article 7=1.qxd 7/9/2004 10:11 AM Page 111

GARY L. ADAMS, Educational Achievement Systems; TIMOTHY A. SLOCUM, Utah State University
with GARY L. RAILSBACK, SCOTT A. GALLAGHER, SARAH A. MCCRIGHT, RANDY A. UCHYTIL, WILLIAM W. CONLON, and JAMES T. DAVIS, George Fox University, Newberg, OR

A Critical Review of Randall Ryder's Report of Direct Instruction Reading in Two Wisconsin School Districts

Abstract: A recent report by Dr. Randall Ryder evaluated the use of Direct Instruction (DI) reading programs in 2 school districts in Wisconsin. In the report, Ryder claimed that students in 1st, 2nd, and 3rd grade who received Direct Instruction scored significantly lower on several standardized tests of reading than students who received more traditional reading instruction. This article examines the validity of the Ryder report. Examination of the report revealed that (a) the quality of implementation of Direct Instruction is highly suspect; (b) the group labeled "Direct Instruction" apparently included numerous students who received an undefined mix of DI and non-DI reading instruction; (c) the selection and assignment of classrooms and students to groups resulted in DI groups that performed substantially below the non-DI groups before the study began; (d) there are numerous ambiguities and contradictions regarding the number of students in various groups in each year of the study; (e) statistical reporting failed to include basic information such as degrees of freedom, means, and standard deviation for some or all analyses; (f) ANCOVA was assumed to control for systematically biased assignment without consideration of the assumptions, limitations, and interpretive difficulties involved; and (g) Ryder fails to report results from subtests on which, in previous reports, the DI group outperformed the non-DI group by a statistically significant margin. As a result of these and other problems, no firm conclusions can be drawn from Ryder's report. We conclude that Ryder's report should be subjected to an independent peer review process, and the results of that process should be publicized as widely as the report has been.

Dr. Randall Ryder, a reading professor at the University of Wisconsin–Milwaukee, has gained notoriety for his report "Results of Direct Instruction Reading Program Evaluation Longitudinal Results: First Through Third Grade 2000-2003" (Ryder, Sekulski, & Silberg, 2003a).

Ryder described his research as follows:

> This report presents the results of a three-year study examining the effect of Direct Instruction on the students' reading achievement. Two studies were conducted. The first was a longitudinal study examined [*sic*] the effects of Direct Instruction as students proceed from first through third grade in the Milwaukee Public School District (MPS) and the Franklin Public School District (FPS). The second study examined the effects of Direct Instruction

Journal of Direct Instruction, Vol. 4, No. 2, pp. 111–127.
Address correspondence to Gary L. Adams at gadams@edresearch.com

article 7=1.qxd 7/9/2004 10:11 AM Page 112

> on first graders in the Franklin Public Schools over three successive years. (Ryder et al., 2003a, p. 18)

Ryder's conclusions were that Direct Instruction was less effective than traditional instruction for teaching reading. Ryder claimed:

> Results on standardized measures of reading achievement revealed:
>
> - Students in first, second, and third grade receiving Direct Instruction scored significantly lower on their overall reading achievement than students receiving more traditional forms of reading instruction. These results were consistent in urban and suburban schools.
>
> - Students in first, second, and third grade receiving Direct Instruction scored significant [*sic*] lower on measures of comprehension than students receiving more traditional forms of reading instruction.
>
> - First graders in an urban school district receiving Direct Instruction scored significantly lower on decoding and comprehension than students receiving more traditional forms of reading instruction and these results were consistent across three consecutive school years.
>
> - Overall, on measures of reading achievement students receiving more traditional forms of reading instruction in urban and suburban school districts display significantly greater gains than students receiving Direct Instruction. (Ryder et al., 2003a, p. 4)

Direct Instruction has been a "hot button" issue for years, and the information in the University of Wisconsin–Milwaukee press release announcing the report was quickly picked up by *The Columbus* (OH) *Dispatch, Education Week* (Manzo, 2004), and the National Education Association Web site (NEA, n.d.), which included a word-for-word copy of the findings described in the University of Wisconsin–Milwaukee press release.

Given the importance of the issues raised by the Ryder report, the fact that its conclusions appear to contradict a great deal of previous research on the effectiveness of Direct Instruction (see reviews by Adams & Engelmann, 1996; American Federation of Teachers 1998a, 1998b, 1999; Borman, Hewes, Overman, & Brown, 2002; Herman, 1999), the fact that the report was apparently released to the public without prior peer review, and the widespread publicity that its conclusions have garnered, a careful evaluation of the report is warranted.

Conceptualization and Implementation of Direct Instruction

In the literature review of Ryder's report, he mentions that many people confuse "direct instruction" and "Direct Instruction." Direct Instruction refers to published curricula developed by Engelmann and associates, whereas direct instruction involves a set of teaching techniques that can be used with any curricula. However, after noting this distinction Ryder proceeds to ignore the difference and confuse the two. For example, Ryder states that "Elements of the direct instruction model began with Brophy and Everett's (1974) work addressing teachers' behaviors and how they are related to student performance" (p. 12). Later, he says, "Following on the work of Brophy and Evertson. [*sic*] Rosenshine (1986) advanced an active teaching method that consisted of presentation in small steps…" (p. 12). It is an interesting history of direct instruction, but Direct Instruction started in the 1960s, a decade before Ryder's history of direct instruction.

article 7=1.qxd 7/9/2004 10:11 AM Page 113

One of the most basic requirements of acceptable research is that the subjects in each treatment group actually receive the prescribed treatment. Ryder describes his study as, "examining the effect of Direct Instruction on the students' reading achievement" (Ryder et al., 2003a, p. 18). Given this focus, it would be reasonable to expect that the students in the "Direct Instruction condition" would have received reading instruction exclusively from a Direct Instruction reading program such as *Reading Mastery* (Engelmann & Bruner, 1995) or *Horizons* (Engelmann, Engelmann, & Seitz-Davis, 1997). However, Ryder did not operationally define the Direct Instruction treatment and did not assure that the students in the Direct Instruction condition actually received Direct Instruction.

