

Socrates on Chicago's Failure

Scene: Hotel lobby

Characters: Socrates and Dr. Sidney Williams, influential education policy-and-management entrepreneur.

Socrates: As I understand it, the Chicago Public School conducted a five-year study designed to improve performance of struggling readers.

Dr. Williams: True. But we refer to these students as “striving readers.”

Socrates: As I understand it, you felt confident that the study would have strong, positive outcomes. Is that correct?

Dr. Williams: Yes, we all felt that way. The research was a landmark in gold-standard design. It addressed everything we know that would benefit striving readers in grades 6 through 8.

Socrates: What are these variables that cause a great difference?

Dr. Williams: Smaller class size, more adults in the classroom, teachers networking and collaborating, parent involvement, increased emphasis on in-service training, a corresponding increase in the training personnel—all bringing a strong focus on evidence-based practices.

Socrates: Are these really variables that cause learning or simply correlations between the performance of better readers and features of that instruction?

Dr. Williams: Probably both. The research was designed to provide the kind of overkill that would guarantee success. Therefore, in all grades, teachers use high-quality, high-interest

material. In all content areas teachers teach comprehension strategies. Teachers explicitly teach more vocabulary. Lead literacy teachers are explicitly trained to use assessment and diagnostic data to adjust instructional practices. There are also after-school programs for students who need more help.

Socrates: I understand the research is conducted in quite a few schools.

Dr. Williams: Yes, over 31 experimental schools and 30 control schools. Furthermore, these schools were randomly assigned the experimental or control treatment.

Socrates: I understand that the cost of this study is the highest the Feds have ever paid for a research project of this type.

Dr. Williams: True.

Socrates: Has the city evaluated this model on a smaller scale before implementing it on such a grand scale?

Dr. Williams: Absolutely. The city's chief instructional officer, Barbara Eason-Watkins, pioneered this effort when she was principal of an inner-city school. She used many of these techniques and her students showed the impressive results that could be achieved. The experimental design of our research closely parallels the one that she used.

Socrates: So if the research proved to be highly successful, the results would corroborate Eason-Watkins' model.

Dr. Williams: Correct.

Socrates: And what conclusions would be drawn if the research failed?

Dr. Williams: Our focus was on making the project successful.

Socrates: With such a large study and the commitment to this all-encompassing approach, the district must have had great confidence that the project would be highly successful.

But am I correct in assuming that if the project failed you planned to disseminate the results as vigorously as you would have for a successful implementation?

Dr. Williams: I am not directly associated with decisions on that level; however, the research evaluation was done by a very reputable and talented group—Metis Associates. And their reports are ready for public inspection.

Socrates: The study has completed its fourth year of implementation. Would you rate the project as a success or a failure?

Dr. Williams: That question seems to present a false dilemma. We learned a great deal from the project, some of which was disappointing.

Socrates: Well, let's consider the results a year at a time. After the first year, were there any statistically significant outcomes in favor of the experimental subjects in grades 6, 7, or 8?

Dr. Williams: As I recall, no.

Socrates: Correct. The Metis Associates reported that **there were no detectable performance differences between the experimental subjects and the controls.** Doesn't that mean that all the provisions for causing reading success failed?

Dr. Williams: Not at all. We don't expect a project to be fully implemented or even well implemented after the first year. At the end of the first year, the practices in the experimental classrooms had not been implemented enough for observers to determine if they were viewing an experimental school or a control group.

Socrates: What was the status after the second year of the implementation?

Dr. Williams: There were far more indications that the experimental schools were implementing the programs. There were more adults in the classrooms, more time spent on vocabulary development, more...

Socrates: Were there performance differences between the two groups?

Dr. Williams: Well, they were doing things differently than they had, and the experimental schools were still learning how to implement effectively.

Socrates: The Metis Associates summary for year two concluded that **“the program has not had a differential impact on any of the subgroups that were included in the analyses, including sub groups based on students' gender, race, and special education status.”**

No differential impact. Doesn't that mean that the project was a complete failure after the second year?

Dr. Williams: No, the schools were still learning new techniques.

Socrates: But there was no data of any differences between the experimental and control groups. The p value was .099, which is five times higher than it would have to be to show the most modest significance. The effect size was .05, which is only 1/3 the smallest acceptable effect sizes.

Dr. Williams: Yes, I'm quite familiar with the data. What's your point?

Socrates: Did you issue cautions? Did the data raise some questions about the apparent discrepancy between Eason-Watkins' accounts and the actual performance in the experimental classrooms? Why was there such a resounding difference, if these classrooms were doing what she did in her classrooms?

