to 6 years. Almost two fifths of the comparisons involved daily exposure of 60 minutes or more to a program, and almost a third involved students who began the program in kindergarten. Thirteen percent of the effects involved maintenance, or postintervention, data, with the average length of the maintenance period equaling a little less than 6 months.

The number of observations used to calculate effects ranged from one in a few single-subject designs to close to 45,000 in some of the very large Follow Through analyses. The average number of cases associated with an effect size was 1,456, but the median was 71, reflecting the extreme positive skew of the raw distribution. As would be expected, however, the distribution of the logged value used in the metaregressions was very close to normal ($M = 4.69$, $Mdn = 4.26$).

The frequency distributions differed slightly from one subarea to another. Thus, as noted above, we carefully analyzed the frequency distributions on each variable within each subarea as part of the initial metaregression procedures.

We also examined correlations among the variables both for the total group and in the subareas. (The full correlation matrix is available on request.) When very strong relationships (.50 or larger) appeared between two or more variables tagged for inclusion in a joint model, a summary variable was substituted. This affected analyses with two of the subareas. In the analysis of effects associated with reading a summary variable combined indicators of using a statistical control design, calculating results from statistical data, and adjusting results for pretest demographic data, with a value of 1 indicating that any of these adjustments had been made ($\alpha = .87$). In the analysis of teacher and parent views, several of the variables identified for initial inclusion (poverty, teacher-delivered, Grade 3, and intervention length) were all strongly related to inclusion in Project Follow Through (correlations ranging from .60 to .82), so the dummy variable related to Follow Through was used in the joint reduced model.

Initial Estimates of Effects

The first lines of each panel of Table 1 give the initial estimates of effects derived from the baseline, intercept-only models. The top panel gives results for the total group and the subareas of reading, mathematics, language, and spelling. The bottom panel gives results for other and multiple academic subjects, single-subject designs, ability/IQ measures, affective outcomes, and teacher and parent views. The first line in each panel gives the estimated effect, essentially the weighted average effect size across all studies; the second line gives the associated standard error; and the third gives the 95% confidence interval around the estimated effect. All of the estimates were statistically significant at well beyond the .001 level of significance. The largest estimates were for the single-subject designs (.83) and spelling (.66), and the smallest were for the ability and affective measures (.34 and .33, respectively).

The ICC, or I^2, indicates that there was substantial variability between the studies. Within the total sample about half of the total variation in effect sizes was between the study designs (ICC = .50). The most consistent results appeared with the single subject designs (ICC = .33), and the most variable were those involving ability measures and multiple academic subjects (ICC = .67).

Metaregression Results

Table 2 gives the model fit statistics for the final reduced joint models for the total sample and each subarea and Table 3 reports the fixed effect coefficients. The results of the −2 log likelihood comparisons (the first four columns in Table 2) indicate that the joint reduced models provided a significantly better fit in all comparisons but the one involving teacher and parent views, a model that included only one independent measure and a control variable.[7] The proportion of between–study design variation explained by the models (the last column in each row of Table 2) differed substantially from one analysis to another, with a low of virtually zero for the analyses of multiple and other subjects, affective outcomes, and teacher and parent views to a high of .32 for spelling.

The estimates for the constant terms and the associated standard errors in the last rows of Table 3 indicate the average weighted effect size net of the other variables within the models. As with the initial estimates, all of the joint model estimates were positive and all but one surpassed the common criterion for educational significance of .25. The estimate for the total sample was .60, with the 95% confidence interval ranging from .54 to .66. Within the subareas, the value for the

TABLE 2

Model fit statistics by group

	−2 Log likelihood				Residual variance		
Subject	Model	Baseline	Change	Degrees of freedom	Baseline	Model	Proportion change
Total group	5749.4	6001.4	252.0	16	0.22	0.20	.06
Reading	2325.3	2493.0	167.7	12	0.16	0.15	.07
Math	799.5	951.8	152.3	10	0.18	0.15	.18
Language	578.0	614.9	36.9	4	0.35	0.31	.10
Spelling	407.4	517.4	110.0	5	0.24	0.17	.32
Multiple and other	242.3	289.8	47.5	5	0.14	0.13	.02
Single-subject	212.6	246.8	34.1	5	0.20	0.20	.04
Ability	116.9	192.5	75.6	9	0.13	0.12	.04
Affective	151.7	165.1	13.4	3	0.13	0.13	−.02
Teacher/parent	173.0	177.6	4.7	2	0.27	0.27	.00