Describing the treatment groups in the Milwaukee Public Schools, Ryder states, "Of the three selected MPS schools, one school used DI-Reading Mastery exclusively for 1_hour [*sic*] blocks each day. The second school used a mixed-method approach where teachers determined the extent to which DI and other instructional methods were used. The third selected MPS school, used for the purpose of comparison, implemented the Houghton-Mifflin reading series" (p. 18). However, Ryder did not further explain what a mixed-DI classroom entailed. Since we do not know what kind of instruction this mixed-DI group received, it is difficult to know how to interpret their results. If we could simply ignore this group and focus on the DI-only group and non-DI group, the problem with the mixed-DI group would not be too serious. Unfortunately, Ryder never states explicitly how he handled the mixed-DI group's data. His data analysis includes only two groups (DI and non-DI), and it appears that Ryder combined the mixed-DI and DI-only groups and referred to the combined group as "DI." Thus, the group that supposedly represents the effects of Direct Instruction in MPS apparently includes numerous students who received an undefined mix of Direct Instruction and other reading instruction methods. The treatment described as Direct Instruction in FPS is equally unclear. Ryder states that the *Reading Mastery* students "were exposed to additional reading curricula (i.e., Guided Reading, Cunningham methods)" (p. 18). Based on this information, it appears that in neither school district did the "Direct Instruction group" receive a true Direct Instruction intervention.

To get clarification on this and other issues, I attended Dr. Ryder's American Educational Research Association (AERA) presentation of his study (Ryder, 2004). When asked to define "mixed-DI" and also to differentiate between the various groups that received some DI along with other methods, Dr. Ryder was unable to answer the question. He said that this was how the schools defined themselves, and he was not able to clarify the nature of instruction that these students actually received.

Since the quality of implementation of the treatment is a crucial contributor to outcomes, research and evaluation studies are expected to clearly document evidence that the treatments were carried out as planned. In this study, such evidence could have included information on training that Direct Instruction teachers received and/or direct observations of their implementation of the programs. Ryder's sole statement about teacher preparation and fidelity of implementation comes in a single sentence: "All DI teachers, in the selected DI and Mixed DI/Non-DI schools, had been previously trained by DI trainers in the use and implementation of the DI reading program" (p. 18). There is no further description of the nature or extent of this training and no data on the quality of implementation are reported.

Ryder created and administered annual Teacher Questionnaires, which included several questions that could give clues about the adequacy of training in Direct Instruction. Relevant items from the 1st-year questionnaire are shown in Table 1. Teachers'

article 7=1.qxd 7/9/2004 10:11 AM Page 114

responses to these items might have provided some level of information on quality of implementation. However, Ryder did not report the results of these items.

At the AERA presentation, Ryder was asked to elaborate on the quality of training and fidelity of implementation. He stated that the implication of poor training and implementation is the excuse that DI proponents always give if program results are unfavorable. When asked how much *Reading Mastery* training was given each year, Ryder said "two or three days." He did not respond to questions about why he did not report results of the questionnaire items that could throw light on this issue. In spite of Ryder's characterization of this critical issue as an excuse, the minimal training provided to teachers does raise serious questions about the quality with which Direct Instruction was implemented. If Ryder is concerned about dismissal of his findings based on this "excuse," a more productive approach would have been to take steps to assure high quality implementa-

Table 1

Items From Ryder's Questionnaire Relevant to the Quality of Training and Implementation of Direct Instruction (based on Ryder et al., 2003a, p. 87)

15. Describe your training in Direct Instruction:

    Number of days you were trained _____

    Name of trainer _____

16. How well trained are you in Direct Instruction?

    (1) no training (2) trained a little (3) moderately trained (4) well-trained (5) very well-trained

17. How confident are you in your ability to use Direct Instruction?

    (1) not confident (2) somewhat confident (3) confident (4) quite confident (5) very confident

18. Did you experience resistance from district personnel in the implementation of Direct Instruction?

    (1) not at all (2) very little (3) somewhat (4) quite a bit (5) extensively

19. How do you feel about the amount of training you have had?

    (1) insufficient (2) somewhat less than adequate (3) adequate (4) more than adequate (5) plenty

20. Have you been evaluated by a Direct Instruction Trainer or Evaluator? Yes_____ No_____

21. If so, how satisfactory was this experience?

    (1) unsatisfactory (2) not very helpful (3) helpful (4) quite helpful (5) very helpful

article 7=1.qxd 7/9/2004 10:11 AM Page 115

tion of the treatment and to document that quality. This approach would have eliminated the issue of low quality implementation as a plausible explanation of his findings. In addition, this minimal training appears to conflict with his research proposal to the Wisconsin Department of Public Instruction (which had been approved by the University of Wisconsin's Institutional Review Board) in which he stated that he would provide 9 days of professional development per year.

