Socrates on Gold Standard Experiments

Setting: Dr. Eugene Emry's office

Characters: Socrates and Dr. Emry, expert on experimental design of educational studies.

Socrates: What Works Clearing House and agencies like the American Institute for Research rejected experimental research that is more than 15 or 20 years old. What is the rational justification for this practice?

Emry: Several reasons. One is that the studies probably do not fit the current more precise analyses. Therefore, there are issues about how to interpret them. The most efficient solution is to either replicate them with more precision or simply ignore them.

Socrates: But why wouldn't the studies be addressed on a case-by-case basis to determine whether they are actually flawed?

Emry: We could do that; however, unless they provided for random assignment of experimental subjects, they would probably be rejected.

Socrates: What evidence is there that random assignment results in more valid outcomes?

Emry: Random assignment is justified on the grounds that it reduces possible confounds that could result from uncontrolled bias. It is one of the steps needed to assure that the only difference between the comparison group and the experimental group is the experimental treatment. For example, if the investigator is unaware of which individuals are assigned to the experimental and control group, a possible bias is

eliminated because the assignments are not directly controlled by the investigator, simply by chance.

Socrates: But how would somebody go about testing the effects of random assignment versus other reasonable methods of matching the performance of subjects in the control and experimental groups?

Emry: Possibly the best one could do would be to construct parallel experimental groups. The only difference between the two groups would be whether random assignment or some other means of assignment was used; everything else would be the same for both treatments.

Socrates: Was this method ever used?

Emry: Certainly. It was done in both medicine and education.

Socrates: And what were the results of these parallel designs?

Emry: (Chuckles.) The outcomes were probably more unusual than we would have predicted.

Socrates: Indeed. It is my understanding that the results of 500 randomized and non-randomized parallel medical studies led to two conclusions. The first was that the random assignment treatment performed higher than the other treatment on some occasions and lower on others. The second conclusion was that the differences were not predictable.

Emry: That's true, but the outcome data doesn't suggest that other assignment practices are as valid as random assignment.

Socrates: But does that same unpredictable relationship show up with educational parallel studies?

Emry: You could say that, yes. For example, comparison of randomized and non-randomized groups of high-risk males in training

programs showed that differences between the parallel groups were large at times, small at times, and as the authors concluded, “Generally unpredictable.”

Socrates: But in every case, you assume that the results achieved by the random-assignment groups are more valid, even though the unpredictable nature of the studies implies that some had serious confounds that were not identified, but that affected the results.

Emry: Yes, all things being equal, the random assignment would provide higher internal validity.

Socrates: The point is that all things are apparently not equal, even in studies that use random assignment and are judged to be of “gold standard” quality.

Emry: That seems to be an extreme conclusion.

Socrates: Well, let’s be specific. A few years back, California spent millions to create smaller classes in the primary grades. What motivated that change?

Emry: I suppose you’re referring to the Tennessee STAR study that compared class size and student performance.

Socrates: Exactly, and that study had large numbers of classrooms and used random assignment of classes. As I recall, a publication issued by the Brookings Institute observed “The STAR experiment offered *convincing evidence* that smaller class size can produce statistically significant and consistent, though modest, gains in student achievement.” Is that an accurate statement of how the project was generally appraised by the professional community?

Emry: Yes, the study seemed very promising.

Socrates: And when California revamped grades K through 3 so classes were smaller, did the change result in statistically significant, though modest, gains in student achievement?

Emry: No. There were no gains. But there are possible reasons for this disparity. The fact that some classrooms in a school had smaller classes while others didn't could have created jealousies and concern about why some classes had fewer students. Also...

Socrates: You are confirming the point I am trying to make. *Now* there are possible reasons, but before the fact there were no possible reasons. So in this case, which was the most accurate predictor of a valid outcome, randomized assignment or "other unidentified factors?"

Emry: That seems to be a redundant question. But you must keep in mind that there is more to internal validity than randomized trials.

Socrates: That's exactly my point, and if the authors of the study and professional evaluators cannot identify these problems before the fact, what possible difference does it make whether random assignment was used to control what seems to be a minor contribution to internal validity?

Emry: You're creating a false dilemma.

Socrates: Not really. There is clearly an uncertainty principle that characterizes the internal validity of educational studies. Given the evidence of uncontrolled variables that affect the outcomes, isn't it possible that in some of the cases, the more valid assessment of student performance was obtained by a study that did not use randomized assignment?

