

THE SCIENCE AND SUCCESS OF ENGELMANN'S DIRECT INSTRUCTION

EDITED BY JEAN STOCKARD

For almost a half-century, Siegfried Engelmann has shown how all children can learn if they are taught effectively. The Direct Instruction (DI) curricular programs he developed reflect the most stringent requirements of the scientific world. They build on sound theoretical understandings of how effective instruction and learning occur, they involve painstaking attention to each detailed step of the instructional process, and they have been validated with rigorous tests of their efficacy. Engelmann's work has transformed the instructional experience of thousands of students and has also led to noted improvements in school behavioral climates and instructional practices. This book is a tribute to the legacy and genius of Siegfried Engelmann and his decades of work in developing the Direct Instruction (DI) curricular programs.

The authors of the chapters in this book represent several generations and multiple disciplines, bringing a variety of perspectives to their analyses of Engelmann's career and impact. Part I of the book documents the extensive research embodied in the development of DI programs, the research that confirms their effectiveness, the unfavorable and short-sighted reactions of the education establishment to the work, and Engelmann's resilience and strength in continuing to develop programs, write essays and books, and promote learning and effective instruction for all students. Part II examines the legacy of his work, including the guidance it gives for transforming schools into effective learning centers for all children and the ways in which it has influenced the tradition of behavioral management in schools. The book ends with a look at the future, the potential for wider acceptance of Engelmann's developments, and the hope for truly solving the problems of achievement in America's schools. This long-awaited survey of DI's history and impact belongs in the collection of all educational researchers, teachers, college libraries, and interested administrators.

Jean Stockard is professor emerita at the University of Oregon and director of research and evaluation for the National Institute for Direct Instruction. She is the author of numerous books in the areas of education and sociology, including *Effective Educational Environments* and *Sociology: Discovering Society*.



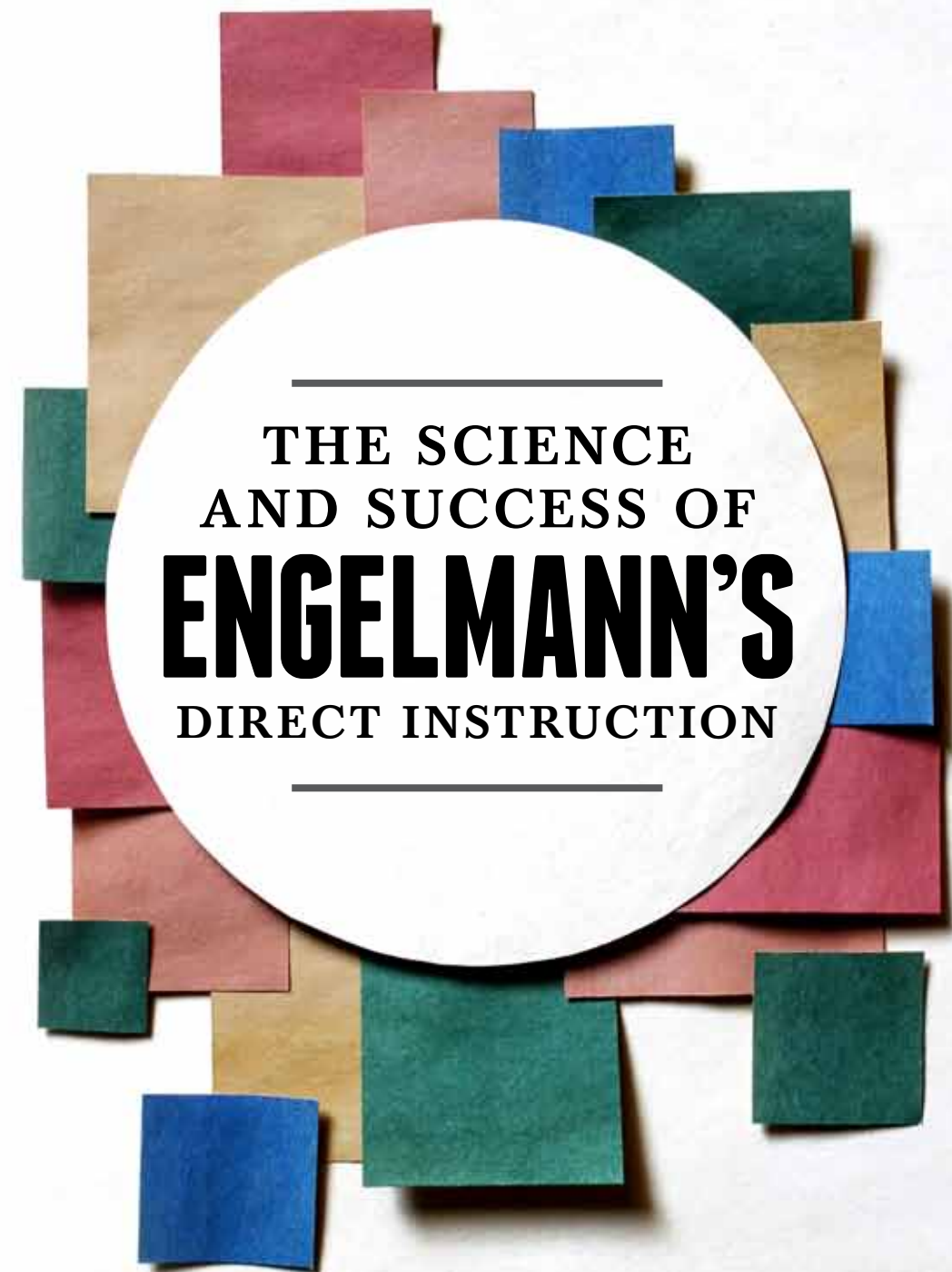
NIFDI Press is a division of the
National Institute for Direct Instruction

805 Lincoln, Eugene, Oregon 97401, USA
877.485.1973 • nifdi.org

THE SCIENCE AND SUCCESS OF ENGELMANN'S
DIRECT INSTRUCTION



NIFDI
PRESS



THE SCIENCE AND SUCCESS OF ENGELMANN'S DIRECT INSTRUCTION

EDITED BY JEAN STOCKARD

**THE SCIENCE AND SUCCESS OF ENGELMANN'S
DIRECT INSTRUCTION**

Edited by Jean Stockard

Design and Marketing Editor: Christina Cox

Copy Editor: Tina Wells

Design: Beth Wood Design

Marketing Coordinator: Courtney Burkholder

The Science and Success of Engelmann's Direct Instruction

Edited by Jean Stockard

Copyright © 2014 by NIFDI Press

A division of the National Institute for Direct Instruction

805 Lincoln

Eugene, Oregon 97401-2810

USA

All rights reserved. No part of the work covered by this copyright hereon may be reproduced or used in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems – without the written permission of the publisher.