Additional information on the issue of the quality of implementation of Direct Instruction comes from a letter from Sara Tarver to the editor of *Education Week* in response to publication of a summary of Ryder's report (Tarver, 2004). Tarver writes:

> I offer a unique perspective on [the Ryder report] because in 1999, SRA/McGraw-Hill, the publisher of Direct Instruction, asked me to meet with Mr. Ryder and others to discuss the study's design and implementation prior to its initiation in 2000. After several meetings, it became clear to me that the study, as planned, was flawed in ways that resulted in an unfair bias against Direct Instruction.
>
> The three major flaws were:
>
> - The faulty conceptualization of Direct Instruction (which destroyed the integrity of the Direct Instruction that was being evaluated);
>
> - The selection of so-called Direct Instruction classrooms in which the reading lessons were to be more like whole language or literature-based instruction than like real Direct Instruction lessons; and
>
> - Grossly inadequate training of teachers in the purposes and use of Direct Instruction.
>
> After several meetings, I became convinced that Mr. Ryder's real intent was to provide so-called "evidence" that could be used to ridicule Direct Instruction. Otherwise, if he were a knowledgeable researcher, why would he propose a study so obviously flawed in both its design and implementation? (p. 38)

Thus, it appears that the importance of clearly defining the Direct Instruction treatment, providing extensive training in Direct Instruction, and documenting the quality with which the program is implemented was raised well in advance of the beginning of this project. This makes Ryder's lack of clarity on these issues and his characterization of the question of treatment fidelity as an excuse even more troubling.

Participants and Groups

The "participants" section of a research or evaluation report normally provides a clear explanation of how participants were selected, the number of participants in each group, and important characteristics of the participants such as location and demographics (American Psychological Association, 2001). Since attrition during the course of a study is considered to be an important potential source of bias, the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001) states that researchers should, "give the total number of subjects and number assigned to each experimental condition. If any did not complete the experiment, state how many and explain why they did not" (p. 19).

Providing clear information on the number of students who participated in the study should be simple and straightforward, but Ryder's description is unclear and apparently self-contradictory. Ryder reports the numbers of students in his Tables 2a through 2d; these tables are reproduced in our Table 2.

article 7=1.qxd 7/9/2004 10:11 AM Page 116

Table 2

Ryder Report Tables 2a–2d (Ryder et al., 2003a, p. 19)

Table 2a, N per Year, (Students that were tested per grade per year)						
Year	2000-01 (1st grade)		2001-02 (2nd grade)		2002-03 (3rd grade)	
Instruction	DI	Non-DI	DI	Non-DI	DI	Non-DI
MPS	97	19	136	85	130	78
FPS	17	91	77	87	0	39

Table 2b, Student Attrition, (Number of students who matched per year)					
Year	2000-01 Sample	Matched Students from 2000-01 to 2001-02	Number of Students Lost from 2000-01	Total Number of Matched Students from 2000-01 to 2002-03	Total Number of Students Lost from 2000-01 through 2002-03
MPS	116	70	46	52	64
FPS	108	98	10	34	74

Table 2c, Number of Students each Year per Instruction (2000-2003) (Students that maintained participant status in study)									
District	DI 00-03 Yrs 1- 3	NonDI 00-03 Yrs 1-3	DI Years 1 & 2	NonDI Years 1 & 2	DI Years 2 & 3	NonDI Years 2 & 3	DI Years 1 & 3	NonDI Years 1 & 3	*Totals*
MPS	42	7	0	0	0	0	0	0	***49***
FPS	3	13	0	3	5	0	5	3	***32***
Totals	***45***	***20***	***0***	***3***	***5***	***0***	***5***	***3***	***81***

Table ?, Total N per Year First Grade Groups						
Year	2000-01		2001-02		2002-03	
Instruction	DI	Non-DI	DI	Non-DI	DI	Non-DI
FPS	17	89	74	0	22	95

Note. The title, Table ?, is copied directly from Ryder's report. The correct title would be Table 2d.

article 7=1.qxd 7/9/2004 10:11 AM Page 117

Ryder described this study as *longitudinal*, implying that a group of students would be followed across a period of time. However, his Table 2a reports that *more* students were tested in 2001–02 (second grade) than were tested in 2000–01 (first grade). For example, the table states that 97 DI students were tested in first grade in the Milwaukee Public School District (MPS), and 130 DI students were tested in second grade in that district. It is not clear how a longitudinal study (i.e., one that tracks a group of students across time) can have more students in the 2nd year than in the 1st year.

Ryder's Table 2a also shows zero DI students in the Franklin Public School District (FPS) in 2002–03 (third grade). If this is accurate, there should, of course, be no results for third-grade DI students in FPS and no complete longitudinal results (first through third grade) in that district. However, as we will discuss in our section on data analysis, Ryder reports outcomes for third-grade DI students in FPS. In addition, his Table 2c reports that three FPS DI students were continuous participants from Year 1 through Year 3. In his Table 2b Ryder reports that the "total number of matched students from 2001-01 to 2002-03" (p. 19) (this combines DI and non-DI) in MPS is 52 and in FPS is 34. However, in Table 2c he states that MPS had 49 continuous participants across Years 1 through 3 (42 DI and 7 non-DI) and FPS had 16 (3 DI and 13 non-DI). This appears to be a contradiction.

In his Table 2a Ryder reports 91 FPS non-DI students in the first grade during the 2000–01 school year; however, in Table 2d he reports 89 FPS non-DI students in first grade during that year. These errors in tabulating and reporting the number of students may not seem earth-shaking in and of themselves; however, these errors in the simplest aspect of data analysis raise questions regarding the accuracy of aspects of data analysis that are more complex and we cannot check so readily.

Ryder does not provide reasons for the unusual aspects of these tables. For example, he does not note the fact that zero students are listed in Table 2a for the third grade FPS DI group, and he does not offer an explanation. Similarly, he does not explain why in Table 2d there are zero non-DI students in 2001–02, nor does he explain why there were 74 DI students in 2001–02 but only 39 in the other 2 years combined.

