

It Feels Fine Once You Get Used To It: Lessons from the 21st Century National Evaluation¹

Remarks Presented at the Brookings Institution's Stakeholder Forum on After-School Programs

**Mark Dynarski
Senior Fellow
Mathematica Policy Research, Inc.
July 10, 2003**

Two events took place concurrently on February 3, 2003: the U.S. Department of Education released Mathematica's first report on the 21st Century program, and the Bush Administration issued its FY 2004 budget calling for a \$400 million funding cut for the program. This confluence of events set waves in motion that continue to swirl, prompting actions by advocates for after-school programs, foundations funding these programs, researchers studying them, and a well-known athlete/actor who had successfully led an initiative to get state funding for after-school programs.

Mathematica's findings showed that students did not attend after-school programs much and, after one school year had elapsed, did not show improvements in academic or developmental outcomes. The study, which included thousands of students attending programs in over 40 school districts and almost 90 schools, used methods designed to ensure that measurement biases were minimized. No research on after-school programs using these approaches and at this scale had been done before, partly because after-school programs are relatively new and, indeed, the 21st Century program received its first infusion of funds only in 1998.

There is no question that studies like this one will become more common in the future. *No Child Left Behind* and other legislation put a major emphasis on rigorous studies like it. As a society, we can only benefit by strengthening the scientific basis for programs to help children. But if education studies all have the same turbulent experience as the 21st Century one, efforts to shore up this type of research could falter. It is not hard to imagine a push to return to education research as management consulting, as Professor Thomas Cook has described it in a series of recent essays.²

¹ I want to thank Isabelle Sawhill, Ron Haskins, and Andrea Kane for organizing the forum and for helpful comments, and Mary Moore, Joanne Pfliegerer, Elizabeth Warner, and Roberto Agodini for comments and suggestions. These remarks do not necessarily reflect the opinions or views of Mathematica Policy Research, Inc. or the Brookings Institution. Any errors are mine.

²See T.D. Cook, "Sciencephobia: Why Education Researchers Reject Randomized Experiments," *Education Next*, vol. 1, no. 3, 2001; T.D. Cook, and M.R. Payne, "Objecting to the Objections to Using Random Assignment in Educational Research," in F. Mosteller and R.F. Boruch (eds.), *Evidence Matters:*

The 21st Century experience has provided many lessons that could be built on in future studies. Like a swimmer jumping into cold water who is asked how it was, I can say that it was cold at first but you get used to it. I offer the following four suggestions as we move forward.

1. *Weight the evidence consistently.* A reader of any one study may believe some aspects of a study's design fall short of a desired standard. If presented with another study with a similar design, but more appealing findings, the reader should apply the standard consistently. The 21st Century experience suggests that standards are applied inconsistently, which may contribute to the public's confusion about what has been learned.

Ultimately, design trade-offs are required in all studies, and the 21st Century one was no exception. After-school programs serving middle-school students were typically under-enrolled, so this part of the study could not use an experimental design. Instead, we used a sophisticated matching approach that was data-intensive but had appealing theoretical properties. To raise the level of generalizability, we selected a random sample of programs serving middle school students, which enabled the results to be nationally representative. Critics noted that the resulting comparison group and the group of program participants differed on some characteristics. Comparison designs generally have this weakness, which probably is the single strongest argument for using experimental designs. However, critics downplayed the fact that the other large evaluations of after-school programs they were promoting had used the same comparison group method, usually at a lower level of sophistication and without the nationally representative feature of the 21st Century study. Applying a standard consistently means that critics should have discounted findings from all these studies, rather than arguing that findings from the 21st Century study were flawed while findings from the other studies provided solid evidence.

On the elementary school side, some after-school programs were over-subscribed, so this part of the study could use an experimental design. Yet again a trade-off was needed. Programs that were over-subscribed were not a random sample of elementary school programs, so findings were not nationally representative (and our report cautioned readers about lack of generalizability). At the same time that critics argued this lack of representativeness meant the findings should not be generalized to all elementary school programs, they also held up positive findings from studies of programs that had operated in only one school district or in one city. If a study must be generalizable for its findings to support a funding cut, as critics appeared to suggest, then generalizability also is required for findings used to support a funding increase. Yet if you accept this premise,

Randomized Trials in Education Research, 2002; and T.D. Cook, "Why Have Educational Evaluators Chosen Not To Do Randomized Experiments?" *Educational Evaluation and Policy Analysis*, in press.

how did the program grow so large between 1998 and 2003 without a generalizable study being done during this period?

To further support their argument that the 21st Century study was flawed, critics held up the overall pattern of positive findings from previous research on after-school programs and said the 21st Century study was an outlier that should not be given much credence. The comparison raises two issues.

First, all studies are not created equal, and study design and methods must always be scrutinized. As recent findings about the effectiveness of hormone replacement therapy point out, one rigorous study that yields markedly different findings from previous research can lead to sharp shifts in thinking about what is effective.³

Second, the 21st Century study's findings differed from the findings that have been *promoted* in literature reviews on the effects of after-school programs. A more thorough and balanced review will find fewer differences. Studies cited to support the program's growth used different methodologies and looked at after-school programs of many different shapes and sizes. The reviews paid little attention to the underlying selection biases occurring at three levels. First, few of these studies used random assignment, so selection bias probably entered into who participated in the programs, how long they stayed, and how frequently they attended. It is well known that more motivated students can have better outcomes than less motivated students, no matter what program they are in. Second, only certain studies, generally ones with positive findings, were cited. But if 100 studies find no effects and 10 studies find some effects, it is imprudent to ignore the 100, focus on the 10, and assert that the evidence shows programs are effective. Furthermore, publication bias is a well-known issue in the research world and always should be a concern in literature reviews.⁴ Third, positive findings were highlighted and insignificant or negative findings were ignored or not discussed. This is a variant of publication bias that can be even more misleading. When a research study states "many studies were reviewed," it conveys a powerful message. But how many findings actually were examined? How large a proportion was positive?