Printed in the United States of America

1 2 3 4 5 6

ISBN 978-1-939851-00-0

CONTENTS

Foreword <i>Stan Paine</i>	v
Introduction <i>Jean Stockard</i>	viii
Part I: The Scientific Basis of Direct Instruction	1
1. Research from the Inside: The Development and Testing of DI Programs <i>Siegfried Engelmann</i>	3
2. Outcomes of Engelmann's Direct Instruction: Research Syntheses <i>Cristy Coughlin</i>	25
3. Blinded to Evidence: How Educational Researchers Respond to Empirical Data <i>Siegfried Engelmann and Jean Stockard</i>	55
4. The Engelmann Corpus of Writings <i>Timothy W. Wood</i>	79
Part II: Translating the Science to Schools	97
5. Creating Effective Schools with Direct Instruction <i>Kurt E. Engelmann</i>	99
6. Direct Instruction and Behavior Support in Schools <i>Caitlin Rasplica</i>	123
7. Debating DI's Future <i>Shepard Barbash</i>	141
Appendices	
A. Engelmann's Direct Instruction Programs: An Annotated Bibliography <i>Timothy W. Wood</i>	163
B. Engelmann's Other Writings: An Annotated Bibliography of Engelmann's Articles, Chapters, and Books <i>Timothy W. Wood</i>	183
C. A Chronology of Engelmann's Career Highlights <i>Timothy W. Wood</i>	241
Author Biographies	246
Acknowledgements	249
Subject Index	250
Name Index	254

CHAPTER 3

The previous two chapters have documented the ways in which DI programs are research-based and validated. In Chapter 1, Engelmann described how the systematic development of DI programs incorporates logical reasoning and ongoing research. Chapter 2 summarized the large amount of empirical research that has documented the efficacy of DI programs and how students make much stronger gains with these curricula than with other programs. Yet the educational community continues to ignore this work, to the detriment of millions of children and their readiness to participate as educated citizens of our society. In this chapter Engelmann and Stockard reflect on this failure. Drawing on their individual research backgrounds, they describe several instances in which educational researchers have ignored the empirical evidence, discuss how this contradicts scientific traditions and logic, and examine how it harms students and the society as a whole.

BLINDED TO EVIDENCE HOW EDUCATIONAL RESEARCHERS RESPOND TO EMPIRICAL DATA

Siegfried Engelmann and Jean Stockard

Both of us have spent decades involved in research, albeit in different academic fields and using different methodological approaches and theoretical traditions. Despite the differences in our academic backgrounds and experiences, our conclusions regarding educational research are very similar. This chapter describes our views and, particularly, the ways in which we have seen educational researchers misconstrue and ignore the empirical data related to Direct Instruction. This chapter has three parts. The first, Is the “Gold Standard” Really Gold?, authored by Engelmann, describes the educational research establishment’s current fascination with so-called “gold standard” research and the ways in which political, professional, or personal bias overrides the research evidence. The second part, A Social Scientist’s View of Educational Research, authored by Stockard, delineates norms that are common to science, describes the ways in which the DI corpus of research conforms to these norms, and presents two very costly examples of educational researchers ignoring these norms and the empirical results. In the conclusion, we reflect on possible reasons for these actions and the costs to students and society.

IS THE “GOLD STANDARD” REALLY GOLD?

S. Engelmann

Much has been made in recent years of a “gold standard” in research, urging the educational community to focus almost exclusively on randomized control trials, supposedly emulating medical research. In reality, when well-conducted studies fail to support favored scenarios or when they support programs that are not in favor, the data are ignored. The first section below, Gold Standard Failure, describes such instances. The second section, Quest for Pristine Internal Validity of Educational Studies, questions the worth of this “gold standard” approach and demonstrates why it is futile.

GOLD STANDARD FAILURE

Our view of the literature is certainly influenced by our perspective about developing and field-testing instructional programs and by the various “low probability”

demonstrations that we have conducted, described in Chapter 1. From this viewpoint, it seems that those who screen and evaluate studies that compare specific instructional approaches set criteria that mimic medical research in a highly selective manner. For instance, medical research endorses a gold standard largely for efficacy studies (tightly controlled lab experiments), but education endorses it for all studies. The centerpiece of the medical gold standard is random assignment of subjects. However, there's also the requirement of blind or double-blind conditions, which means that neither the subjects nor those who analyze the data know whether a given subject is in the experimental group or is receiving a placebo.

The double-blind criterion is not easily satisfied in educational experiments that compare programs. Certainly studies that may be completed in two weeks could be conducted in a “lab” setting. There may also be hope of conducting some educational double-blind, “gold standard” experiments; however, it is unlikely studies involving year-long programs can meet the double-blind requirements. Imagine students in a school who don't know that they are in a program quite different from the program students in the grade above them went through the last year. Consider the improbability of a teacher who doesn't know that this year's program is not the one the school used last year.

Possibly duplicity in how results are used provides the greatest difference between medical and educational research. If a gold-standard medical study provides conclusive evidence that something works, it is promoted; if the study provides evidence that something does not work, the ineffective process is publicly rejected. In contrast, if a gold-standard educational study provides conclusive evidence that a favored program worked, there would be a strong effort to publicize this fact; however, if the study showed that a favored approach failed, there would be no attempts to publicize or to caution use of the approach that caused the unfavorable outcomes. Also, differential treatment occurs if an unfavored approach has evidence of effectiveness. If studies of an unfavored approach do not meet gold standards but have been reported in respected journals, the evidence is rejected.

There are four large studies that support these assertions:

1. The national evaluation of Project Follow Through (Kennedy, 1978; Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977);
2. The Rodeo Institute for Teacher Excellence (RITE) studies in Dallas (Carlson & Francis, 2003);
3. The state of Tennessee's Student/Teacher Achievement Ratio (Project STAR) study on class size (Word et al., 1990, 1994); and

4. Chicago's Striving Reader Initiative (Metis Associates, 2011).

Tale of Two Successful Unfavored (and Unpublicized) Approaches

The reactions of the educational establishment to two very large and successful implementations of Direct Instruction—Project Follow Through and the Rodeo Institute work—illustrate how highly successful, but unfavored, approaches are ignored and unpublicized.

Project Follow Through

The evaluation of Project Follow Through (the largest educational research study ever conducted) compared the performance of 22 different approaches to teaching at-risk students in Grades K through 3. Direct Instruction was the only approach in which students had positive gains in all of the outcomes measured, both academic and affective, and the DI model produced the highest scores on all these outcomes (Bereiter & Kurland, 1981–82; Stebbins et al., 1977; Watkins, 1997). Long-term follow-up studies documented lasting impacts of the program, as students in the DI program maintained higher levels of achievement through the high school years and were more likely to finish high school and go on to college (Meyer, 1984). There is no scientific basis for rejecting Follow Through outcomes; yet, the study was ignored at its completion and has been rejected by the What Works Clearinghouse (WWC) on the grounds that it is too old to be valid (WWC, 2007).