A fundamental issue regarding participants and grouping is whether the groups were comparable before the initiation of the intervention. Random assignment of participants to treatments is the gold standard for creating groups; however, in many applied evaluation settings, random assignment is not practical. Nonrandom assignment of intact groups (e.g., classrooms or schools) to treatments is a lesser alternative that can provide reasonable control under some circumstances. When nonrandomly constituted intact groups are used, great care must be taken to assure that the processes of selection and assignment of these groups do not introduce biases, and these processes must be clearly described so that readers can evaluate the adequacy of the methods. Further, the research must provide specific evidence that before the interventions were implemented, the groups were comparable in terms of achievement and other characteristics that may impact their future rate of learning.

Ryder used intact groups in his longitudinal evaluation; however, he does not provide a description of *how* these groups were selected. His most detailed description of the selection process is, "the schools participating in the three-year longitudinal study were selected from urban and suburban school districts" (p. 18). The report does not state how particular schools were selected for inclusion. We do not know, for example, whether the non-DI schools were among the highest achieving, lowest achieving, or typically achieving schools in their districts. In addition to this lack of

article 7=1.qxd 7/9/2004 10:11 AM Page 118

information on the process of assignment, Ryder does not report achievement at these schools in the years before the treatments were initiated. This information could have provided evidence regarding the overall comparability of the schools.

Ryder does provide demographic information on the participating classrooms. He states, "Demographic information of the classrooms from the first and second year of the study is illustrated in Appendices 1a and 1b" (p. 18). These appendices are reproduced in our Tables 3 and 4. His Appendix 1a (see Table 3) supports the claim that the classrooms *within* each school district are broadly comparable in terms of ethnicity and percent of students eligible for free or reduced-cost lunch. Appendix 1b (see Table 4), however, does not provide any demographic information. It merely reports the grade levels, numbers of students in each classroom, and number of students

Table 3

Ryder Report Appendix 1a (Ryder et al., 2003a, p. 74)

Appendix 1a.
First Year 2000-01

Demographic Data for Classrooms									
	Class-room	Students Enrolled	Students with Valid Tests	Ethnicity					% of Students Eligible for Free or Reduced Lunch
				African American	Asian	White	Other	Not Identi-fied	
MPS 1 (DI)	1111 1112 1113 1114 1115	19 19 20 17 20	15 15 13 16 10	15 15 13 15 4			 1 	 6	100 100 92 94 100
MPS 2 (DI)	1211 1212 1213	23 23 16	14 3 11	14 3 11					100 100 100
MPS 3	1321 1322	18 18	8 11	8 10			 1		100 100
Franklin 1 (partial DI)	2111 2112 2113	20 19 20	18 18 19	2 1 1	 3	16 15 15	 	 2 	11 11 5
Franklin 2	2221 2222 2223	16 20 20	15 20 18	3 4	1	11 16 18	 	 	13 10 5
Totals		308	224	119	4	91	2	8	

article 7=1.qxd 7/9/2004 10:11 AM Page 119

with valid tests, omitting any information on demographic variables. In addition, the numbers of students with valid tests given in Appendix 1b do not match the numbers listed in Tables 2a and 2d for either district. For example, Appendix 1b shows a total of 70 MPS second-grade students with valid tests (out of 230 students enrolled in those classrooms) in the 2nd year of the study, but Table 2a shows a total of 221 "students that were tested" as second graders in MPS. If Ryder's phrase "students that were tested" meant "students with valid test scores," then there is a very large discrepancy in the numbers (70

Table 4

Ryder Report Appendix 1b (Ryder et al., 2003a, p. 75)

Appendix 1b.
Second Year 2001-02

Demographic Data for Classrooms				
School	**Classroom**	**Grade**	**Students Enrolled**	**Students with Valid Tests**
MPS 1 (DI)	1116 1117 1119 1118 1110	2nd 2nd 2nd 2nd 2nd	16 13 16 16 21	8 6 8 9 11
MPS 2 (Mixed DI/ Non-DI)	1213 1214 1215	2nd 2nd 2nd	26 20 12	5 7 5
MPS 3 (Non-DI)	1325 1323 1324 1327 1326	2nd 2nd 2nd 2nd 2nd	17 19 19 19 17	1 3 5 0 2
FPS 1 (Non-DI)	2225 2226 2224	2nd 2nd 2nd	16 16 16	16 15 16
FPS 2 (Mixed DI/ Non-DI)	2116 2115 2114	2nd 2nd 2nd	19 20 19	15 18 18
FPS 3 (DI)	2411 2412	1st 1st	18 19	16 16
FPS 4 (DI)	2311 2312	1st 1st	21 22	21 22

article 7=1.qxd 7/9/2004 10:11 AM Page 120

versus 221); if the phrase did not imply that these students had valid test scores that were included in the analysis, then the column heading in Table 2a is extremely misleading.

Appendix 1a (and Table 2e) shows the vast demographic differences between the two districts: The MPS schools serve large numbers of African-American students from low-income families, and the FPS schools serve a mostly white population with relatively few low-income families. If students in the DI and non-DI treatments were drawn from the two districts in approximately equal numbers, then this contrast between the districts would provide for a valuable opportunity to examine the effects of DI in districts with differing demographics. However, participants in the two groups were not drawn from the districts in approximately equal numbers. For example, Table 2a indicates that in the 1st year (which, in a longitudinal study provides the basis for all subsequent years' sample), approximately 85% of the DI group was drawn from the urban district, but only 17% of the non-DI group came from that district. This, along with the demographic data provided in Appendix 1a (and Table 2e), is direct evidence that the DI and non-DI groups were *not* demographically comparable; in fact, they were *extremely* different in terms of demographics.