2. *Evaluate programs when they are small.* Donald Campbell's well-known aphorism to "evaluate programs when they are proud" is often cited in the program evaluation literature. I think the intent of his message is that evaluators should wait for

³See Susan Okie, "Hormone Treatment Is Called Harmful: Menopause Study Cites Health Risks," *Washington Post*, July 10, 2002; Gina Kolata, "Hormone Therapy, Already Found to Have Risks, Is Now Said to Lack Benefits," *New York Times*, March 18, 2003. The trial was stopped due to the risks of the treatment and final results recently were published; see J. Manson et al., "Estrogen Plus Progestin and the Risk of Coronary Heart Disease," *New England Journal of Medicine*, vol. 349, no. 6, August 7, 2003.

⁴If any filter is used, I suggest that studies with weak designs be downplayed and studies with strong designs be the focus, following the example set by Professor Lawrence Sherman and his colleagues in their study of what works in preventing crime. The study can be downloaded at www.preventingcrime.org/report/ (accessed 9/2/03).

some time to elapse so that new programs can iron out any problems they may have as they are getting started, and then evaluate them. Evaluators are in Donald Campbell's debt for his many contributions to the field and we continue to read his work and benefit from its thoughtfulness and rigor. But in the policy research arena, his advice presents two problems.

First, what exactly does it mean for programs to be proud? How do we observe pride? Is the advice simply to wait for implementation to stabilize? If so, are programs ever finished tinkering with implementation? Won't they always be working on a few things and able to say it is not yet time for an evaluation?

Programs should not be determining the appropriate time for evaluation. They do not view the quality and effectiveness of their services through the same skeptical lenses that evaluators typically wear, and we should not be surprised if program staff believe evaluation is unnecessary. It is up to someone else—funders, most often—to ask for evidence of program effectiveness. And for programs that possibly have negative effects (for example, some of the results in the 21st Century study suggest that older students misbehaved more when they attend after-school programs), the case for early evaluations is even stronger.

The second problem with Donald Campbell's advice is that even if it could be followed, the time lapse may be too long. As noted, the 21st Century study was criticized for looking at the program early in its implementation, but the public might wonder when else the program was supposed to be studied. Funding was increasing dramatically. Moreover, how would we know the program needed modifications without evaluating it? Case studies of programs—formative evaluations—can provide valuable information about program practices, but they cannot provide information about effectiveness. For that, we have to measure effects.

Studying programs when they are small fits well within the trade-offs of public policy. Demonstration programs and small-scale implementations are excellent opportunities to look at efficacy—whether programs *can* work. If the findings are promising, full-fledged programs can then be launched. Perhaps policymaker patience is sorely tested by having to wait for researchers and program designers to learn what is going to work. However, a central principle of medical research is that it takes time for evidence to accumulate, as studies progress through limited efforts to large-scale clinical trials. In the medical arena, the public has proved willing to wait for information about treatment effectiveness. In education, being patient in developing sound programs is likely to be worth it too.

3. *Be clear about program goals and likely effectiveness.* What does it mean to say a program is effective? Effective at what? What do we want the program to do? For after-school programs, is the goal to care for children after school? To improve their academic skills? To develop emotional and social competencies? Is one objective more

important than another? Clear direction from program designers and funders about a program's primary and secondary goals provides a basis for specifying a precise set of outcomes for study.

Clear expectations about how large an effect programs can generate also would be useful. Evaluation designers set target effect sizes for what their evaluations will be able to detect, and much larger evaluations are needed if programs are expected to achieve small effects. However, if funding is related to potential to achieve large effects, program proponents have an incentive to overstate likely effects. If the actual effects are then measured and found to be smaller--and perhaps statistically insignificant--those who had expected much larger effects may be disappointed or frustrated, but their expectations may not have been sensibly grounded in what programs could accomplish within their resources and designs.

4. Increase objectivity in education research. Currently, many people view findings as implicit expressions of what the study's researchers would like to see policy do. Media flare-ups around the release of findings contribute to a sense that a study in question, like others before it, is simply a form of politicking. The media's appetite for conflict fuels the reporting of findings as more politics than science.

In light of the widely held view that education research lacks objectivity, mechanisms to increase objectivity may need to be as stringent as those in place in the hard sciences. Ways to promote public confidence in findings as objective scientific interpretations rather than subjective value judgments are worth considering. The Department of Education's What Works Clearinghouse will be crucial in this area. In the end, rigorous methods--especially random assignment--and replication contribute to objectivity. However, random assignment is rare and replication even more rare. Promoting rigorous studies of moderate size with different teams of researchers, all analyzing the same program and its outcomes, may be worth considering.

In his recent address to the American Educational Research Association, Robert Slavin said that "Evidence-based policies could finally set education on the path toward the kind of progressive improvement that most successful parts of our economy and society embarked upon a century ago."⁵ He implored researchers to "set to work generating the evidence that will be needed to create the schools our children deserve." We do not know how much of our experience with the first 21st Century report reflects growing pains from this early stage of doing studies of education programs using new and rigorous methods. Like jumping into cold water, the episode soon may seem just a temporary discomfort, until we get used to it. In any event, jumping in will have been worth it if it contributes to a stronger basis of evidence about effective education programs and progressive improvement for a vital piece of our society.

⁵Robert Slavin, "Evidence-Based Education Policies: Transforming Educational Practice and Research," 2002 DeWitt Wallace Distinguished Lecture, *Educational Researcher*, vol. 31, no. 7, pp. 15–21.