

Critique of: *Results of Direct Instruction Reading Program Evaluation Longitudinal Results: First Through Third Grade 2000-2003*, by Ryder, Sekulski, and Silberg.

A study conducted by Randall J. Ryder and two colleagues from the University of Wisconsin at Milwaukee (October, 2003) draws the conclusion that Direct Instruction (DI) produces lower gains in student performance than “traditional methods” for teaching reading. The report is over 100 pages, most of which are single-spaced, with numerous references, extensive discussions of ANCOVAs (analysis of covariance), the history of DI, a tedious discussion of characteristics of good teachers, many tables, many instruments for measuring teacher responses, and much information about the classrooms that participated in the program (including the designated room number of each classroom).

Superficially, the study appears to be the quintessence of academic precision and thoroughness. If one takes the time necessary to plow through the discussions and the methodology, however, it becomes apparent that the study is so seriously flawed that it could be summarized in a few paragraphs.

The basic design of the study was to compare DI with traditional methods in two districts—Milwaukee Public Schools (MPS) and Franklin Public Schools (FPS). The study covered three years, 2000-01 through 2002-03, and three grades, 1, 2 and 3.

The most serious problem is that the population identified as DI students has only a small percentage of students who were taught

using strict DI procedures. During the first year of the study, three MPS schools and two FPS schools participated.

Of the three MPS schools, one used DI; one used what the study calls “a mixed-method approach where teachers determined the extent to which DI and other instructional methods were used. The third . . . used . . . the Houghton-Mifflin reading series” (Ryder et al., p. 18). *Both the DI school and the “mixed method” school are categorized in the data as DI schools.*

FPS schools are in a “suburban” district with a school population that is 91% white. According to the study, students in this district “score at a more proficient reading level,” and DI was used as a “compensatory model specifically for students that scored low” in reading. *The problem with FPS is that none of the teachers used strict DI.* According to the report, all students receiving DI in FPS “were exposed to additional reading curricula (i.e., Guided Reading, Cunningham Methods)” (p. 18). At least one of the teachers also used *Reading Recovery*.

As noted above, the study identifies three categories—DI, mixed methods, and Non-DI. All the data, however, refer to only two categories: DI and Non-DI, with all of the “mixed methods” put in the DI sample. (How DI students were identified in FPS is not clear from the study since none of the schools used undiluted DI.)

The report does not show performance by school, so it is difficult to draw conclusions about the relative gains of the children in the DI school compared to the mixed method.

Given the problems of treatment assignment, the rest of the report is moot. It purports to be a comparison of DI with Non-DI, but

the DI group has been attenuated so severely that it is difficult to draw conclusions about the effects of DI, rather than the effects of what Ryder has decided is to be labeled DI.

The presentation and analysis of data is probably more obscure than the logic for assigning students to the treatment groups. Here is a table that summarizes the treatment groups in first grade for three years.

Gates-MacGinitie Reading Achievement Scores

	Pretest	Post-test
Non-DI 2000-2001	375.18	444.79 N = 91 FPS N = 19 MPS
DI 2000-2001	327.5	408.41 N = 17 FPS N = 97 MPS
Non-DI 2001-2002		N = 0
DI 2001-2002	373.34	411.82 N = 74 FPS
Non-DI 2002-2003	390.07	436.87 N = 95 FPS
DI 2002-2003	327.48	384 N = 22 FPS

The 2000-01 pretest scores for DI and Non-DI are greatly different (327 to 375). The reason has to do with the composition of the groups. For the DI group, 97 of the subjects came from MPS (urban), but only 17 came from FPS (suburban). In contrast, the Non-DI sample was composed of 91 students from FPS and only 19 from MPS. The loading of the DI group with MPS students means that the anticipated rate of progress of the DI group would be less than that of

the Non-DI group, which had 91 of its 110 subjects from FPS. Interestingly, however, the DI group gained 81 points, and the Non-DI group gained only 70 points.

In year 2001-02, there are no Non-DI students. The reason is that the first-grade scores for MPS were recorded for only one year—00-01. For the two remaining years, all of the students were from FPS. There were 74 students in the 01-02 sample and all came from FPS. For some unexplained reason all of the students in the FPS classrooms were assigned to DI, even though the DI pretest scores for 01-02 were suspiciously high (373 versus 327 for 00-01). They are as high as the Non-DI scores for 00-01, when only 20% of the population came from MPS. In contrast, the DI pretest scores for both 00-01 and 02-03 were 327, which is 46 points lower than it was in 01-02. This huge discrepancy is not explained by Ryder.