Rodeo Institute for Teacher Excellence (RITE)

More recently, the Rodeo Institute for Teacher Excellence (RITE) funded an extensive implementation of Direct Instruction reading programs for at-risk students in the primary grades. RITE also conducted evaluations of the programs that involved over 9,300 students and 277 teachers. Again, the results were overwhelmingly positive (Carlson & Francis, 2003). The study has been rejected by the WWC, and the intervention has received no recognition as a highly successful effort (WWC, 2007). RITE and the Follow Through studies involved an “unpopular” program (Direct Instruction), and the studies were rejected.

Tale of Two Unsuccessful Favored Approaches

The state of Tennessee's Student/Teacher Achievement Ratio (STAR) study of class size (Word et al., 1990, 1994) and the Chicago Striving Reader Initiative involve approaches that are in accord with currently popular views about causes of academic problems and the solutions. Both these studies were “gold standard” in that both used random assignment of subjects. Yet educators have drawn conclusions from these studies that are not supported by the evidence.

Project STAR

The STAR study involved reducing the number of students in regular classrooms to 23. The total number of students in the study was 1,650, over a four-year period. The study used random assignment of classrooms in the same school, with half the classes smaller than the others. Students in the smaller classes had achievement scores that were higher than those in the larger classes. However, the associated effect sizes, an average of 0.21 (Mosteller, 1995, p. 121), were only a fraction of those found in Project Follow Through's 1.40 average (Adams & Engelmann, 1996, p. 52; Bereiter & Kurland, 1981–82) and below the level typically used to denote “educational importance,” 0.25 (Tallmadge, 1977).

Despite the poor STAR results, the findings were widely publicized. An article on the STAR project that appeared in *The Future of Children* ended with these confident remarks:

Because a controlled education experiment ... of this quality, magnitude, and duration is a rarity, it is important that both educators and policymakers have access to its statistical information and understand its implications. Thought should be given by both public and private organizations to making sure that this information is preserved and well documented and that access to it is encouraged. The ... statewide controlled experiment is a valuable device for assessing educational interventions and, thereby, improving school systems. (Mosteller, 1995, p. 127)

The book *Evidence Matters: Randomized Trials in Education Research* (Mosteller & Boruch, 2002) featured only one educational study, STAR. California was convinced by the unqualified endorsements of the STAR project and reduced class size at the state level. Around the same time Mosteller's article (quoted above) was published, however, California was looking at disappointing data on the performance of its students. The 1994 National Assessment of Educational Progress (NAEP) assessment showed that only 18% of California's students were rated proficient or advanced in reading. California's national ranking on NAEP was second from last in reading (NAEP, 1996). There were obviously flaws in the design of the STAR study, but these were not identified before the fact. The biggest lesson that came from STAR is that random assignment cannot compensate for poor internal validity.

The Striving Reader Initiative

Smaller class size received (and continues to receive) considerable press (Krueger & Whitmore, 2001; Mosteller, 1995). In contrast, a colossal five-year study in Chicago,

the Striving Reader Initiative, was a gold-standard showcase for a considerable number of variables endorsed by the literature. The study was initiated in 2007 when Arnie Duncan (who later became Secretary of Education) was CEO of Chicago Public Schools. This study has received virtually no press. The probable reason is that the initial reports show not only that it didn't do as well as expected, but that it produced no apparent positive results.

The Striving Reader Initiative was designed to remedy reading failure of students in Grades 6–8. The project had random assignment of 31 treatment schools and 32 control schools. The initiative involved a three-tier intervention model that featured a full-school, 90-minute, daily immersion in reading comprehension and a focus on subject areas. The experimental schools initiated longer school days, after-school activities, more adults in the classroom, smaller class size, small-group differentiated instruction, staff development, teacher collaboration, counseling, high interest reading material, and frequent assessments.

The year four report issued in 2011 is over 180 pages and contains many tables (Metis Associates, 2011). Although the experimental schools were judged to be well implemented, results of the first year assessment showed no significant differences for any grade or any of the subgroups. After two years there were no differences. After three years, “results indicate that there was no detectable overall impact of the program on Tier 2 and 3 students” (p. 120). In addition, “there was no overall treatment effect for all students in the ITT sample at the end of the fourth project year ...” (p. 156).

Over the four years, only one of the 24 subgroups obtained a statistically significant outcome in favor of the experimental group. That achievement occurred in the fourth year; however, the p value was 0.048 (which would not be considered significant if the significance criterion was 0.01 rather than 0.05). More than one outcome would have been significant by chance alone (5 per hundred measures at the 0.05 level). Also, the effect size was 0.174, far below what is required for generally recognized educationally important effects (0.25; Tallmadge, 1977). Finally, the normal curve equivalent (NCE) scores for both groups were far below the mean of sixth graders and differed by less than 2 points (Control: 36.513 and Treatment: 38.333). If students had performed at grade level, they would have had scores in the 60s, which means both controls and experimental students were years below the norm for sixth graders. These are obviously not the results Arnie Duncan and others anticipated.

The fact that the results are not publicized, however, suggests that the system is deeply prejudiced and unscientific in its orientation. The Chicago study represents a

milestone for educational studies with its large numbers of students and gold-standard experimental design. The results should therefore be unquestionably valid. Why wouldn't it follow that the field should embrace the results and conclude that either there is no way to remedy the underachievement of these striving students or that the way Chicago approached the problem has been documented to be the wrong way?

Those seem to be the only conclusions, and they seem to be highly relevant to the instruction for Striving Readers. Shouldn't teachers and administrators know that the Chicago approach was inert? Shouldn't the specific instructional material and practices used in this effort be recognized as being ineffective or at least questionable? In a broader sense, shouldn't those who proposed this apparently naïve intervention be recognized as lacking credibility?

The ultimate question may be: If no conclusions are to be drawn from a study that fails to show significant results, wouldn't it have been wiser to design an intervention that did not meet the gold standard? In that way, there would be a modest reason for not publicizing the results.

Bias

In summary, there seems to be an overriding bias in recognizing the worth of the four large studies. Follow Through, the RITE study, and the Striving Reader Initiative had results that were not consonant with current prejudices about instruction. The results were ignored or actively suppressed. The STAR project had outcomes consistent with the current prejudices about instruction and the results were embraced and continue to be embraced, even after the California results strongly contradicted the STAR conclusions (NAEP, 1996) and despite weak outcomes. For instance, a Center for Public Education report on class size and student achievement (CPE, n.d., para. 6) concluded, "The most influential contemporary evidence that smaller classes lead to improved achievement is Tennessee's Project STAR."