The FPS first-grade study was conducted exclusively in FPS. This eliminates the problem of disproportionality across the two districts. However, in the second study the students who participated in the DI and non-DI groups were not selected for comparability; instead, *the DI students were assigned to the DI group based on the fact that they were struggling with reading*. Ryder describes this selection process: "The first grade teachers within these classrooms utilized DI and Non-DI methods of instruction for reading, with lower-ability readers receiving both DI and Non-DI methods and higher-ability readers just receiving Non-DI methods" (p. 18). It is not clear whether the higher ability readers constituted the non-DI group or whether this group was drawn from other classrooms, but it is very clear that the participants in the DI group were selected for low performance, and the participants in the non-DI group were not selected in a similar manner.

Thus, the selection and assignment of students to groups appears to be strongly biased in both studies. In the longitudinal study, DI groups appear to be made up of a disproportional number of students from the urban district which, based on demographics, would be expected to produce lower test scores. And in the FPS first-grade study, students were assigned to DI based on the fact that they were identified by teachers as showing lower reading ability, with no apparent attempt to identify a comparable non-DI group.

Appendices 1a and 1b raise further questions about the samples of students whose test scores were analyzed. Appendix 1a shows that across all the classrooms, only 73% of the students enrolled had valid tests; in MPS only 63% of the enrolled students had valid tests, and in one classroom only 13% (3 of 23) of the enrolled students had valid tests. These high proportions of missing tests constitute a level of selection of participants. We do not know whether this selection creates a bias in the samples because Ryder does not mention the issue of missing tests and does not discuss the reasons for the large differences between numbers of students enrolled and numbers of students with valid tests. Appendix 1b shows even higher proportions of missing tests. One classroom of 19 students yielded no valid tests, and over one third (5 of 13) of the MPS classrooms had less than 20% valid tests. One contributor to this low yield of valid tests in the 2nd year may be the fact that Ryder was tracking only those students who had valid tests from the 1st year. However, Ryder does not suggest this explanation, much less explain how much of the discrepancy is a result of attrition, nor does this explanation correspond very well with the numbers in Table 2a.

article 7=1.qxd 7/9/2004 10:11 AM Page 121

Statistical Analysis

The first and most basic question regarding data analysis is what data were analyzed. The answer to this question is not at all obvious. As we have mentioned above, Ryder does not state whether the mixed-DI groups were combined with DI groups in the longitudinal analyses, although this seems to be the case. Further, Ryder's tables do not provide a clear indication of how many students were actually included in the analyses. Some of these uncertainties could be clarified somewhat by examining the *n* or degrees of freedom associated with the descriptive and inferential statistics that are reported in the analysis; however, Ryder gives neither *n* nor degrees of freedom for *any* of his descriptive or inferential statistics.

A preliminary step in most analyses of longitudinal data is to examine the pretest scores. Ryder, however, does not report any descriptive statistics (e.g., means or standard deviations) on pretest scores for the longitudinal study. As a result, we cannot know how the problems with selection are reflected in participants' reading skills immediately before the study began.

Ryder does report pretest means for the FPS first-grade study. As an example, the data from Ryder's Graph 5, which reports reading composite scores from the Gates-MacGinitie test, are reproduced in Table 5. The pretest means confirm that the DI group entered the study with lower reading skills than the non-DI group. In the 1st year (2000–01) the discrepancy between the DI and non-DI groups was approximately 48 points, and in the 3rd year (2002–03) the difference was approximately 63 points. We cannot compute an effect size to put these differences in better perspective because Ryder does not report standard deviations for these results (or any other results in the report). However, we can put the magnitude of the differences in context by noting that the non-DI group gained about 70 points across the 2000–01 school year and 47 points during the 2002–03 school year. Thus, the initial difference between the DI and non-DI groups appears to be roughly equivalent to a year's growth.

Table 5 also shows two unexplained anomalies in the pretest scores for the DI group. In the 1st and 3rd years, the pretest scores are remarkably similar (327.5 and 327.48). These means may even be exactly the same and differ only in how they were rounded. This may suggest an error in analysis and/or reporting of the results. The second issue is that the 2nd year pretest mean (373) is sub-

Table 5

Pretest, Posttest, and Gain Scores for First Graders in FPS (based on Ryder et al., 2003a, p. 30, Graph 5)

	2000-01			2001-02			2002-03		
	Pre	Post	Gain	Pre	Post	Gain	Pre	Post	Gain
DI	327.5[1]	408.41	80.91	373.34	411.82	38.48	327.48	384[1]	56.52
Non-DI	375.18	444.79	69.61				390.07	436.87	46.8
Difference	47.68	36.38					62.59	52.87	

[1] Ryder does not explain whether these values are rounded or whether trailing zeros have been dropped.

article 7=1.qxd 7/9/2004 10:11 AM Page 122

stantially different from those from the 1st and 3rd years. In fact, this pretest score is quite similar to the non-DI groups in the 1st and 3rd years. Recall that the FPS first-grade groups, unlike the non-DI groups, were made up exclusively of lower performing students. It is difficult to reconcile Ryder's statements about the selection of this group with the pattern of pretest results. This suggests that the sampling and assignment procedures were different in Year 2 than in the other years; however, Ryder gives no explanation of this in the report. As a result, we know that the DI group was different in the 2nd year, but we know very little about the sampling and assignment in this year, and this is the year that provides 65% of the DI data for this FPS first-grade study.

For the longitudinal study, Ryder reports Gates-MacGinitie total scores and comprehension scores. As an example of his results, Table 6 reproduces the data from Ryder's Graph 4, in which he reports results for the comprehension section of the Gates-MacGinitie test. One point to note from this table is that the FPS DI group is shown to have a score of 25.1 in the third grade; however, in his Table 2a Ryder reports zero students in this group. No explanation of this apparent anomaly is given in the report.