Note. The change in −2 log likelihood values from baseline to the joint model was significant at the .001 level in all cases but the analyses of affective measures and teacher/parent views. The change in the −2 log likelihood statistic for affective measures was significant at the .01 level, and the analysis of teacher/parent views was significant at the .10 level.

affective measures (.14) was not statistically significant. Values for all other subareas were statistically significant and ranged in magnitude from .37 for language to over 1.0 for spelling, single-subject designs, ability measures, and teacher and parent views.

Most of the publication-, methodological-, and sample-related variables were not included in the final joint reduced models or were significant in only one analysis. Three variables were included in more than two of the joint reduced models, but the nature of the relationship varied from one subarea to another. Effect sizes based on normed assessments were significantly larger in the analysis for the total group and language but significantly smaller for analyses of math, spelling, and single-subject data. Effects calculated from odds ratios were significantly smaller for the total group and single-subject analyses but significantly larger for the analysis for multiple and other academic subjects. And effects for urban samples were significantly smaller for language achievement but significantly larger in single-subject studies or those involving ability measures. In general, there appeared to be no consistent or strong pattern of association in the metaregressions between publication-, methodological-, or sample-related variables and effect estimates.

Slightly more consistent results appeared with the intervention-related variables, although the pattern varied from one subarea to another. In support of expectations, effect sizes were significantly larger with greater dosage: when students were exposed for more years (for the total group and reading), had a longer period of daily intervention (math), or began their work with DI in kindergarten

TABLE 3

Reduced joint metaregression results, total group and subgroups

Variables	Total group	Reading	Math	Language	Spelling	Multiple/ other	Single-subject	Ability	Affective	Teacher/ parent
Publication related										
Decade 4	—	−.26***	—	—	—	—	—	—	—	—
Decade 5	—	−.18*	—	—	—	—	—	—	—	—
Design-related										
Random assignment	—	−.20*	—	—	—	—	—	−.48***	—	—
Statistical control design	−.33***	—	—	—	—	—	—	—	—	—
Assessment-related										
Normed assessment	.09**	—	−.67***	.56**	−.94***	—	−.38***	—	—	—
Curriculum-based measures assessment	.25***	—	—	—	—	1.72***	—	—	—	—
State assessment	—	—	−.50***	—	—	—	—	—	—	—
Data used in calculations										
Statistical data	.28***	—	.81***	—	—	—	—	—	—	—
Percentiles/normal curve equivalent scores	−.12**	−.31***	—	—	—	—	—	—	—	—
Data adjusted for pretest differences in demographic	—	—	—	—	—	—	—	−.59***	—	—
Statistical adjustment scale	—	−.21**	—	—	—	—	—	—	—	—
Calculations of effects										
Transformation of statistical data	.31***	.52***	—	—	—	—	—	—	.50***	—
From odds ratios	−.17**	—	—	—	—	.80**	−.52***	—	—	—
Extra calculations needed	−.06*	−.11**	—	—	—	—	—	—	—	—
Sample										
More than one class	—	—	.24*	—	—	—	—	−.38**	—	—
More than one school	−.11*	—	—	—	—	—	—	—	—	—
More than one district	—	—	—	—	—	—	—	.80***	—	—

(continued)

TABLE 3 (continued)

