

May 12, 2009

Dear Colleagues,

Many of you have probably heard about the disappointing results from the reading comprehension study funded by the federal Institute of Education Sciences (IES) through a contract awarded to Mathematica Policy Research (Supplementary Reading Programs Found Ineffective, *Education Week*, May 13, 2009). Project CRISS was one of four reading comprehension programs selected to participate. The results from the first year of implementation produced no positive effects for any of the programs studied. The results of this study add to a growing body of research sponsored by IES which have produced no effects and may give rise to an examination as to whether the methodology used in these large scale evaluations is suited to the real world of classrooms. So many factors are impossible to control—teacher buy-in, administrative support, degree of project implementation, length of treatment, student populations, sensitivity of the assessment measure. (See *Education Week*, "No Effects" Studies Raising Eyebrows, April 1, 2009, attached.)

After CRISS was selected from a pool of programs, we realized features of the research design were problematic for us. First, the subjects were fifth graders (ten year olds). Most fifth graders can use aspects of CRISS, but understanding Project CRISS at a deep level requires a certain level of maturity. Fifth graders need more time to internalize the self-directed aspects of CRISS with more modeling and teacher-directed exposure. The schools in the study were from high poverty areas with a predominance of second language students having difficulty reading in English. In the research site where I worked, teachers taught students CRISS strategies using Spanish texts in science and social studies. While they were able to implement strategies in their mother tongue, many couldn't read English well enough to adapt what they learned to English texts. To make matters worse, the final assessments, as well as the test-taking instructions, were in English making it impossible for students to apply the comprehension strategies they had learned.

Second, there was a major problem with buy-in. In a randomized experimental design, teachers do not choose to participate but are randomly assigned to a treatment. In this study, teachers who refused to use CRISS (or the other treatments) were still counted and their students' scores were averaged into the final data results. Because the effects for each treatment were combined across all sites, it is impossible to examine effects by individual sites. Our trainers working in these different sites reported tremendous variability in buy-in and implementation. For example, I worked in one school where one of the experimental teachers refused to cooperate at all. He remained totally disengaged through the initial CRISS workshop and was always conveniently absent on my subsequent visits to the school. According to the principal and other teachers, he refused to implement any part of CRISS. Even though I let project managers at Mathematica Policy Research know about the situation, they refused to drop him from the study. CRISS is based on the idea that teachers actually want to improve the quality of their instruction by helping their students learn how to learn. In order for the project to take hold, it needs broad support from both administrators and staff. The project provides professional development—not software, a workbook, or a basal program that teachers

"pace" their students through—which makes it even more vulnerable to teacher buy-in issues.

Third, the assessments didn't measure what CRISS does. We voiced our concerns to the research group and suggested using a variety of assessment tools (e.g., longer reading selections, short and long writing responses, student reflections, process assessment) relevant to a deep level of understanding. They refused to allow us any input into their choice of assessments and haven't even let us see the instruments even at the completion of the study. We do know, however, the instruments were in English and they were standardized multiple choice tests. Teachers at one site told us one section of the test provided students with a set of sentences where all spaces between the words had been removed. The students' "comprehension" task was to draw lines within the sentence where the spaces should be. In the research site where I worked, most of the Hispanic students would not have been able to even read the selections on the tests. Over the years as we have refined the assessment tools used to evaluate the effectiveness of CRISS, we found that multiple choice tests are not a sensitive measure; strong effects occur, however, when students read a longer passage, study the selection, and then take a free recall test after a 24 hour delay. Because CRISS-taught students learn how to reorganize and transform the information from text, they use deeper level processing than required to answer questions on a multiple choice test. Not having any input about the type of assessments turned out to be extremely problematic for CRISS.

Fourth, we are extremely upset by the way *Education Week* reported the results from the study. The headline was "Supplementary Reading Programs Found Ineffective." When a randomized study finds no effect, it pointedly does not mean the programs were ineffective and one is not allowed to make conclusions in a null hypothesis test. A null hypothesis design inherent in a general linear model of statistics only allows one to make probabilistic conclusions when there is a positive effect. The absence of an effect can be explained in many ways and conclusions cannot be drawn from such data. The entire design of the experiment had a high likelihood of failure to find a positive effect as have all the large-scale, randomized controlled studies supported by IES over the last several years. Far too many sources of uncontrolled variation occurred to make this a valid experiment. In particular, the failure of their tests to assess the skills taught by CRISS (and, perhaps, the other treatments as well); the failure to control the overwhelming effect of teacher variability; and the failure to control the quality of program implementation have doomed any meaningful interpretation of the study. Once again, it is the assessment process that has failed; not our program. It is unfortunately quite common, particularly in education, that poorly understood or even blatantly bad science is used to make inflammatory and incorrect conclusions.

Carol Santa, Ph.D.  
Developer, Project CRISS