*The apparent stability of the DI pretest scores over the other two years implies that there is some kind of mix of higher and lower performers in 01-02.*

The post-test scores for 01-02 are strikingly low compared to what happened in the previous year, when the DI students' gain was 50 points greater than it was in 01-02. In 00-01, DI students had post-test scores that were only 4 points lower than the 01-02 group even though the 00-01 students began 50 points behind.

A suspicious relationship exists between the pretest scores for the three years. *The pretest scores of the DI 01-02 group were nearly as high as the Non-DI scores for 00-01 (373 to 375); however, FPS was supposed to use "DI" only for students who "scored low."* This inconsistency is further complicated by statements from the report.

Ryder contradicts the idea that DI was used exclusively as a “compensatory model” for lower readers. He indicates that there was an FPS “administrative decision to limit the use of Direction Instruction [*sic*] to students identified as lower-ability readers *after the 2001-02 school year*” (p. 18). This statement would imply that DI was not strictly used for lower-ability readers before that year; however, if this is true, there are extensive contradictions in the data.

In 02-03, there are 95 Non-DI students and 22 DI students, all from FPS. The DI pretest score is once more where it was in 00-01, at 327. This time, however, the gain is only to 384, a much smaller gain than there was when the sample included MPS. In 02-03, the Non-DI pretest score was 390, which is 63 points higher than the DI pretest score for that year. However, the Non-DI group increased only 47 points, while the DI group increased 57 points. These scores are consistent with the possibility that FPS teachers are not as adept at achieving student gains as the MPS teachers are. The study draws the conclusion that DI is apparently not as well suited to the suburban population as it is to an urban population. The 02-03 scores suggest that if DI is not well suited to suburban populations, Non-DI is equally ill suited, even for students who have pretest scores much higher than the DI group.

Returning to the 01-02 conundrum, we can generate an approximation of what kind of population mix would generate the strange pretest score. We assume that the pretest scores for “low performers” is around 327 and that FPS Non-DI classrooms have a pretest score of around 390 (as it was in 02-03). If we combine those scores at the ratio of 80% Non-DI to 20% DI, we get a score of 377,

which is pretty close to 373. If this were the case, there actually was a Non-DI group in 01-02 and it comprised around 80% of the 74 students. This is one of several possibilities. In any case, the 01-02 DI population was clearly not composed entirely of low performers. Furthermore, the treatment did not occur after the school year of 2001-02. By implication, the group would not necessarily be composed exclusively of low performers, in which case there are serious inaccuracies in both the analysis and the discussions.

In summary, DI students are compared with students who have much higher entry scores. There is no comparison group for one of the years and no explanation of how the pretest scores could be about as high as the Non-DI population from the preceding year, when about 80% of the Non-DI population came from FPS and is assumed to be high performing. At the same time, the pretest score for Non-DI in 02-03 had no MPS students and was 63 points higher than the DI sample from 02-03.

In addition to the study of first-graders, Ryder conducted a longitudinal study that followed students from grade 1 through grade 3. Its problems were largely those of group designation and missing data points. The DI and Non-DI groups were not comparable; there are no data by school. The number and percent of students assigned to DI varied wildly from one year to the next. For instance, in grade 3, all 130 students in the DI sample were from MPS and none were from FPS. In contrast, 39 of the 117 Non-DI students were from FPS.

### Number of Students for Longitudinal Analysis

	00-01 Grade 1		01-02 Grade 2		02-03 Grade 3	
	DI	Non-DI	DI	Non-DI	DI	Non-DI
MPS	97	19	136	85	130	78
FPS	17	91	77	87	0	39

Ryder’s analyses use two dependent variables, the condition (DI versus Non-DI) and the year. The independent variable is the difference or gain in performance. The most straightforward observation for the first-grade data is that there is an obvious interaction by district, with FPS producing very low rates of progress for all and MPS producing larger gains. It seems that there is also a very large effect created by the one MPS classroom that used uncompromised DI. So why wouldn’t Ryder use the classrooms by year and treatments as dependent variables?