Variables	Total group	Reading	Math	Language	Spelling	Multiple/ other	Single- subject	Ability	Affective	Teacher/ parent
Urban area	—	—	—	−.50***	—	—	.30***	.28**	—	—
Western United States	—	—	—	—	—	—	—	−.24**	—	—
American Indian students	—	.61***	—	—	—	—	—	—	—	—
African American students	—	—	—	—	—	—	—	−.36***	—	—
Kindergarten	—	—	.17*	—	.37**	—	—	—	—	—
Grade 4	—	—	.27***	—	—	.64***	—	—	—	—
Project Follow Through	—	—	—	—	—	—	—	—	—	−.54*
Intervention										
DISTAR	−.16*	—	—	—	—	—	—	—	—	—
Reading Mastery/Horizons	−.29***	—	—	—	—	—	—	—	—	—
Multiple programs	−.18**	—	—	—	—	—	—	—	—	—
Intervention length (years)	.06***	.08***	—	—	—	—	—	—	—	—
Daily 60 minutes or more	—	—	.22**	—	—	—	—	—	—	—
Started Direct Instruction in K	.21***	.15***	.43***	—	.55***	—	—	—	—	—
Trained and/or coached	—	—	—	.53***	—	—	—	—	—	—
Maintenance length (months)	—	—	−.01**	−.01**	—	—	—	−.001	—	—
Controls										
Maintenance data	−0.01	−0.10	—	—	−0.20	−0.03	0.06	—	−0.27	–—
Log of *N* in comparison	−.02*	−.04**	−.02	.05	.06	−.05	−.19**	−.08***	.03	−.10
Constant										
Estimate	0.60***	0.74***	0.75***	0.37*	1.07***	0.52***	1.02***	1.18***	0.14	1.03**
SE	0.06	0.08	0.12	0.19	0.21	0.15	0.10	0.14	0.15	0.37

Note. *DISTAR* = Direct Instruction System for Teaching Arithmetic and Reading.
*p < .05. **p < .01. ***p < .001.

(total group, reading, math, and spelling). However, these significant relationships involved only the core academic subjects and not the other subareas that were examined. The expected significant relationship of effects with teacher training and coaching was found only in the analysis of language. Effect sizes in maintenance periods were significantly lower in two analyses (math and language), but the magnitude of these coefficients was relatively small (0.01).[8] Effects were significantly smaller when the sample size was larger in only 3 of the 10 analyses (total group, reading, and ability measures).

In general, relatively few of the independent variables tested in the metaregressions were significantly related to the effect sizes. Most important, when adjusted for variations in these independent variables, the estimates of effects remained substantial. In 8 of the 10 analyses (all but in the subareas of language and affective measures), the adjusted effect sizes given in Table 3 were larger than the initial estimates shown in Table 1. The only adjusted effect size that was not significant was the one associated with affective measures.

Sensitivity Analyses

As the final step in our analysis we examined the extent to which our results were sensitive to a variety of methodological constraints. The coefficients associated with the intercept (the estimate of the effect size) for the additional tested models are in Table 4. The first four rows contrast the results obtained with different groupings at level two of the mixed model: study design (the results shown in Tables 1, 2, and 3) and study (collapsing designs within studies).[9] The next set of results report estimates when alternative estimation methods were used (restricted maximum likelihood estimates and an unstructured covariance matrix, rather than maximum likelihood estimates and independent covariance structures). This is followed by estimates when control variables were added to the model for studies using data from Project Follow Through sites and for sites that were examined by several different authors or studies. Finally, we examined results when a larger sample of effects was used, including the outliers and studies for which there were quality concerns.

When outliers were included in the sample, several of the effects, especially those involving the total group, reading, math, spelling, and single subjects, were substantially larger. This outcome would be expected given the overwhelming preponderance of positive values in the distribution of the outliers.[10] In all other cases the estimates from the joint reduced models were very similar. All of the 117 estimates included in the table were positive, and all estimates, except for those from the joint models regarding affective outcomes, exceeded the common .25 threshold of educational significance. In addition, all but those involving the joint models related to affective measures and some of those regarding language were statistically significant. Of the 104 significant results, the vast majority (*n* = 88) were significant at the .001 level.