Emry: From an argumentative point of view, yes. But we will never really know because we obviously have no way to compare the

performance of either group to the “theoretically true performance.” Therefore, we have to base our judgment on careful scrutiny of the internal validity of each study.

Socrates: I completely agree. Our agreement, however, implies that older studies should not be categorically dismissed, but should be evaluated on a case-by-case basis, with a clear understanding of the possibility that some older studies may actually be closer to the “theoretically true performance” than some later studies that are unquestioned. This is particularly the case with older studies that involved very large numbers of students, because we presumably agree that larger numbers tend to equalize irregularities that occur in smaller populations.

Emry: Your conclusion about older studies is not in accord with current thinking in the field; you’re overlooking important factors in how life experiences of students have changed over the last 20 to 40 years.

Socrates: Did schools use normed tests 40 years ago?

Emry: Yes.

Socrates: Wouldn’t it therefore be possible to use norms as an indicator of how current children compare to those 40 years ago?

Emry: Yes but the achievement tests are renormed as populations change.

Socrates: Have the tests been renormed so that 50th percentile today is higher or lower than it was 40 years ago?

Emry: Overall, lower.

Socrates: It would seem to follow, therefore, that older experimental studies involving at-risk students would be particularly meaningful today because a larger proportion of today’s students would

be in that low range. The older studies from 40 years ago that would have far less potential application today would be those that evaluated gifted students, because there would be proportionally far fewer students in that range.

Emry: Well, that's an interesting interpretation, but it doesn't consider the issue of stimulation differences between now and then.

Socrates: But doesn't learning to read and learning elementary math today require the same skills they did back then?

Emry: Well, essentially yes.

Socrates: And don't some programs achieve large gains both now and then?

Emry: Even if that were true, I'm not sure I see your point.

Socrates: These studies serve as a credibility bridge. If something works then and works now, it confirms that the older studies are probably as valid as the current ones.

Emry: But if the later studies were flawed, that formula wouldn't work.

Socrates: Let's try a different angle. In medicine, psychology and other fields, older studies are recognized. For instance, the works on memory, like memorizing a list of nonsense words, are still recognized as being valid. The U shaped curve that represents the general order of which items in the list are learned earlier and later is still valid. And what about the studies that identified the causes of scurvy and yellow fever, or the results of doctors not washing their hands before delivering babies? By your standards, these studies would have long since been discarded and Madam Curie, Semmelweis, and Salk would have been written off as oddities of the past, while education would have been officially born as a

legitimate scientific endeavor around 1980. Doesn't that strike you as very inconsistent and without apparent reason?

Emry: It seems we're covering the same ground.

Socrates: Did it occur to you that only one educational approach is seriously affected by removing earlier studies?

Emry: No, it hasn't.

Socrates: As I understand it, Direct Instruction was validated by more than 50 studies that occurred before the cut-off date. No other approach has more than one or two scientific studies. Direct Instruction is also supported by more recent studies that confirm that the approach is as effective now as it was then. Some critics of the ban on referring to older studies suggest that the cut-off date was specifically designed to divest Direct Instruction of its rich data base.

Emry: That's absurd. I know of no such machinations.

Socrates: Possibly not, but I presume you do know that What Works Clearinghouse indicates that DI has virtually no evidence of effectiveness in teaching beginning reading, when in fact there are more than 100 studies that attest to DI's effectiveness in teaching reading.

Emry: You'll have to excuse me if I don't agree with your conspiracy theory.

Socrates: You seem to be denying what appears to be a clear pattern. But let's move on and look at the problem of what works from the standpoint of external validity. As I understand it, external validity is really nothing more than a rational argument about the extent to which the results of an experiment could be generalized beyond the population used in the study.

Emry: Yes, I suppose you could call it an argument.

Socrates: So what are the threats to external validity?

Emry: Well, in your terms a threat would be established by any argument that identifies a confound in the experimental procedures that could limit the generalizability of the outcomes.

Socrates: What are the *unique* threats to external validity in the context of what works?

Emry: I don't understand what you mean by "unique threats." The threats to external validity are aptitude-treatment interactions, pre-test effects, post-test effects, the Hawthorn..."

Socrates: Those aren't unique. The unique threats are those that have to do with the relationship between the experimental conditions and the conditions of the person who is looking for a program that works.

Emry: I'm not sure I understand.