QUEST FOR PRISTINE INTERNAL VALIDITY OF EDUCATIONAL STUDIES

The STAR study had very obvious problems of internal validity that were not on a handy list of possibilities. Because the study used random assignment of classrooms in the same school, two teachers teaching the fourth grade discover that one has a smaller class. The other teacher understandably asks, "Why does that teacher have less work than I have?" One possible response is for the slighted teacher not to work as hard. Scores in that classroom go down compared to the other classroom, and a spurious

difference is created. The probability of investigators knowing everything to control is apparently slim, but even obvious details are not controlled well in many studies.

Medical research recognizes that the greatest cause of poor internal validity in effectiveness studies is that patients don't provide accurate statements about whether they took the medication as scheduled (Glintborg, Hillestrom, Olsen, Dalhoff, & Poulsen, 2007). Parallel problems occur in educational studies. Studies often have inadequate provisions for documenting whether teachers are providing the instruction scheduled. Teacher reports of having started on time, ending on time, and teaching the material according to particular procedures are potential causes of distortions.

Are results distorted that much by not using random assignments? With the Chicago Striving Reader study, there would be no difference in any results if matched-pairs rather than random assignment were used to create the groups.

Unlike most medical trials, studies that evaluate instructional programs do not articulately describe the treatments. Studies that evaluate an experimental drug reveal an extreme difference between instructional and medical research. The drug, its schedule, and related information about dos and don'ts can be summarized succinctly. In contrast, a description of the schedule and related information for an instructional program would require many pages and would describe contingencies that have no clear parallel in trials that evaluate drugs. The reason is that an instructional treatment is the product of three variables—program, teaching, and placement of students. (See Chapter 1, Tables 1.1 and 1.2.) Unless all are controlled, serious distortions may occur.

This is not to say that nothing is revealed from studies that have flaws. The better program should produce better results. So if the numbers are large enough, the more effective programs should produce solid evidence of causing superior student performance. The caution is that the results are based on a correlation, so they don't imply what someone needs to do to cause positive outcomes in a particular classroom.

Large numbers of studies seem to be the key for evaluating any approach, but DI is the *only* approach that has large numbers of older studies that document its effectiveness. Unfortunately the DI studies that have large numbers are judged by the WWC to be invalid. Medical studies are not stricken from the record if they are more than 20 years old; however, the WWC rejects studies older than 20 years. This provision affects only one approach—DI (Stockard, 2010).

Should DI "replicate" these studies to provide more current findings?

No.

The field should follow the lead of medical research and recognize reasonably well-designed studies whether or not the results are recent or consistent with current prejudices. Furthermore, the field should encourage research that challenges the various universal assertions that appear in the literature.

A purging based on analytic induction, the logical approach described in detail in Chapter 1, would disclose the technical nature of the causes of learning that will otherwise remain hidden in correlational half truths that dominate much of current educational literacy. The ultimate goal would be the institutionalization of practices that guarantee continued use of programs that stand up to mandated program-referenced testing (also described in detail in Chapter 1). The beneficiaries would not only be students, but also teachers and administrators who engineer scientific instruction.

A SOCIAL SCIENTIST’S VIEW OF EDUCATION RESEARCH

Jean Stockard

In the following discussion, many of the points made by Engelmann in the first part of this chapter are echoed from a different perspective, that of a social scientist reflecting on educational research from outside the discipline. In the first section, key normative elements of science common to the social, biological, and physical sciences are described. The second section details ways in which the Direct Instruction tradition of program development and research conforms to each of these elements and how the DI tradition’s adherence to these norms contrasts sharply with the vast majority of educational research. The third section describes two of the most egregious examples of ways in which the educational research community has violated basic rules of science and misrepresented research results.

NORMATIVE TRADITIONS OF SCIENCE

In introductory classes to the field, and in more in-depth courses required of majors, undergraduate students in the social sciences study the methodological procedures and norms that have guided the disciplines for generations. While the subject matter of the social sciences is, of course, different and, many would argue, more complex than the biological and physical sciences, the normative traditions are similar, if not identical (Kemeny, 1969; Nagel, 1961). The discussion below focuses on four key norms of science.

1. Science Is Theory Based

Science is guided by theoretical speculations about how the world works. These speculations form the basis of all scientific work, from a chemist theorizing about the way in which chemical compounds will interact, to economists speculating about how changes in monetary policy affect inflation rates, or sociologists hypothesizing about relationships between changes in the demographic composition of communities and rates of lethal violence. No matter what the topic of study, scientists are guided by theoretical understandings regarding their areas of interest, generally building on theoretical traditions in their own and/or other fields. This theoretical work is both inductive and deductive in nature.

Good scientists use both inductive and deductive approaches. They compare their theoretical speculations with the data, revise these speculations when empirical results indicate that they should, test developing hypotheses, and then, as needed, revise their theories. As a field of study matures, the speculations become conceptual models and explanatory theoretical systems. Parsimonious theoretical explanations, those which use the fewest underlying assumptions or variables, are seen as the most elegant and the ultimate goal (Cohen, 1989; Einstein, 1934; Kaplan, 1964; Kemeny, 1969).

2. Science Is Cumulative in Nature

We use the findings of past work to guide our future work. Using the Popperian notion of falsification, hypotheses that are proven false are discarded, and those that receive support (or technically have not yet been falsified) are retained (Cohen, 1989; Popper, 1962). We continue to test hypotheses, but expand our analyses to see if results hold under varying conditions.

A variety of terms have been used to describe this cumulative process. One is the notion of a cumulative research program, or a phased model of research, involving the gradual development of understandings and causal inferences, typically moving from rather small and highly focused controlled experimental designs to tests with more varied settings, subjects, and outcomes and employing a range of methods and techniques (Cohen, 1989; Huitt, Monetti, & Hummel, 2009). Another is the “grounded theory of generalized causal inference” described by Shadish, Cook, and Campbell (2002), which involves the systematic comparison of results from a variety of settings, samples, and research approaches. The third and perhaps most commonly cited in the current literature is the meta-analytic tradition, discussed in Chapter 2 of this volume, which uses quantitative techniques to summarize large sets of research results (Bornstein, Hedges, Higgins, & Rothstein, 2009; Wolf, 1986). Despite the differing terms, all of the approaches are based on the notion that scientific understandings are

cumulative in nature, building on multiple tests of hypotheses. As Cook and Campbell, authors of the most widely cited works on research design, put it,

We 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. (1979, p. 31, emphasis in original)

3. Good Science Is Flexible

The best social science theories and the most reliable accumulations of results are based on data derived through multiple methods, in a wide range of settings, using both inductive and deductive reasoning, employing qualitative and quantitative analyses, and adjusting the research design to the conditions.