Discussion

The classic methodological literature emphasizes the importance of cumulating evidence to develop scientific conclusions (e.g., Popper, 1962). As Cook and Campbell, authors of the most widely cited works on research design, put it, “We

TABLE 4

Sensitivity analyses: estimated effects by model, total group, and subgroups

Model	Total group	Reading	Math	Lang.	Spelling	Multiple/ other	Single-subject	Ability	Affective	Teacher/ parent
Intercept-only models										
Intercept-only, Design Level 2	0.54***	0.51***	0.55***	0.54***	0.66***	0.41***	0.83***	0.34***	0.33***	0.40***
Intercept-only, Study Level 2	0.55***	0.50***	0.59***	0.54***	0.72***	0.40***	0.83***	0.30***	0.33***	0.39***
Reduced models										
Design Level 2	0.60***	0.74***	0.75***	0.37*	1.07***	0.52***	1.02***	1.18***	0.14	1.03**
Study Level 2	0.52***	0.72***	0.70***	0.32	0.97***	0.60***	1.02***	1.18***	0.14	1.04**
Reduced models, alternative estimation methods										
Unstructured covariance	0.60***	0.74***	0.75***	0.37*	1.07***	0.52***	1.02***	1.18***	0.14	1.03**
Restricted maximum likelihood	0.60***	0.74***	0.75***	0.37*	1.06***	0.50**	1.02***	1.21***	0.12	1.02*
Reduced models, additional controls										
Project Follow Through	0.61***	0.73***	0.77***	0.40*	1.07***	0.46***	—	1.16***	0.16	—
Site 1	0.60***	0.73***	0.69***	—	—	0.50**	—	—	—	—
Site 2	0.60***	0.74***	—	—	—	—	—	—	—	—
Site 3	0.60***	0.74***	—	—	—	—	—	—	—	—
Site 4	0.61***	0.74***	—	—	—	0.51***	—	—	—	—
Site 5	0.61***	0.74***	0.75***	0.37*	1.05***	0.52***	—	1.15***	0.15	—
Reduced models, larger sample including										
Outliers	0.77***	0.81***	1.12***	0.39	1.30***	0.52***	1.59*	.98***	0.14	1.03**
Moderate quality issues	0.58***	0.65***	1.00***	0.22	1.14***	0.51***	1.08***	0.96***	0.12	0.94**
Serious or moderate quality issues	0.57***	0.66***	1.00***	0.21	1.11***	0.40**	1.07***	0.96***	0.11	0.94**

Note. All models, except as noted, used maximum likelihood estimates, independent covariance structure, and study design as the Level 2 variable. Models that included both study and design as level variables did not converge.

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

stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results" (Cook & Campbell, 1979, p. 31, emphasis in original). Later writings by the authors emphasize the importance of cumulating findings that allow one to "generalize . . . inferences . . . over variations in persons, settings, treatments and outcomes" (Shadish, Cook, & Campbell, 2002, p. 32). The results presented above could be seen as meeting this goal. They were derived from 328 studies and almost 4,000 calculated effects. They involved general academic achievement, as well as specific subjects, measures of ability, affective outcomes, and teacher and parent views. Both group and single-subject analyses were included. The studies appeared over a 50-year period and involved a wide range of subjects, settings, comparison groups, and methodological approaches.

Our results support earlier reviews of the DI effectiveness literature. The estimated effects were consistently positive. Most estimates would be considered medium to large using the criteria generally used in the psychological literature and substantially larger than the criterion of .25 typically used in education research (Tallmadge, 1977). Using the criteria recently suggested by Lipsey et al. (2012), 6 of the 10 baseline estimates and 8 of the 10 adjusted estimates in the reduced models would be considered huge. All but one of the remaining six estimates would be considered large. Only 1 of the 20 estimates, although positive, might be seen as educationally insignificant.

Except for the analysis of affective measures, estimates remained statistically and substantively significant when a broad range of control variables were introduced. The initial estimates suggested that effects were weaker for ability measures and views of teachers and parents, but with the introduction of control variables in the metaregressions the estimates in these areas increased substantially to equal or surpass those associated with other subareas. The only areas in which estimates declined from the initial estimates to the joint models were language and affective measures.

