Reading Research Plagued by Poor Designs and Misleading Conclusions

Once again *Education Week* has published a summary of a study that, with one small exception, suggests "no effects" of programs designed to improve student engagement and depth of reading comprehension *(Ed Week*, May 6, 2010). This follows a large string of "no effect" studies coming from the federal Institute of Education Science (IES). (See *Ed Week*, April 1, 2009.) For those programs involved in these studies the suggested conclusions are frustrating and misleading. In fact, when a "randomized" study finds *no effect*, it pointedly does not mean the treatment was ineffective. A test of the null hypothesis in statistics only allows one to make probabilistic conclusions when there is a *positive* effect. Unfortunately, the data from these null result experiments are often reported with the implication that programs are ineffective, when it is more likely that the design and measurements were poorly controlled and failed to provide a test of the reading program effectiveness. There are many ways to obtain no results in large scale quasi scientific experiments.

To illustrate let's say we want to study the effectiveness of Prozac on symptoms of depression. We randomly assign half the subjects to a treatment condition in which they take Prozac daily and half to a placebo condition. If we find no significant differences between the experimental group (Prozac) and the control group (placebo) can we conclude that Prozac is not effective in treating depression? What if we also know that only 30% of the subjects in the treatment condition actually took Prozac daily, 30% took it when they remembered, and the rest didn't bother? What if we also find out that several subjects in the control condition were actually taking Prozac instead of the placebo? The answer is that such a study would never be considered for publication nor could it be to determine the effectiveness of Prozac. This example of "research" is almost a direct parallel to the latest large scale "studies of reading program effectiveness."

As developers of Project CRISS, one of the experimental treatments in the study, we agreed to participate because we wanted to support the idea of independent empirical verification of program effectiveness. However, we believed the design and implementation would provide a fair test of the program, and results would be fairly reported.

Only after signing contracts and committing to the research did we discover that key components of the research design doomed the study to failure and rendered the data useless in terms of evaluating the effectiveness of Project CRISS. We raised these concerns repeatedly with the evaluation staff, but were given no voice in the final experimental design or implementation.

The design flaws were many and to comment on all of them would be too technical and tedious, but we would like to highlight several major problems that render these data meaningless. First, we did not realize that a significant portion of the experimental schools would involve immigrant Hispanic students who lacked fluency in English. For example, Carol Santa, Ph.D., was the teacher trainer in one of the experimental sites in which 97% of the participants were Spanish speaking (including most of the teachers) with a high turnover of students whose families had recently entered this country. Dealing with English language learners is not necessarily a problem for Project CRISS, but it is unfair to assess students with English language multiple choice assessments when they receive the majority of their instruction in Spanish.

Second, we had no input into the types of assessment tools and, for simplicity, the evaluators selected several standardized multiple choice reading assessments. This form of assessment is much less effective in assessing deep understanding, organization, and comprehension than tests that involve real world tasks that require oral or written tests of recall. When recall assessments

are used to assess the effectiveness of Project CRISS, evaluators found consistently strong positive effects of the program. The data are reliable and replicable over a span of twenty years in more than 21 different comparison groups.

Of even more concern, the design was supposedly randomized, but in order to maintain equal sample sizes, the Project Managers of Mathematica Policy Research insisted that those schools and teachers assigned to the "experimental condition" must remain in the study . . . even if the school and/or the teachers were not willing to participate. In the research site Dr. Carol Santa facilitated, the principal did not attend the full CRISS workshop and one of the two experimental teachers refused to participate and was absent for every follow-up visit. When observing his classroom, Dr. Santa saw no sign of CRISS implementation. Even though we let the Project Managers know about the situation, they refused to drop him from the study. Unfortunately, this was not an isolated event—superficial participation or non-participation was repeated in many experimental sites. Districts received government money for the study even in situations of minimal participation. Because Project CRISS provides professional development—not software or a workbook—it cannot be effective if teachers and their administrators will not participate and actually use the instructional methods that provide the basis of CRISS.

Finally, *Education Week* must stop reporting poorly-designed, null result studies with inflammatory headlines such as "Supplementary Reading Programs Found Ineffective." These headlines should instead read "Millions of Federal Dollars Wasted on Poorly Designed Study of Reading."

Carol Minnick Santa, Ph.D.
Founder and Co-Owner Project CRISS
Past President of the International Reading Association

John L. Santa, Ph.D.