As we have noted above, in both of the studies the DI and non-DI groups were substantially different in terms of expected achievement levels. In the longitudinal study this is a result of the DI group being drawn disproportionately from the urban school district. In the FPS first-grade study, this difference is a result of systematic selection of lower performing students and placement of these students in the DI group. Thus, the initial differences between groups are not a result of the fact that classrooms that appear to be generally similar may have subtle differences. Instead, they are the result of obvious and systematic processes that will clearly lead to noncomparable groups. Ryder uses the analysis of covariance (ANCOVA) to attempt to adjust for these differences. This use of ANCOVA is Ryder's only method for dealing with the vast differences between groups.

The use of ANCOVA to "equate" intact groups that differ as a result of nonrandom processes has been discussed extensively by statisticians and is covered in statistics textbooks. For example, Stevens (1999) writes:

> It should be noted that some researchers (Anderson, 1963; Lord, 1969) have argued strongly against using analysis of covariance with intact groups. Although we do not take this position, it is important that the reader

Table 6

Ryder's Adjusted Posttest Scores (Estimated Marginal Means) From the Longitudinal Study (based on Ryder et al., 2003a, p. 26, Graph 4)

	First Grade	Second Grade	Third Grade
MPS Non-DI	3.1818	-6.5455	3.1818
MPS DI	-24.7778	-30.1944	-29.5556
FPS Non-DI	35.6667	14.1667	24.875
FPS DI	44.3	24.4	25.1

article 7=1.qxd 7/9/2004 10:11 AM Page 123

> be aware of several limitations and/or possible dangers when using ANCOVA with intact groups. First, even the use of several covariates will *not* equate intact groups, and one should never be deluded into thinking it can. The groups may still differ on some unknown important variable(s)….Third, the assumptions of linearity and homogeneity of regression slopes need to be satisfied for ANCOVA to be appropriate….A fourth issue that can confound the interpretation of results is differential growth of subjects in intact or self-selected groups on some dependent variable….A fifth problem is that of measurement error….In non-randomized studies measurement error can seriously bias the treatment effect. (pp. 322–323)

Virtually all textbooks echo this need for caution in the application and interpretation of ANCOVA. Howell (2002) states, "Anyone using covariance analysis, however, must think carefully about her data and the practical validity of the conclusions she draws" (p. 637). Several authors make the point that the potential problems with ANCOVA increase when the groups are substantially different at the outset. Pedhazur and Schmelkin (1991) comment, "The farther apart the two groups are on the initial measure, the potentially more severe the interpretive difficulties" (pp. 291–292). Glass and Hopkins (1996) state, "When groups differ widely on some confounding variable X, these differences imply that the interpretation of an adjusted analysis is speculative rather than definitive" (p. 606). All of these authors (and those of virtually all statistical textbooks) make the point that to use ANCOVA correctly, the analyst must check whether the data meet the assumptions of the procedure and carefully consider whether it is actually performing appropriately.

Ryder provides a basic description of the fact that ANCOVA adjusts group means based on scores on the covariate (pretest); however, he does not mention any of the assumptions, limitations, or potential interpretive problems associated with the procedure. He does not give any indication that he has tested any of the assumptions or even considered these issues. Considering the fact that virtually all of his data analysis and all of his conclusions regarding the effects of DI on student achievement depend on the application of ANCOVA to groups that systematically differ by large margins on pretests, it is shocking that he is so cavalier in his use of this procedure.

We cannot test the assumptions of ANCOVA from the results available from the report; however, we can gain some insight into whether the groups have been properly adjusted for some of the differences that are known to exist. The main source of differences between groups in the longitudinal study is the disproportionate sampling from the two school districts. If the ANCOVAs successfully adjusted group means in such a way that the differences between students in the two districts were eliminated (leaving only the treatment as a potential cause of differences between groups), then we would expect that results from the two districts would be similar. However, Ryder's results indicate statistically significant differences between districts even after use of ANCOVA to adjust for pretest differences. This difference can be seen in his Graph 4 (our Table 6) which provides group means that have been adjusted by ANCOVA. The two FPS groups score well above the two MPS groups even after adjustment. This demonstrates that ANCOVA has *not* eliminated the differences between districts and therefore *has not compensated for the unequal assignment of students to groups*. The same pattern is evident in results for Gates-MacGinitie total scores.

Table 6 clearly illustrates the problem with unequal sampling from the two districts. Since both FPS groups scored higher than both of the MPS groups (even after ANCOVA adjustment), a treatment group with a higher proportion of FPS students will tend to have a

article 7=1.qxd 7/9/2004 10:11 AM Page 124

higher mean than a treatment group with a lower proportion of FPS students. Thus, by mixing samples with different proportions of students from the two districts, we could obtain any desired outcome including one directly opposite of that which Ryder obtained. In this light, the fact that the DI group was initially constituted with 85% of its students from MPS, while the non-DI group drew only 17% of its students from MPS, looms large.

Unfortunately, we cannot make a similar examination of the results from the second study because it includes only one district. Ryder does not discuss how the hugely different numbers of participants in the two groups (see his Table 2d), combined with systematic differences in achievement level, might affect ANCOVA results. Further, Ryder does not discuss how the ANCOVA might be impacted by the fact that 65% of the DI students participated in 2001–02 and that no non-DI students were tested in that year.