The metaregressions and sensitivity analyses indicated that the results were robust, with no systematic impact of variables related to the nature of the publication, methodological approach, or sample. The strong positive results were similar across the 50 years of data; in articles, dissertations, and gray literature; across different types of research designs, assessments, outcome measures, and methods of calculating effects; across different types of samples and locales, student poverty status, race-ethnicity, at-risk status, and grade; across subjects and programs; after the intervention ceased; with researchers or teachers delivering the intervention; with experimental or usual comparison programs; and when other analytic methods, a broader sample, or other control variables were used. This conclusion regarding consistent results is bolstered by comparing the ICC values (Table 1) to the proportion of residual variance explained by the joint models (Table 2). Such a comparison shows that except for the analyses of spelling and mathematics, very little of the variance in effect estimates between studies was explained by variables examined in the metaregressions. In other words, the results were very consistent across all of the variables examined.

Earlier literature had led us to expect that effect sizes would be larger when students had greater exposure to the programs, and this hypothesis was supported for most of the analyses involving academic subjects. Significantly stronger results appeared for the total group, reading, math, and spelling for students who began the programs in kindergarten; for the total group and reading for students who had more years of intervention; and for math students with more daily exposure. Although we had expected that effects could be lower at maintenance than immediately postintervention, the decline was significant in only two of the analyses (math and language) and not substantial in either. Similarly, while literature across the field of education has suggested that reported effects would be stronger in published than in unpublished sources (Polanin et al., 2016), we found no indication of this pattern.

Contrary to expectations, training and coaching of teachers significantly increased effects in only one analysis (language). We suggest that readers interpret this finding cautiously for we suspect that it reflects the crude nature of our measure—a simple dummy variable noting if teachers were reported as receiving any training or coaching. We hypothesize that a more precise measure of teacher preparation, including fidelity to all the various technical elements of the programs and training specific to the programs taught, would produce different results (see Benner, Nelson, Stage, & Ralston, 2010; Carlson & Francis, 2002; Gersten, Carnine, & Williams, 1982; Gersten, Carnine, Zoref, & Cronin, 1986; Ross et al., 2004; Stockard, 2011; see also Kennedy, 2016).

Limitations and Future Research Directions

Even though our study involved a comprehensive examination of a very large data set, it was not without its limitations, many of which reflect the size and heterogeneity of the sample. For instance, we did not attempt to compare the results of each of the DI programs with specific other approaches. Nor did we examine outcomes in subdimensions within the various subject areas, such as differentiating reading fluency and comprehension. In addition, many of our measures were less precise than could be considered optimal. The studies differed, often substantially, in the nature and amount of information given. To preserve degrees of freedom we included cases with missing data on a measure in the reference category. Meta-analyses that focus on much smaller parts of the literature could, potentially, include such information and have more precise measures. Our study was also, of course, limited by the range of information that was available. Although the number of studies and designs included in each of our analyses was not small, it is clear that there was more information about some subareas than others, particularly those involving nonacademic outcomes.

Even though the literature on the effectiveness of DI is substantial and consistent, there are areas in which additional work could be helpful and informative. For instance, there were many more studies regarding reading outcomes than other areas. Studies of mathematics, especially of programs aimed toward older students, such as Essentials of Algebra and Corrective Mathematics, were especially rare, as were studies, at least in recent decades, of programs that teach English language skills to those with other first languages. Studies of preschool students were less common than those of primary grade children. Most of the results regarding affective outcomes and teacher and parent views were reported as ancillary information in studies of academic outcomes. The field could be well served by studies that explicitly focus on these important nonacademic areas. The relationship between academic growth with exposure to DI, student affective outcomes, and teacher and parent views has been addressed theoretically (e.g., Engelmann, 2014c). However, at least to our knowledge, it has not been subjected to systematic empirical analysis.

Implications for Policy and Practice

The findings of this meta-analysis reinforce the conclusions of earlier meta-analyses and reviews of the literature regarding DI. Yet, despite the very large body of research supporting its effectiveness, DI has not been widely embraced or implemented. In part this avoidance of DI may be fueled by the current popularity of constructivism and misconceptions of the theory that underlies DI. As explained in the first part of this article, DI shares with constructivism the important basic understanding that students interpret and make sense of information with which they are presented. The difference lies in the nature of the information given to students, with DI theorists stressing the importance of very carefully choosing and structuring examples so they are as clear and unambiguous as possible. Without such clarity students will waste valuable time and, even worse, potentially reach faulty conclusions that harm future progress and learning.