Among scientists, belief in the experiment as the *only* means to settle disputes about causation is gone, though it is still the preferred method in many circumstances. Gone, too, is the belief that the power experimental methods often displayed in the laboratory would transfer easily to applications in field settings. (Shadish, Cook, & Campbell, 2002, p. 30, emphasis in original)

Note that this call for multiple methods and approaches that match the needs of the situation is in complete agreement with Engelmann’s conclusions regarding the utility of the so-called “gold standard” approach for educational research discussed in the first part of this chapter.

4. Good Science Is Honest and Open

In the social sciences we have strong traditions of blind peer review, with scholars who do not know the identity of the authors reviewing and critiquing all material before publication. We also have strong norms regarding the sharing of data so that others can check our results, perhaps with different analytic strategies. Replication is encouraged, for only through multiple tests can we be assured that results are valid. In fact, without such openness and honesty none of the other scientific norms actually matters.

DIRECT INSTRUCTION AND NORMATIVE SCIENCE

The Direct Instruction corpus of work is a classic example of each of the elements of normative science.

1. Strong Theoretical Base

First, DI has a strong theoretical base. As detailed in Chapter 4 and Appendix B, Engelmann and his colleagues have written extensively on the theory of learning that underlies the development of the DI instructional programs with works such as *Conceptual Learning* (Engelmann, 1969), *Theory of Instruction* (Engelmann & Carnine, 1991), and *Inferred Functions of Performance and Learning* (Engelmann & Steely, 2004).

While the theoretical writings are intellectually connected with the long-established and classical tradition of the logical empiricists (see Engelmann and Carnine, 2010), Engelmann and Carnine's Theory of Instruction (1991) stands by itself as a fully articulated analysis of how children learn and can be effectively taught. Each element of the theory was carefully tested as it was developed, illustrating the interplay of inductive and deductive analyses mentioned above.

2. Cumulative in Nature

Second, the work is cumulative in nature. In fact, the Direct Instruction literature can be seen as a classic example of the phased model of research with the gradual expansion of focus and methodologies. The development of curricular programs grew from the initial work with preschoolers through the elementary grades and to programs for those in the upper grades and adults. The academic subjects involved moved from reading and mathematics to science and social studies (see Appendices A and C).

Detailed guidelines for behavioral management and implementation of the programs have been developed, as discussed in Chapters 5 and 6 of this volume. The curriculum has been developed through lengthy and detailed experiments with different populations of students and in different settings, using the methods described in Chapter 1. Extensive work has validated the principles of learning and instruction that provide the theoretical base. Field-testing with large and varied populations of students and teachers has examined the extent of the programs' effectiveness across many different settings (Engelmann, 2007; Engelmann & Carnine, 1991; Huitt, Monetti, & Hummel, 2009). Perhaps most important, evidence of the efficacy of the curriculum has continued to expand over the years with, as described in Chapter 2, highly consistent results.

3. Varied Approaches

Third, research related to Direct Instruction has taken many different forms and approaches, embodying the flexibility that is the hallmark of a mature science. While there are dozens of randomized control trials testing the efficacy of the DI programs,

there are also many field-based studies employing large samples as well as so-called “single subject designs” looking at more unique and specialized populations. Scholars have examined the impact of Direct Instruction with many different groups, from gifted students to those with severe disabilities. They have employed quantitative and qualitative methods; and they have looked at not just impacts on achievement, but at impacts on perceptions of self-confidence and self-efficacy of students and teachers as well as ways in which the programs can best be implemented. The result is a strong cumulative body of work that embodies consistent documentation of the ways in which DI programs promote high achievement among students with a wide variety of demographic characteristics, in many different settings, and at all ability levels.

4. Honest and Open

Finally, the body of DI research conforms to the norm of honest and open research. The work has appeared in peer-reviewed journals around the world, having been checked and cross-checked by independent researchers. In the context of a meta-analysis project, Borman, Hewes, Overman, and Brown (2003) documented the extent of these differences and commented on the relatively large amount of work on DI conducted by researchers not affiliated with the programs’ development. Borman and colleagues (2003) identified 49 studies of DI compared to a median of only 4 studies for the other 28 models they examined.

THE EDUCATIONAL RESEARCH COMMUNITY AND DIRECT INSTRUCTION

Even though Direct Instruction research conforms closely to the norms of science and the cumulated body of work is strong and consistent, the educational research community has, for many years, ignored and actively resisted these findings. Two egregious examples are examined below. Many could have been chosen, but these were selected because they involve the expenditure of vast amounts of money with the full endorsement of the federal government. Their beginnings are separated by almost four decades but show disturbing parallels of what could be seen as purposeful manipulation of research results to hide the ways in which DI programs promote strong student learning.

Project Follow Through

Project Follow Through has been called the largest educational experiment that was ever conducted. Beginning in 1967, as an element of President Johnson’s War on Poverty, it continued until the summer of 1995, with a total price tag of about one

billion dollars (Grossen, 1996). The project was built on the assumption that low levels of educational achievement were a major factor in the perpetuation of poverty from one generation to another. More than 20 different educational models, including Direct Instruction, were implemented in 170 high poverty communities throughout the country as a way to test the relative efficacy of each approach. Carefully designed assessment procedures were developed by two independent evaluation agencies to determine which of the models would be most successful in developing students' achievement and their self-esteem. Assessments included comparisons to "control" schools with similar characteristics as well as comparisons to national norms.

The results of the evaluation were clear cut. Direct Instruction was the clear winner. It was the only approach that resulted in students having positive changes in all of the measured outcomes. Some of the programs even had negative results, with students having worse outcomes at the end of the intervention (see Adams, 1996; Becker & Engelmann, 1996; Bereiter & Kurland, 1981–82; Grossen, 1996; Watkins, 1997).

In a world in which the norms of science are valued and honored, one would expect that these findings, based on such a well-designed and extensive experiment, would lead the educational research community to embrace Direct Instruction. In fact, however, just the opposite occurred. Most of the other programs in the experiment were programs favored by the educational establishment, such as those discussed in the first part of this chapter. To a large extent, many of these programs are still popular today, albeit sometimes having slightly altered names, such as "whole language" to describe the "language experience" approach and "developmentally appropriate" instead of "Open Education" (Grossen, 1996, p. 9).

In a series of maneuvers that are almost impossible to believe but which are, in fact, well documented, the educational research community changed the research question. While originally interested in which program produced the most favorable results, the question morphed into asking, "What was the aggregate result of the programs?" In other words, the results of all the different models were simply lumped together (House, Glass, McLean, & Walker, 1978). Given that 20 of the 22 models had negative or null results, the new conclusion was that the programs had no effect. Educational researchers wrote extensive critiques of the Follow Through design, criticizing the use of quantitative analyses and comparisons and suggesting that an ethnographic and case-study descriptive approach would be more appropriate (see Grossen, 1996, p. 7). Even though the evidence—from the original rules of the game—was

overwhelming, the educational research community chose to ignore the evidence and alter the rules so that this would seem legitimate.