Socrates: Let's say we are a school district that uses reading program X. We use achievement test A. The test results over the past few years show that our at-risk first graders perform in the 18th percentile at the end of the year. We want to find a better reading program. We read about a study that involves a population something like ours but with a lower percentage of blacks. The study used random assignment and is listed as meeting the *gold standard* of experimental design. The first-graders in the experimental group entered higher than our children but they scored considerably higher than our children at the end of the first grade. Also, the results are based on achievement test B, not A. To what extent are those results valid *for us*?

Emry: Assuming there are no confounds, quite valid.

Socrates: Wouldn't anything that is different between our situation and those in the experiment present a possible confound?

Emry: Yes, there are situational-specific confounds, but I'm not sure how they would apply here.

Socrates: We use achievement test A. The study used achievement test B. Is that a situational confound?

Emry: I suppose it could be if the norms for the two tests aren't the same, but you're talking a small difference if both instruments are normed.

Socrates: So you don't seriously question whether the achievement tests are valid measures of achievement.

Emry: No. If they are properly constructed, with all items having construct validity and being correlated with the total score, there is no reason to question the tests.

Socrates: But would the study have more validity or less validity for *us* if the study used our achievement test?

Emry: Possibly more.

Socrates: And could this confound possibly counter whatever benefits in validity were created by random assignment?

Emry: I have trouble with that conclusion. We don't know the magnitude of either variable.

Socrates: Let's try another item. The description of our hypothetical study indicated that the population in the experiment was ethnically different from ours. Couldn't that create a confound for us using the program?

Emry: Yes, I suppose so.

Socrates: So now we have two possible variables that could counter the possible benefits of random assignment.

Emry: I don't like the idea of keeping score in this manner.

Socrates: Here's another difference. The pretest performance of the students in the experiment is higher than ours. Couldn't that create a very serious confound?

Emry: I suppose you could argue that it's a confound, but I don't know about it being very serious.

Socrates: As I understand the instructional law of populations, a program that works well with lower performers will always work well with higher-performers, but I know of no data that suggest that if programs work well with higher performers, they will consistently work with lower performers.

Emry: Since I'm not familiar with studies that address this issue, I'll concede that a confound is possible.

Socrates: So on the negative side we have at least three possible confounds that should make us cautious about adopting the program, and logically these situational differences could more than offset the fact that the study used random assignment. In other words, it is not a gold-standard study for us. It might not even be a brass standard.

Emry: Mmmm.

Socrates: Do you agree that these context differences imply how we could conduct an experiment that would have gold-standard validity for *our* context?

Emry: I'm not sure I follow you.

Socrates: We simply design a study that avoids all the problems of external validity the reported study had. We run a trial that uses a representative sample of our at-risk students; we measure the performance on achievement test A; and we compare the results of the study to the performance of our current population. For the study to be perfectly valid for us, we wouldn't even need a comparison group, only an experimental group that uses the program we're considering adopting.

Emry: Now I think you've gone too far. You must have a concurrent comparison group if you hope to conduct a proper experimental trial.

Socrates: In terms of what you have said the comparison group is unnecessary.

Emry: What did I say that could possibly lead you to that conclusion?

Socrates: You said that you had great faith that appropriately normed instruments are reliable and valid measures of achievement. You also said that data based on our achievement test would probably be more valid for us than data on another normed test.

Emry: That's true but I don't see how those facts support the conclusion that you wouldn't need a comparison group.

Socrates: Well, let's say we have a comparison group that is perfectly representative of our at-risk population. At the end of the first-grade year, what do we do to show how well this group performed?

Emry: This question seems very elementary. We test the students, record the scores, and then analyze them to determine the gain.

Socrates: So is it fair to say that the only way we use the comparison group at the end of the experiment is to reduce these children to scores and numbers that are then analyzed?

Emry: That's an argumentative way to state it.

Socrates: Possibly. But could you describe how the data for the comparison group and the experimental group would be used differently if we didn't have a comparison group but simply made up a series of scores that are representative for our at-risk students?

Emry: There would be no difference in the calculation, but the scores are not authentic, whereas they are authentic if data are based on actual responses.

Socrates: True, but they are representative of our current student population. Therefore, what difference would it make if we used scores, instead of reducing live bodies into scores that we use in exactly the same way?

Emry: The procedure that you describe is not consistent with the intent of experimental trials. A real comparison is necessary.

Socrates: Really? What would we say if a real experimental group had representative scores on the pretest, but the real comparison group did not have representative pretest scores?

Emry: That the groups are not sufficiently matched.