Many current curriculum recommendations, such as those included within the Common Core, promote student-led and inquiry-based approaches with substantial ambiguity in instructional practices. The strong pattern of results presented in this article, appearing across all subject matters, student populations, settings, and age levels, should, at the least, imply a need for serious examination and reconsideration of these recommendations (see also Engelmann, 2014a; Morgan, Farkas, & Maczuga, 2015; Zhang, 2016). It is clear that students make sense of and interpret the information that they are given—but their learning is enhanced only when the information presented is explicit, logically organized, and clearly sequenced. To do anything less shirks the responsibility of effective instruction.

Another reason that DI may not be widely used involves a belief that teachers will not like it or that it stifles teachers' ability to bring their own personalities to their teaching. Yet, as described in earlier sections, proper implementation of DI does not disguise or erase a teacher's unique style. In fact, the carefully tested presentations in the programs free teachers from worries about the wording of their examples or the order in which they present ideas and allow them to focus more fully on their students' responses and ensure their understanding. Recall that effect sizes associated with teachers' perceptions of the program reached as high as 1.04 in our analyses. Fears that teachers will not enjoy the programs or not be pleased with their results do not appear to be supported by the evidence.

Lipsey et al. (2012) have suggested that effect sizes based on performance gaps among demographic groups could be a useful benchmark in evaluating the potential impact of an intervention. Using data from the National Assessment of Education Progress, they calculated performance gaps in reading and math and found that the difference between more and less privileged groups corresponds to effect sizes ranging from 0.45 to 1.04 (Lipsey et al., 2012; p. 30; see also Bloom, Hill, Black, & Lipsey, 2008). These values are quite similar to the effects found in our analysis. In other words, the effects reported in this analysis, and calculated from 50 years of data on DI, indicate that exposure to DI could substantially reduce current achievement disparities between sociodemographic groups. Moreover, as noted above, at least for the academic subjects, greater exposure would be expected to result in even larger effects. There is little indication that the effects would be expected to decline markedly after intervention ceased; the positive effects are long-term.

Certainly our nation's children deserve both effective and efficient instruction. As one of the anonymous reviewers of our article put it, "Researchers and practitioners cannot afford to ignore the effectiveness research on DI."

Notes

Some of the work on this article was completed while the authors were employed on a part-time basis by the National Institute for Direct Instruction, a nonprofit organization that provides support for schools implementing DI programs. We thank Kerry Hempenstall for his assistance in identifying relevant materials; the staff of the University of Oregon Interlibrary Loan Office, Alexa Engelmann, Ricky Carrizales, and Ashly Vanderwall for their assistance in finding literature; and Douglas Carnine, Kurt Engelmann, and anonymous reviewers of this journal for providing feedback on drafts of the analysis. Any errors and all opinions in the article are the sole responsibility of the authors.

[1] Authors who were contacted and/or were the focus of specific bibliographic searches included W. Becker, C. Bereiter, D. Carnine, C. Darch, M. Flores, J. Ganz, R. Gersten, B. Grossen, J. Lloyd, N. Marchand-Martella, R. Martella, M. Vitale, and P. Weisberg.

[2] In a few cases our calculated effects differed from those reported by the authors. This almost always involved our desire to use a consistent effect measure (e.g., *d* rather than eta).

[3] It is possible that there were other studies that involved the same community but could not be discerned from the reports. As explained in the sensitivity analysis, including these variables had no impact on the estimates of effects.

[4] In the subarea analyses in which all single-subject analyses were within one group, the reference group was other posttest-only designs.

[5] The measures of subject matter were not used in the metaregressions as independent variables but instead denoted the subareas used in the analysis. The dummy variables regarding studies related to Project Follow Through and those for which multiple studies involved the same site were added in the sensitivity analysis. Substantive results were identical when these variables were instead used as independent variables in the metaregressions.

[6] Eighteen of the identified studies involved results from Project Follow Through. Sixteen of these involved data from specific sites, one involved analyses with the entire group, and another involved comparisons of DI with alternative programs. We divided the reports of data from Project Follow Through in this manner to provide groupings that were most homogeneous in nature.

[7] There were no studies of teacher/parent views that involved maintenance data, so only the log of the sample size was used as a control.