What Works Clearinghouse

Over three decades after Follow Through began, the U.S. Department of Education established the What Works Clearinghouse (WWC). The stated purpose of the WWC is to be “a central and trusted source of scientific evidence for what works in education ...to provide accurate information on education research” (WWC, 2013a). Like Project Follow Through, the WWC has been extraordinarily expensive, with costs of well over 75 million dollars in its first 10 years of existence (U.S. General Accountability Office, 2010, p. 25). Yet, like Follow Through, the WWC has failed to live up to its promise, presenting very misleading reports to the public, analyses that denigrate strong programs, such as Direct Instruction, and promote far weaker ones. Many of the problems stem from the criteria that the WWC uses to choose studies for review as well as outright errors and misinterpretations of the studies that are accepted.

Criteria for WWC Acceptance

While the conclusion of the Follow Through project incorporated criticism of quantitative research and experimental designs, the What Works Clearinghouse, somewhat ironically, made an about-face. It strongly promotes randomized control trials, seeing these as the “gold standard” discussed in the first half of this chapter and virtually dismisses almost all other research designs. Studies that are accepted for the WWC review must typically meet a long list of methodological criteria, such as pretest scores within given ranges for each group, extensive data on attrition of subjects, and publication within the last two decades (WWC, 2013b).

Not surprisingly, very few studies have met the WWC criteria. The U.S. General Accountability Office (U.S. GAO) reports that less than 10% of all studies examined pass their requirements (U.S. GAO, 2010, p. 13). The acceptance rate for studies of DI is even lower. Of the more than 200 studies of Direct Instruction that the WWC claimed to have examined by mid 2013, only 10 (5%) were found to meet their criteria either fully or in part.

As described more extensively in other writings (e.g., Stockard, 2010, 2012, 2013a; Stockard & Wood, 2012), the elements of the DI literature that reflect its maturity and should, given the norms regarding the cumulative nature of science, add weight to judgments of its merit instead appear to have worked against its acceptance by the WWC. For instance, the automatic exclusion of studies conducted more than 20 years ago effectively discards the large number of randomized control trials that were at

the basis of the early development of the programs. To date the WWC officials have refused to provide any scientific basis for this decision. While they have claimed that this policy ensures that results will be valid for today's students, they have provided no evidence to suggest that the way in which students learn has altered over time (Stockard, 2008). The preference for small, tightly restricted, randomized control trials effectively excludes most field-based studies of larger populations and those that use advanced statistical methods for controls. Such larger, field-based trials (like the RITE work analyzed by Carlson and Francis, 2003, and discussed earlier in this chapter) are especially important for ensuring external validity and have been much more common in the DI literature as the field has matured.

Having such strict criteria for studies to be reviewed would be important if this made a difference in the results. In other words, if the criteria affected results of studies such that those that met the criteria had more accurate estimates of a program's efficacy, one could argue that it would be important to maintain these "standards." Given the large number of studies of Direct Instruction, it is possible to test this hypothesis, and I did so with a sample of works that examined the use of *Reading Mastery* with students with learning difficulties. The extent to which the studies met the WWC's criteria varied, with some meeting most of them and some meeting very few. (None met all the criteria, and all met at least one criterion.) In an extensive statistical analysis, I found that none of the criteria had a significant relationship to estimates of the effect of DI. In other words, whether or not the criteria were met, or whether a few or many of the criteria were met, the estimates of the effect of *Reading Mastery* on students' achievement remained unchanged—positive and significant, similar to the effects reported in Chapter 2, and well beyond the levels usually seen as educationally important (Stockard, 2013a).

WWC Errors in Interpretations

Given that the WWC's reports are based on only a small fraction of the extant literature it is crucial that decisions regarding the inclusion or exclusion of studies, as well as the interpretation of those that are accepted, be as accurate as possible. In other words, because their reports are based on a very small fraction of the extant literature it is crucial that such reports be accurate. However, in other writings I have documented numerous errors (see Stockard, 2008, 2010, 2012, and 2013a; Stockard & Wood, 2013). These involve errors in decisions regarding the exclusion of studies, such as the decision to exclude RITE's large scale evaluation discussed by Engelmann in the first part of this chapter. The WWC claimed that the inclusion of teacher training and instruction in behavior management, both key elements of DI programs, introduced confounds to the study and made it invalid.

I have also found serious errors in decisions regarding studies that were included in their reviews, documenting serious problems in 4 of the 10 studies of DI programs that the WWC had accepted for review by 2013. For example two studies accepted for a review of *Corrective Reading* used only selected elements of the program and the authors explicitly stated that their analysis should not be used to evaluate its efficacy. Analyses of study results that were accepted for review have been far from error free, with summary reports that sometimes distort and misrepresent the findings of the authors. For instance, an article that compared two highly similar Direct Instruction reading programs (*Reading Mastery* and *Horizons*) found that both produced achievement gains that were significantly greater than national norms would expect and that the impact of the two programs was similar. The WWC chose to ignore the comparison to the national norms and instead focused on the lack of difference between the two programs, concluding that there was no evidence that *Reading Mastery* was effective. (See Stockard, 2008, 2013a, and Stockard & Wood, 2012, for more extensive discussion.) To date, the WWC has refused to admit to any of these documented errors and has concluded only that some DI programs have “potentially positive” effects, a conclusion in stark contrast to the cumulated weight of scholarly research, such as the material summarized in Chapter 2.

The errors in the WWC procedures have extended to misrepresenting results for weak programs and providing positive ratings when the research evidence indicates that a different conclusion would be appropriate. For instance, the WWC has given high rankings to the Reading Recovery program, a short-term tutoring intervention. One of the studies used to justify this rating compared the standard Reading Recovery program to a “modified” program that included explicit instruction in phonological skills. Students in both the unmodified Reading Recovery program and the modified program (including instruction in phonologically based elements) eventually caught up with the other children, but the students in the modified program were able to discontinue tutoring much earlier. The standard Reading Recovery program was found to be 37% less efficient than the program that included instruction in phonics. In addition, students in the modified program continued to have higher levels of achievement and higher rates of learning at the end of the school year. The authors provide an extensive discussion and additional analyses that demonstrate the fallacy involved in Reading Recovery’s assumptions about the ways in which word recognition skills develop. They clearly conclude that Reading Recovery is not an efficient method for teaching children to read and that phonological training is superior.

The WWC chose to ignore any results regarding the comparison group that received phonological training and had superior achievement, “because it was a

modified version of the standard program” (Stockard, 2008, pp. 13-14). In correspondence explaining their decision they noted that the results with the other comparison groups were mentioned in a technical appendix of the WWC report, implying that such information could be available for those who were interested. Of course, the chance of a parent or school official accessing a technical appendix to find information that contradicts the inaccurate information given in the major pages of the web site is extremely remote, and most users of the website would reach the erroneous conclusion that the Reading Recovery program was effective. (See Stockard, 2008, for a fuller discussion and copies of correspondence with the WWC.)