The ultimate product of this study is the analysis that leads to “Estimated Marginal Means Total Difference Scores” for each grade. Ryder concludes that there was a statistically significant difference between DI and non-DI on the Gates-MacGinitie norms, a statistically significant difference within districts, and one on comprehension.

The label of “longitudinal” for this study is gratuitous. There was little continuity from level to level, whereas there was significant discontinuity. No DI students from FPS went through the three-year sequence. No more than 17 went through grades 1 and 2. No more than 19 Non-DI students from MPS completed the sequence, but 97 MPS DI students could have completed the sequence (by virtue of the design,

not actual attrition). There are a total of 94 data points (Ns) for the FPS DI students compared with 214 for FPS Non-DI. In contrast there are 363 data points for MPS DI, but only 182 for Non-DI.

By treating the categories simply as DI and Non-DI, Ryder is able to hide the fact that there are great gaps in the data and significant loading in the composition of the DI group and the Non-DI group at each grade level. If DI in grade 3 has no students from FPS, and if only 17 students from FPS could complete the grade 1-2 sequence, a proper correction of the data would have thrown out the FPS data or used it in a far more conservative manner. In the same way, the opportunities for only 19 MPS students to complete the Non-DI sequence renders the advertised Ns of the study misleading. Did the students who suddenly appear at a particular grade level go through the same sequence as students who are being evaluated as part of that treatment and that grade level?

Ryder's design obscures this question by treating DI as if it is a homogeneous treatment that is not affected either by how it is implemented or the extent to which students actually went through earlier levels of the program.

Ryder declares that the scores had been "adjusted" to account for the difference in pretest scores at each grade. The problem is that the kinds of adjustments that are made in this manner tend not to take into account the trends for the content that is tested. For instance, comprehension scores in the third grade are affected differentially by the performance level of the children. Low performers have a serious performance drop when reading material requires "rich language," which occurs around the third grade. In the same way,

students who are below norm are expected to progress at a rate that is consistent with the rate they have achieved historically.

The only safe way to perform a fair evaluation, therefore, is *to study effects on comparable students*. The latest trend in experimental design is demanding random assignment of schools of comparable demography for measuring school-wide effects. Although this demand is probably specious, it is many times more specious to try to draw conclusions about effectiveness by comparing students who are not even close in pretest performance.

In any rigorous study, the FPS data should be thrown out. The main reason is that there is no fidelity in following DI tenets. The MPS sample should not include the Mixed treatment because there are no rigorous means to determine what that treatment is and therefore the extent to which the Mixed treatment provided in one classroom is the same as the Mixed treatment in another. One “mixed” classroom is credited with being 100% whole language, and one is one-half DI, one-half whole language.

The two generally accepted scientific standards for a comparative study are:

- The report presents number of subjects, means, and standard deviations for all groups (experimental and comparison).
- The groups being compared are closely matched with respect to entry scores, demography (male/female, SES, etc.) and time of treatment.

The Ryder study fails on both counts. It does not provide standard deviations. It does not have matched subjects. Also, the numbers are strange. For example, the subjects were apparently

screened according to whether they had “valid test scores.” The percentage of students who had valid scores was over 90% for FPS DI schools, but was 73% for the pure DI school in MPS and was 47% for the “Mixed” school. Clearly, the populations are not matched within MPS DI. So even if the study had proper group designations, it would fail to meet basic scientific standards on several counts.

A large part of the Ryder document focuses on qualitative evaluations of teacher opinions. Possibly these responses will be convincing to somebody, but what they tend to show is that many of the teachers involved in the study are naïve. Their responses clearly reveal that they are extremely misinformed about teaching reading effectively.

I spent the time necessary to sort through this report and write about it because you can bet that those who are into “traditional teaching methods” will pick up on this study and make much of it. That’s unfortunate because it is obviously a bad study that would not stand up to any kind of careful scientific scrutiny. In several places, including the executive summary, Ryder points out that there was not really a lot of difference between the DI and Non-DI classrooms, because they all tended to do the same thing. If that’s true, why is he apparently so determined to promote his version of “the same thing”?

Wisconsin would be well advised to present the study to a statistician and have the analyst point out the crippling methodological flaws in the study design and analysis. Wisconsin should also demand a reanalysis of the data by classroom and by strict designation of treatment.

Siegfried Engelmann, University of Oregon