[8] Calculations based on the coefficients in the reduced joint models indicate that for language, which had a smaller constant, the estimated effect size, all other variables remaining equal, 1 year after the end of intervention would be 0.25 [$0.37 - (12 * 0.01) = 0.37 - 0.12 = 0.25$]; for math the estimated effect 1 year later would be estimated to be 0.63 ($0.75 - 0.12$).

[9] Models did not converge when both study and design were included as level variables, no doubt because of the incomplete nesting of the data set. (Many studies had only one design.)

[10] Sixty-seven of the 71 outlying effects (94%) were positive.

References

Adams, G. (1996). Project Follow Through: In-depth and beyond. *Effective School Practices*, *15*, 43–56.

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

American Institutes for Research. (1999). *An educator's guide to schoolwide reform*. Arlington, VA: Author.

Barbash, S. (2012). *Clear teaching: With Direct Instruction, Siegfried Engelmann discovered a better way of teaching*. Arlington, VA: Education Consumers Foundation.

Becker, W. C., & Gersten, R. (1982). A follow-up of Follow Through: The later effects of the Direct Instruction model on children in fifth and sixth grades. *American Educational Research Journal*, *19*, 75–92.

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.

Bereiter, C., & Engelmann, S. (1966). *Teaching disadvantaged children in the preschool*. Englewood Cliffs, NJ: Prentice Hall.

Bereiter, C., & Kurland, M. (1996). A constructive look at Follow Through results. *Effective School Practices*, *15*, 17–32.

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: Wiley.

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*, 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*, 141–166.

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.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Coughlin, C. (2011). *Research on the effectiveness of Direct Instruction* (NIFDI Technical Report 2011-4). Eugene, OR: National Institute for Direct Instruction

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.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Rev. ed.). New York, NY: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., . . . Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102.

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. (2007). *Teaching needy kids in our backward system: 42 years of trying*. Eugene, OR: ADI Press.

Engelmann, S. (2014a). The dreaded standards. In T. W. Wood (Ed.), *Engelmann's Direct Instruction: Selected writings from the past half century* (pp. 414–421). Eugene, OR: NIFDI Press.

Engelmann, S. (2014b). Research from the inside: The development and testing of DI programs. In J. Stockard (Ed.), *The science and success of Engelmann's Direct Instruction* (pp. 3–24). Eugene, OR: NIFDI Press.

Engelmann, S. (2014c). *Successful and confident students with Direct Instruction*. Eugene, OR: NIFDI Press.

Engelmann, S., Becker, W. C., Carnine, D., & Gersten, R. (1988). The Direct Instruction Follow Through model: Design and outcomes. *Education and Treatment of Children*, *11*, 303–317.

Engelmann, S., & Carnine, D. (1991). *Theory of instruction: Principles and applications* (Rev. ed.) Eugene, OR: ADI Press. (Originally 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.

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

Grossen, B. (1996). The story behind Project Follow Through. *Effective School Practices 15*(1), 4–9.

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

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: Vol. 3. Building a common knowledge base* (pp. 73–98). Mahwah, NJ: Lawrence Erlbaum.

Kalaian, H., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 227–235.

Kennedy, M. M. (1978). Findings from the Follow Through planned variation study. *Educational Researcher*, *7*(6), 3–11.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*, 945–980.

Kinder, D., Kubina, R., & Marchand-Martella, N. E. (2005). Special education and Direct Instruction: An effective combination. *Journal of Direct Instruction*, *5*, 1–36.

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.

MacIver, M. A., & Kemper, E. (2002). The impact of Direct Instruction on elementary students' reading achievement in an urban school district. *Journal of Education for Students Placed at Risk*, *7*, 197–220.

Meyer, L. A. (1984). Long-term academic effects of the Direct Instruction Project Follow Through. *The Elementary School Journal*, *84*, 380–394.

Morgan, P. L., Farkas, G., & Maczuga, S. (2015). Which instructional practices most help first-grade students with and without mathematics difficulties? *Educational Evaluation and Policy Analysis*, *37*, 184–205.

National Institute for Direct Instruction. (2016). *Writings on Direct Instruction: A bibliography*. Eugene, OR: Author. 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.