Violating the Norms of Science

Although several decades separate the beginnings of Project Follow Through and the What Works Clearinghouse, the parallels in the ways in which they have violated the norms of science are striking and, to a social scientist, chilling. It could be suggested that, in its original conceptualization and design, Project Follow Through conformed to three of the four key norms of science described earlier:

1. Follow Through was based on the theoretical assumption that higher educational achievement was a key element to combatting generations of poverty.
2. Follow Through incorporated the notion of cumulating evidence by testing programs in multiple sites and over a period of years.
3. Follow Through could be seen as incorporating flexibility through its use of community and parental involvement and many different educational approaches.

Unfortunately, Project Follow Through failed to meet the most important scientific norm of honesty and openness:

4. Follow Through changed the research question and altered the presentation of the results to disguise the clear superiority of Direct Instruction.

The WWC lauds itself as “a central and trusted source of scientific evidence” (WWC, 2013), however, it fails to conform to all four key norms of science.

1. To date, the WWC has provided no theoretical basis for their approach, although the methodological literature, such as that cited above, could provide a great deal of guidance for such endeavors.
2. The very restrictive approach the WWC has taken to the literature, with an arbitrary cut-off date of studies to be reviewed and stringent criteria regarding the design of acceptable studies, has made a full examination of the cumulative literature virtually impossible.

3. The WWC acceptance protocols fly in the face of the flexibility advocated by the classic methodological tradition.
4. Finally, and most disturbing, the evidence suggests that, like the ultimate purveyors of the Follow Through results, the WWC has violated the scientific norms of honesty and openness. Violations of this norm appear in the inclusion and exclusion decisions, as well as the summaries of studies described above.

Other problems noted in analyses of the WWC work include failures to inform users of the ways in which their findings conflict with the established literature (e.g., the ways in which their conclusions regarding Direct Instruction contradict the meta-analytic findings described in Chapter 2), having no external peer review process, and failing to be transparent in procedures and processes, even though, as a federally funded program, it is required to do so. Finally, when errors occur, the WWC has failed to post public retractions and corrections of erroneous reports, allowing misrepresentations to remain in the public eye.

CONCLUSION

In the sections above we have given several examples of the ways in which the educational community, and particularly educational researchers, have ignored empirical evidence. In contrast to scientific norms that call for flexibility in approaches and building a strong cumulative tradition of research, the current fad in educational research involves the so-called “gold standard” approach of randomized control trials conducted under severely restricted conditions. We assert that this approach is inappropriate for most work in education and produces results that are neither internally nor externally valid (Stockard, 2013a, b). Even more disturbing than the current preoccupation with one narrow approach to research is the concerted pattern of those in power of misconstruing research results. For less favored programs, such as Direct Instruction, strong and positive results are hidden or misrepresented. In contrast, for favored programs, such as “whole language” and “constructivist” approaches, much smaller positive results are publicized and promoted and negative and null results are hidden or misrepresented.

While our analysis has focused to a large extent on issues related to logic and scientific design, we should not lose sight of the ultimate purpose of educational research—helping students. What could have been averted if the evidence from project Follow Through had been given the political attention that had been promised? What could happen if the What Works Clearinghouse presented an honest and complete compilation of the research evidence?

Two longitudinal studies of the impact of Direct Instruction provide indications of the potential. Linda Meyer (1984) looked at the long-term impact of Project Follow Through, comparing the rates of high school graduation and college acceptance of students in a Direct Instruction Follow Through program at the start of their schooling career to the rates of students in the comparison group in a nearby school with very similar demographic characteristics. The Follow Through students were over one and a half times as likely to graduate from high school and twice as likely to apply to and be accepted at college. More recently, Stockard, Carnine, Rasplica, Paine, and Chaparro (2014) looked at the high school experiences of students with varying exposure to Direct Instruction in their elementary years. Those with more exposure were significantly more likely to be prepared for higher education: more than twice as likely to take advanced college preparatory mathematics classes; more than twice as likely to take advanced placement and/or college entrance exams, such as the SAT or ACT; and ranked significantly higher in their high school class.

Reams of evidence in the social sciences document the importance of educational attainment in promoting higher incomes and occupational status, lower crime rates, greater family stability, and better health. In fact, this finding was the impetus for starting the Follow Through program. The results of the Meyer (1984) and Stockard et al. (2014) studies, separated by almost three decades, show that the potential is great. If the results of Project Follow Through had changed the course of education, as originally envisioned, the adult fates of millions of young people, and the face of the nation, would be substantially different.

Given the dramatic nature of this potential impact, one must ask why the educational research community has so vociferously resisted the accumulated data and research results. The discussions above suggest that they have not just ignored the evidence but have, in direct violation of the norms of science, actively manipulated their presentation of findings to hide the efficacy of Direct Instruction programs from the public. One could suggest that these actions reflect, either intentionally or unintentionally, attempts to maintain the privilege and power of educational researchers as well as those with vested interests in ineffective curricular programs. By promulgating the fiction that no educational approach is especially effective, they conclude that much more work is needed to develop effective curricula. In other words, they create a justification for their continued employment and for the continued expenditure of millions of dollars in the educational research enterprise. They also justify the continued use of ineffective programs for students throughout the nation, enriching the coffers of authors and publishers of these ineffective programs.

The research community, individual researchers, publishing firms, and authors of these ineffective approaches continue to garner grants, prestige, and money. In the meantime, generations of students have been, and continue to be, denied access to highly effective curricula. The nation has lost the potential contribution of generations of talented young people. The gap between the powerful and the less powerful remains and, in fact, in recent years has widened. The reasoning behind this process may or may not be consciously self-serving. But, in the end, the motive doesn't matter because the result is the same. The educational research community, and the publishing empires to which it is connected, continue to enrich themselves. The vast majority of students they purport to serve continue to lose.

Both of us, as authors, have approached this issue from our own perspective: Engelmann as the author of the Direct Instruction programs and a participant in education for many years, Stockard from a career as a social scientist. Yet, we agree on the bottom line. The educational research community appears to have willfully ignored the research evidence. They have done this for many years and in direct contradiction to logic and to the norms of the scientific community. The big losers are the students and the society as a whole, as the strong evidence regarding the effectiveness of Direct Instruction is degraded and ignored and ineffective programs are promoted. The bias of educational researchers trumps the research evidence, and students and the society are the losers.