Information From Other Sources

In the Method section, Ryder indicates that the Word Decoding and/or Word Knowledge subtests of the Gates-MacGinitie were administered to students at all grade levels; however, he does not report results for these subtests in this report, and he gives no explanation for their absence. He did report results for these subtests in his Year 2 report (interestingly, Ryder reports degrees of freedom, means, and standard deviations in this earlier report):

> A statistically significant effect of Direct Instruction on the change in decoding subtest scores from 2001 to 2002 was found for students in FPS ($\underline{F}(1,94)=4.92$, $\underline{p}=.0290$), suggesting that Direct Instruction ($\underline{M}=57.07$, $\underline{SD}=25.30$) leads to a greater understanding of letter-sound correspondences than other reading instructional methods ($\underline{M}=35.21$, $\underline{SD}=36.71$). It should be noted, however, that 15 out of the 82 students not receiving Direct Instruction demonstrated a decrement in their 2002 decoding scores when compared to their 2001 decoding scores. Removal of these 15 cases from the data results in the effect of Direct Instruction on change in decoding scores as no longer statistically significant ($\underline{F}(1,79)=1.45$, $\underline{p}=.2322$). There was no statistically significant effect of Direct Instruction on the decoding scores from 2001 to 2002 for students in MPS ($\underline{F}(1,67)=0.13$, $\underline{p}=.7204$). (Ryder, Sekulski, & Silberg, 2003b, p. 27)

There was no mention of test administration problems in the Year 1 report; those tests were not removed at that time. Nor does the decrement necessarily reflect incorrect results. Since these scores are relative to test norms, a decrement merely means that the students did not keep pace with the rate of growth seen in the test's norm group. Thus, it is not clear that there was a legitimate rationale for dropping these cases. It was only after Ryder conducted his statistical analysis for Year 2 and found that the DI Group had statistically significantly higher scores, that he decided to remove the data. Subsequently, Ryder did not even report results from this subtest.

In order to get another look at reading performance at these schools, we accessed the Wisconsin Department of Instruction Web site. This site reports performance of third graders on the Wisconsin Reading Comprehension Test for all public schools in Wisconsin. Results from the three MPS schools that participated in Ryder's evaluation are shown in Table 7 (Wisconsin Department of Instruction, n.d.). FPS scores are not shown because no FPS schools maintained a DI program through third grade. These data suggest that the DI school had far more students in the *advanced* category and far fewer in the *mini-*

article 7=1.qxd 7/9/2004 10:11 AM Page 125

mal and *basic* categories than did the mixed-DI and non-DI schools. These results must be considered with some caution because, as with Ryder's data, questions can be raised regarding the percentage of students not tested. Further, these data are based on different samples than Ryder's and a different test of reading. Nonetheless, these results strongly suggest that Ryder's results are not the only results that can be derived from comparisons of these schools. Given the problems and uncertainties with Ryder's methods and analysis, this alternative dataset is particularly instructive.

Conclusions

The University of Wisconsin–Milwaukee's (2004) press-release title was unambiguous: "Study: Direct Instruction Not Best Way to Teach Reading." This headline was followed by the first sentence, "A three-year study of methods of teaching reading shows that highly scripted, teacher-directed methods of teaching reading were not as effective as traditional methods that allowed a more flexible approach" (para. 1). Our analysis of Dr. Randall Ryder's report suggests that these conclusions are unwarranted and, as a result of severe problems with methods, data analysis, and reporting, no firm conclusions can be drawn from this report. To summarize the main problems

1. The quality of implementation of the DI treatment is highly suspect. The amount and nature of training in DI is not reported, and from what can be inferred, it appears to have been minimal. At least one consultant quit the project because of minimal training that was being offered. No data are reported on fidelity of implementation of the treatment, and though survey questions that could throw light on this question were administered, results from these questions are not reported. Thus, we do not know the degree to which DI methods were actually used in the DI classrooms.

2. Results from a mixed-DI treatment, a treatment that is ill-defined but is clearly not a pure DI model, appear to have been combined with results from the DI group, and the number of subjects exposed to DI versus mixed-DI is not clear. Thus, in the data analysis the results attributed to DI appear to actually represent a combination of DI and mixed-DI treatments, and there is no way to know how much of the effects are due to each of these treatments. In addition, all conclusions regarding DI appear to conflate the two treatments.

3. The assignment of participants to groups is clearly and explicitly biased against the DI group. In the longitudinal study, the DI

Table 7

Wisconsin Reading Comprehension Test Results From the Three MPS Schools Involved in Ryder's Evaluation

			Proficiency Level			
Group	School	Not Tested	Minimal	Basic	Proficient	Advanced
DI	Clarke St.	13.3%	3.3%	16.7%	48.3%	18.3%
Mixed-DI	Franklin	2.7%	16.2%	29.7%	48.6%	2.7%
Non-DI	Hopkins St.	2.4%	28.0%	26.8%	39.0%	3.7%

article 7=1.qxd 7/9/2004 10:11 AM Page 126

group was drawn disproportionately from an urban school district, and the non-DI group was drawn disproportionately from a suburban district. In the study of first-grade achievement in FPS, students were assigned to the DI group only if their teachers identified them as lower performing, and the remaining students appear to have been assigned to the non-DI group. Therefore, biased selection appears to be a very plausible threat to the validity of conclusions.

4. There are numerous ambiguities and contradictions in reporting the number of participants in each treatment and across the 3 years of the study. Thus, it is not clear how many participants contributed to each analysis. Further, although Ryder repeatedly describes one of the studies as longitudinal, it is not clear whether the individual participants in the study were tracked across the 3 years. In addition, contradictions regarding the number of participants raise questions about the care and precision with which other aspects of the data were handled.

5. In many of the analyses, pretest and/or posttest means were not reported. Basic descriptive statistics such as *n*, standard deviations, and effect sizes were not reported for *any* of the analyses. Thus, the reader cannot independently check many of the analyses, and there is no report of the magnitude of differences between groups on either pretest or posttests.