O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth Independent school district. *Journal of Education for Students Placed at Risk*, *7*, 167–195.

Polanin, J. R., Tanner-Smith, E. E, & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207–236.

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Basic Books.

Przychodzin, A. M., Marchand-Martella, N. E., Martella, R. C., & Azim, D. (2004). Direct Instruction mathematics programs: An overview and research summary. *Journal of Direct Instruction*, *4*, 53–84.

Przychodzin-Havis, A. M., Marchand-Martella, N. E., Martella, R. C., Miller, D. A., Warner, L., Leonard, B., & Chapman, S. (2005). An analysis of *Corrective Reading* research. *Journal of Direct Instruction*, *5*, 37–65.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: Russell Sage Foundation.

Ross, S. M., Nunnery, 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*, 357–388.

Schieffer, C., Marchand-Martella, N. E., Martella, R. C., Simonsen, F. L., & Waldron-Soler, K. M. (2002). An analysis of the *Reading Mastery* program: Effective components and research review. *Journal of Direct Instruction*, *2*, 87–119.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simonsen, F., & Gunter, L. (2001). Best practices in spelling instruction: A research summary. *Journal of Direct Instruction*, *1*, 97–105.

StataCorp. (2011). Stata Statistical Software: Release 12. College Station, TX: Author.

Stockard, J. (2010). Promoting reading achievement and countering the "fourth-grade slump": The impact of Direct Instruction on reading achievement in fifth grade. *Journal of Education for Students Placed at Risk*, *15*, 218–240.

Stockard, J. (2011). Direct Instruction and first grade reading achievement: The role of technical support and time of implementation. *Journal of Direct Instruction*, *11*, 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*, *47*, 2225–2257.

Stockard, J., & Wood, T. W. (2017). The threshold and inclusive approaches to determining "best available evidence." *American Journal of Evaluation*, *38*, 471–492.

Tallmadge, G. K. (1977). *The Joint Dissemination Review Panel idea book*. Washington, DC: National Institute of Education.

Vitale, M. R., & Joseph, B. L. (2008). Broadening the institutional value of Direct Instruction implemented in a low-SES elementary school: Implications for scale-up and reform. *Journal of Direct Instruction*, *8*, 1–18.

Vitale, M. R., & Kaniuka, T. S. (2012). Adapting a multiple-baseline design rationale for evaluating instructional interventions: Implications for the adoption of Direct Instruction reading curricula for evidence-based reform. *Journal of Direct Instruction*, *12*, 25–36.

Watkins, C. (1996). Follow Through: Why didn't we? *Effective School Practices, 15*, 56–66.

White, W. A. T. (1988). A meta-analysis of the effects of Direct Instruction in special education. *Education & Treatment of Children*, *11*, 364–374.

Wood, T. W. (2014). *Engelmann's Direct Instruction: Selected writings from the past half century*. Eugene, OR: NIFDI Press.

Zhang, L. (2016). Is inquiry-based science teaching worth the effort? *Science & Education*, *25*, 897–915.

Authors

JEAN STOCKARD is a quantitative sociologist and professor emerita at the University of Oregon. In addition to sociology of education, her current research projects involve issues related to cohort variations in lethal violence, gender-related climates in the academic sciences, and the role of leisure in mid-life development and change.

TIMOTHY W. WOOD is a master's candidate in historic preservation at the University of Oregon. His previous research examined the history and significance of Direct Instruction, and he currently specializes in cultural resource management, survey and inventory, and section 106 review.

CRISTY COUGHLIN completed her PhD in school psychology at the University of Oregon and has worked as a research consultant and program evaluator for educational projects based in the United States, Australia, and Africa. Her interests include educational assessment, school-based behavioral intervention, and practical application of educational research.

CAITLIN RASPLICA KHOURY received her PhD in school psychology and masters in special education from the University of Oregon. She currently works as a licensed psychologist supporting children and families in a pediatric primary care clinic in Portland, Oregon. Her research interests include early literacy assessment and intervention, school readiness and kindergarten transition, and behavioral interventions for children with disruptive behavior disorders.

View publication stats