REFERENCES

- Adams, G. L. (1996). Project Follow-Through: In-depth and beyond. *Effective School Practices*, 15(1), 43–55.
- Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle, WA: Educational Achievement Systems.
- Becker, W. C., & Engelmann, S. (1996). Sponsor findings from Project Follow-Through, *Effective School Practices*, 15(1), 33–42
- Bereiter, C., & Kurland, M. (1981–82). A constructive look at Follow Through results, *Interchange*, 12, 1–22.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73(2), 125–230.
- Bornstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York: Wiley.
- Center for Public Education (CPE). (n.d.). *Class size and student achievement: Research review*. Alexandria, VA: Author. Retrieved from <http://www.centerforpubliceducation.org/Main-Menu/Organizing-a-school/Class-size-and-student-achievement-At-a-glance/Class-size-and-student-achievement-Research-review.html>
- Carlson, C. D., & Francis, D. J. (2003). Increasing the reading achievement of at-risk children through direct instruction: Evaluation of the Rodeo Institute for Teacher Excellence (RITE). *Journal of Education for*

Students Placed At Risk, 7(2), 141–166.

Cohen, B. P. (1989). *Developing sociological knowledge: Theory and method* (2nd ed.). Chicago: Nelson-Hall.

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

Glintborg, B., Hillestrom, P., Olsen, L., Dalhoff, K., & Poulsen, H. (2007). Are patients reliable when self-reporting medication use? Validation of structured drug interviews and home visits by analysis and prescription data in acutely hospitalized patients. *Journal of Clinical Pharmacology*, 47(11), 1440–1449.

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

Einstein, A. (1934). *Essays in Science: The authorized translation from the volume originally published as: "Mein Weltbild."* New York: Philosophical Library. (Republished in 2011 ebook format, New York: Philosophical Library/Open Road).

Engelmann, S. (1969). *Conceptual learning*. San Rafael, CA: Dimensions Publishing Company.

Engelmann, S. (2007). *Teaching needy kids in our backward system*. Eugene, OR: ADI Press.

Engelmann, S., & Carnine, D. (1991). *Theory of Instruction: Principles and Applications*. Eugene, OR: ADI Press. (Originally published 1982, New York: Irvington Publishing).

Engelmann, S., & Carnine, D. (2010). *Could John Stuart Mill have saved our schools?* Verona, WI: Full Court Press.

Engelmann, S., & Steely, D. (2004). *Inferred functions of performance and learning*. Mahwah, NJ: Lawrence Erlbaum.

House, E., Glass, G., McLean, L., & Walker, D. (1978). No simple answer: Critique of the FT evaluation. *Harvard Educational Review*, 48(2), 128–160.

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

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. Scranton, PA: Chandler.

Kemeny, J. G. (1969). *A philosopher looks at science*. Princeton, NJ: D. Van Nostrand.

Kennedy, M. M. (1978). *Findings from the Follow Through planned variation study*. U.S. Office of Education. Retrieved from <https://www.msu.edu/~mkennedy/publications/docs/Federal Programs/Follow Through/Kennedy 78 FT findings.pdf>

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Economic Journal*, 111, 1–28.

Metis Associates. (2011). *Chicago Public Schools Striving Readers Initiative: Year four evaluation report*. New York: Author. Retrieved from <http://www2.ed.gov/programs/strivingreaders/chicagoeval32011.pdf>

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

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 5(2), 113–127.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Press.

- Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace, & World.
- National Assessment of Educational Progress (NAEP). (1996). *NAEP 1994 reading report card for the nation and the states*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York: Basic Books.
- Shadish, W. R., Cook, T.D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cervia, T. R. (1977). *Education as experimentation: A planned variation model (Vol IV-A)*. Cambridge, MA: Abt Associates. Retrieved from: http://www.eric.ed.gov/ERIC-WebPortal/search/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=ED148490&ERICExtSearch_SearchType_0=no&accno=ED148490
- Stockard, J. (2008). *The What Works Clearinghouse beginning reading reports and rating of reading mastery: An evaluation and comment*, NIFDI Technical Report 2008-4. Eugene, OR: National Institute for Direct Instruction.
- Stockard, J. (2010). An analysis of the fidelity implementation policies of the What Works Clearinghouse. *Current Issues in Education*, 13(4).
- Stockard, J. (2012). *A Summary of concerns regarding the What Works Clearinghouse*. Eugene, OR: National Institute for Direct Instruction.
- Stockard, J. (2013a). *Examining the What Works Clearinghouse and its reviews of Direct Instruction programs*. NIFDI Technical Report 2013-1. Eugene, OR: National Institute for Direct Instruction.
- Stockard, J. (2013b). Merging the accountability and scientific research requirements of the No Child Left Behind Act: Using cohort control groups. *Quality and Quantity: International Journal of Methodology*, 47(2013), pp. 2225–2257.
- Stockard, J., Carnine, L., Rasplia, C., Paine, S., & Chaparro, E. (2014). *The long term impacts of Direct Instruction*. Eugene, OR: National Institute for Direct Instruction.
- Stockard, J., & Wood, T. (2012). *Reading Mastery and learning disabled students: A comment on the What Works Clearinghouse Review*. NIFDI Technical Report 2012-2. Eugene, OR: National Institute for Direct Instruction.
- Stockard, J. & Wood, T.W. (2013). *The What Works Clearinghouse review process: An analysis of errors in two recent reports*. NIFDI Technical Report 2013-4. Eugene, OR: National Institute for Direct Instruction.
- Tallmadge, G. K. (1977). *The Joint Dissemination Review Panel ideabook*. Washington, DC: National Institute of Education and U.S. Office of Education.
- U.S. Government Accountability Office. (2010). *Department of Education: Improved dissemination and timely product release would enhance the usefulness of the What Works Clearinghouse (GAO-10-644)*. Washington, DC: Author.
- Watkins, C. L. (1997). *Project Follow Through: A case study of contingencies influencing instructional practices of the educational establishment*. Cambridge, MA: Cambridge Center for Behavioral Studies.
- What Works Clearinghouse. (2007). *Beginning reading topic report*. Washington, DC: U.S. Department of Education. Retrieved from: <http://bit.ly/rHCevF>
- What Works Clearinghouse. (2013). About Us. Washington, DC: U.S. Department of education. Retrieved from <http://ies.ed.gov/ncee/wwc/aboutus.aspx>.

What Works Clearinghouse. (2013). *Procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education. Retrieved from: <http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19>

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Thousand Oaks, CA: Sage.

Word, E., Johnston, J., Pate Bain, H., Fulton, D. B., Boyd Zaharias, J., Lintz, M. N., et al. (1990). *Student-teacher achievement ratio (STAR): Tennessee's K-3 class-size study*. Nashville, TN: Tennessee State Department of Education.

Word, E., Johnston, J., Pate Bain, H., Fulton, D. B., Boyd Zaharias, J., Lintz, M. N., et al. (1994). *The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report 1985-1990*. Nashville: Tennessee State Department of Education. Retrieved from: <http://d64.e2services.net/class/STARsummary.pdf>