6. ANCOVA was invoked to statistically control for the systematically biased selection and assignment of participants; however, there was no discussion of the assumptions, limitations, and potential interpretive difficulties with this technique. None of the assumptions of ANCOVA were tested for plausibility. Therefore, we do not know whether ANCOVA was a valid analytic technique for these data. Inferences from reported results suggest that ANCOVA did not adequately control for the large initial differences between groups.

7. Ryder gives no explanation for failing to report results of a subtest of the Gates-MacGinitie on which the DI group outperformed the non-DI group by a statistically significant margin.

One might wonder how a report with so many serious flaws could be published and taken seriously by the educational community. This research was not reported in a peer-reviewed journal; it was merely released at a press conference and with no apparent prior peer review. The release of the report circumvented a critical aspect of the scientific process—scrutiny of other researchers in the field. The process of peer review is designed to safeguard against poor quality research and faulty conclusions being represented as legitimate scientific products. In bypassing the peer review process, Ryder and others responsible for the report have done the educational community a disservice by releasing potentially misleading information and unwarranted conclusions in the guise of scientific research.

Unfortunately, the release of the report cannot be undone and its impact cannot be reversed. However, unless some sort of systematic peer review takes place, the conclusions of the report will continue to be believed by many who have not examined their basis. For this reason, I recommend that officials from the Wisconsin legislature, officials from the University of Wisconsin–Milwaukee, educational researchers, and others concerned that educational practices be built upon a firm foundation of high-quality research, call for an independent review of Ryder's report by a team of qualified educational researchers. Such a review could be conducted under the auspices of the AERA. We suggest that the individuals and institutions responsible for the Ryder report and the distribution of its claims should assure that the results of this peer

article 7=1.qxd 7/9/2004 10:11 AM Page 127

review receive publicity equal to that which has been generated for the report.

References

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

American Federation of Teachers. (1998a). Seven promising reading and English language arts programs. Retrieved May 25, 2004, from www.aft.org/edissues/downloads/seven.pdf

American Federation of Teachers. (1998b). Six promising schoolwide reform programs. Retrieved May 25, 2004, from www.aft.org/edissues/downloads/six.pdf

American Federation of Teachers. (1999). Five promising remedial reading intervention programs. Retrieved May 25, 2004, from http://www.aft.org/edissues/downloads/remedial.pdf

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. A. (2002). *Comprehensive school reform and student achievement: A meta-analysis.* CRESPAR Technical Report No. 59. Baltimore, MD: Center for Research on the Education of Students Placed at Risk, Johns Hopkins University. Retrieved May 25, 2004, from http://www.csos.jhu.edu/crespar/techReports/Report59.pdf

Brophy, J., & Evertson, C. (1974). *Process-product correlations in the Texas teacher effectiveness study: Final report* (Research Report 74-4). Austin, TX: University of Texas, R & D Center for Teacher Education. (Eric Document Reproduction Service No. ED099345)

Engelmann, S., & Bruner, E. C. (1995). *Reading Mastery*. Columbus, OH: SRA/McGraw-Hill.

Engelmann, S., Engelmann, O., & Seitz-Davis, K. L. (1997). *Horizons learning to read*. Columbus, OH: SRA/McGraw-Hill.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn and Bacon.

Herman, R. (1999). *An educators' guide to schoolwide reform.* Washington, DC: American Institutes for Research. Retrieved May 25, 2004, from http://www.aasa.org/issues_and_insights/district_organization/Reform/index.htm

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbuy (Thomson Learning).

Manzo, K. K. (2004, January 28). Study challenges direct reading method [Electronic version]. *Education Week, 23*(20), 3.

National Education Association. (n.d.). Study says Direct Instruction not best way to teach reading. Retrieved May 25, 2004, from http://www.nea.org/reading/directinstruction.html

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenshine, B. (1986). Synthesis of research on explicit teaching. *Educational Leadership, 43*(7), 60–69.

Ryder, R. J. (2004, March). *A longitudinal examination of the effects of Direct Instruction on beginning readers*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Ryder, R. J., Sekulski, J., & Silberg, A. (2003a). Results of Direct Instruction reading program evaluation longitudinal results: First through third grade 2000-2003. Milwaukee, WI: School of Education. Retrieved May 25, 2004, from http://www.uwm.edu/News/PR/04.01/DI_Final_Report_2003.doc

Ryder, R. J., Sekulski, J., & Silberg, A. (2003b). Results of Direct Instruction reading program evaluation first through second grade 2000-2002. University of Wisconsin–Milwaukee. Retrieved May 25, 2004, from http://www.soe.uwm.edu/pages/welcome/Departments/Education_Outreach/Programs___Courses/reading_symposium/readings

Stevens, J. P. (1999). *Intermediate statistics: A modern approach* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Tarver, S. G. (2004, February 25). Direct Instruction: Criticism of a Wisconsin study [Letter to the editor] [Electronic version]. *Education Week, 23*(24), 38.

University of Wisconsin–Milwaukee. (2004, January 14). Study: Direct Instruction not best way to teach reading. Retrieved May 25, 2004, from http://www.uwm.edu/News/PR/04.01/Reading.html

Wisconsin Department of Instruction. (n.d.). School performance report. Wisconsin reading comprehension test: An assessment of primary-level reading at grade three. Test results. Retrieved May 25, 2004, from http://www.dpi.state.wi.us/dpi/oea/spr_wrct.html

Editor's Note:

We recognize the stylistic inconsistencies that exist within this article. In the preparation of the article we followed two principles: (a) to adhere to the stylistic guidelines set forth in the American Psychological Association's (APA; 2001) *Publication Manual of the American Psychological Association 5th Edition*, and (b) to print all quoted material verbatim even if the quoted material does not conform to APA style. Thus, there are some inconsistencies in some aspects of style.