Socrates: And what would we say if the groups were matched at pretest but the comparison group had post-test scores that were much higher than the post-test scores of our at-risk population?

Emry: Probably that there was some kind of confound, assuming that the control subjects used the same material that the rest of the population used.

Socrates: So in both conditions, if you don't like the numbers, you reject them or effectively change them through some kind of statistical

"adjustment." Obviously, you feel you have license to change scores you don't like and substitute "unauthentic scores" for them. Yet, you reject the notion of having unauthentic scores for a comparison group even though these scores require no further adjustment, and are perfectly representative of our population. So is it simply an arbitrary decision on your part about when made-up scores are permissible?

Emry: No. Our scores are not made up. They are adjusted in a manner that does not jeopardize the internal validity of the study.

Socrates: Neither do the made-up scores I propose. Consider the internal validity of the study if the experimental group was perfectly representative of our at-risk population with respect to demography and scores, and the only difference in reading instruction was that the experimental group used program Z, not X. The instructional time and time of day were the same, the number of periods was the same, the experimental teachers had taught program X and obtained typical scores. How could the outcome of using program Z be caused by anything other than the program itself?

Emry: Well, since you're describing a setting in which the internal validity would be high, it would be high. But that doesn't mean that there would not be problems of external validity. There could be a novelty effect, a teacher-motivation effect and other possible confounds.

Socrates: You make an excellent point, but it is unlikely that any novelty effects would involve the students. For them the program is new, whether it is the experimental program or program X. There could be a novelty effect for the teachers, which could be estimated by having some of the other classrooms do something different, such as having visitors observe the reading period every Thursday. If these classrooms have a 2-point higher score than they had traditionally, we deduct two points from

whatever score the experimental group had. In fact, however, whatever novelty effect the teacher might experience by having a new program would be more than offset by unfamiliarity with the program. Studies involving Direct Instruction teachers show that their performance improves during the first three years they teach the program.

Emry: The major problem I see with your analysis is that it does not take into account the fact that what you are describing is a norm-referenced design, which simply compares experimental scores to the norms. This design had been used somewhat in the 1970s, but has later been rejected.

Socrates: Two points. We are limiting our discussion to what works for us. In this context, the norm-referenced experiment has high internal validity. The external validity is also high because we are the ones who will use the results. The second point is that G. Kasten Tallmage did several analyses of studies that showed good correspondence between results of true experiments with the results that would be obtained if they had only an experimental group that was referenced to the test norms.

Emry: What disturbs me is your apparent rejection of the gold standard. The standard has been established by thoughtful analysts who consider both the philosophy of science and psychology to create standards for *proper* experiments.

Socrates: As you may have gathered, I believe that you are depriving school districts of simple formulas that they could apply to perform experiments that do not require doing things much differently than the way they do them now. Yet, these experiments would provide quite valid information about what works in specific districts. These districts do not need random assignment or even comparison groups to perform experiments that are tailored to each district, its history, its

achievement tests, and the specific details of its demography. Districts are intimidated by the aura of gold-standard "experiments" when in fact the task of discovering gold-standard information about what works is less complicated than installing a new reading program. Norm-referenced studies let districts do something constructive with their achievement test data, rather than going through the ritual of administering the tests and then filing away the results, with the hope that everybody will forget about them.

Emry: Well, I would prefer to think that the field would be much farther ahead if it conducted studies that have a sound scientific base.

Socrates: It appears however, that you are extremely selective with what you choose to recognize as scientific. For example, why would one have faith in the norms of tests, and yet fail to recognize the most obvious implication of the norms—if you have data on your students, and if this data is expressed in terms of norms, there are some contexts in which you don't need comparison groups.

Emry: I certainly don't agree with that conclusion. Our requirement for the use of randomly assigned comparison subjects is consistent with the philosophy of scientific inquiry.

Socrates: With respect to the philosophy of scientific inquiry, are you familiar with the works of Abraham Kaplan?

Emry: No.

Socrates: In 1964, he wrote *The Conduct of Inquiry.* In it, he postulated the Law of the Instrument. Are you familiar with that law?

Emry: No.

Socrates: Here's the law: "Give a small boy a hammer and he will find that everything he encounters needs pounding." I think that law

describes the field's preoccupation with variables like random assignment, which are often impractical and are trivial in the overall scheme of what works.

Emry: It appears that you and I see the issues from irreconcilably different perspectives.

Socrates: I agree.

End