Dr. Williams: No, we did not get into that degree of speculation. We figured that it took time to implement fully. The new procedures were being incorporated in the experimental schools. We supposed that the results would follow.

Socrates: Did positive results follow in year three?

Dr. Williams: I suspect you already know the answer.

Socrates: Indeed. According to the year-three report by Metis Associates, the analyses did not reveal any significant overall impact of the Striving Readers Initiative on students' reading performance, as measured by their spring 2009 ISAT reading scale scores. In another place, the investigators reported that "there was no detectable overall impact of the program on Tier 2 and 3 students" Don't you find it amazing that there would be no performance differences after three years?

Dr. Williams: Of course.

Socrates: Then why didn't you publicize this egregious failure?

Dr. Williams: I don't dictate policy in those areas.

Socrates: But just from a humanitarian standpoint, people need to know that the Eason-Watkins scheme was either a hoax or her students happened to be greatly different than those in all 61 schools that took part in the experiment.

Dr. Williams: I think you're overstating the case.

Socrates: If you put in provisions for teaching vocabulary and comprehension and there are no resulting differences in these areas, how effective was the instruction?

If the instruction was totally ineffective, what is the possible justification for continuing it?

If the results of 31 experimental schools are the opposite of those achieved by Eason-Watkins, what is the probability that Eason-Watkins obtained her high scores by following the model she promoted?

Dr. Williams: Are you saying that she cheated and created false scores?

Socrates: No. I'm saying that the likelihood of her cheating and creating false scores is extremely high. Don't you agree?

Dr. Williams: Of course not. She is a person of high standards and I believe that her reported scores were accurate. We would not have embarked on this approach unless we had complete confidence in her veracity.

Socrates: But if not one of the 31 experimental schools showed any detectable improvement after the third year of the implementation, is there any reason to believe that the project procedures are resulting in improved student performance?

Dr. Williams: I think that's a philosophical issue beyond the scope of our discussion.

Socrates: Did you issue cautions based on the data and designed to alert teachers, schools, and researchers that the multifaceted approach Chicago implemented is ineffective?

Dr. Williams: As I indicated before, all the reports are available on line and may be found by a search of Chicago's Striving Reader Program.

Socrates: Teachers and administrators don't know that there is a serious problem with the experiment. They don't know that some of the techniques they may be using have been shown to be inert. You know that if the results of the experiment had been positive, you would have publicized the results extensively.

Dr. Williams: The data on the project is on public record, available for anybody who wants to report on it.

Socrates: If the study has gold standard design, doesn't a negative outcome have implications as strong as a positive outcome?

Dr. Williams: I understand your argument. I'm sure that the Chicago board considered these issues before deciding on a strategy for disseminating information. Furthermore, the study completed the fourth year, and in that year there was definitely improvement.

Socrates: I don't recall any material differences in the fourth year. The report indicated, "There was no overall treatment effect for all students in the ITT sample at the end of the fourth project year ...". No effects means no improvement, no differences

Dr. Williams: In the fourth year one of the 6 experimental subgroups achieved a statistically significant gain over the comparable control subgroup.

Socrates: Was this significance at the .01 level of chance?

Dr. Williams: No, it was the .05 level of chance.

Socrates: Doesn't that value tell how frequently the outcome would occur by chance?

Dr. Williams: Yes, it would occur by chance in 5 of every 100 measures.

Socrates: That's 1 for every 20 measures.

Dr. Williams: Correct.

Socrates: After four years, how many subgroups were measured?

Dr. Williams: 24.

Socrates: That's more than 20. Doesn't that mean that the significant outcome in year 4 could have occurred by chance and probably shouldn't have been reported for that reason?

Dr. Williams: I don't know that we viewed it that way.

Socrates: Given that's how it should be viewed statistically, in four years, the project did not manufacture one measurement that showed superiority of the experimental group over the control group.

Dr. Williams: Mmmm.

Socrates: Most important, the NCE scores for both the control and experimental 6th graders after year four were incredibly 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, not the 30's. Nothing of instructional importance occurred in this study.

Dr. Williams: (Shakes head.) I have a presentation in four minutes. I'll have to cut our discussion short.

Socrates: I believe that the title of your address is, "Gold-standard Studies, the Hope for Scientifically Based Instruction." I hope that some of the important things we discussed here find their way into your address.

Dr. Williams: Thank you for sharing your ideas with me.

End