STATEMENT RELEASED BY

Members of the Scientific Advisory Board for the 21st Century Community Learning Center Evaluation

Joan S. Bissell, Christopher T. Cross, Karen Mapp, Elizabeth Reisner, Deborah Lowe Vandell, Constancia Warren, Richard Weissbourd

May 10, 2003

A great deal of attention has been given in recent weeks to the first year report of the 3-year evaluation of the 21st Century Community Learning Centers (CLC). As researchers and educators who served as independent technical advisors on that study, we feel it is imperative that there be a full understanding of a number of issues.

We believe that the first year report, released on February 3, 2002, has serious methodological problems that call into question its findings and that violate basic principles governing how evaluation should be used to guide policy and affect program budgets.

To provide appropriate and sound information to policy makers, program evaluations must be based on rigorous research methods and a substantive understanding of the issues central to the evaluation. In the text that follows, we delineate our concerns as part of the professional scrutiny and critique that scientific inquiry entails. A practice that is, indeed, central to the scientific process.

I. Nonequivalent Treatment and Comparison Samples in the Middle School Evaluation

Our first concern pertains to substantial baseline differences between the middle school treatment and comparison groups. In the Draft Interim Report (dated September 12, 2002), significant differences in standardized test scores were reported between the two groups. As shown in Table III.1 of the Draft Interim Report (p. 57), the mean percentile reading scores for the treatment subjects at baseline was 39.7, whereas the mean percentile reading score for the comparison subjects was 50.0 at baseline, $p < .00$***. The mean percentile score in math for the treatment group at baseline was 34.3 percentile versus 43.6 percentile for the comparison group, $p < .00$***.

At its meeting with MPR and the Department of Education in January 2002, several TWG members advised that these test score differences indicated that the comparison students were substantially advantaged relative to the treatment group. Given the baseline test score differences, members of the TWG further advised MPR and the Department of Education of the necessity of addressing this fundamental problem. Otherwise, biased results would occur.

This problem was not addressed in the First Year Report. As indicated in Table B.2 of the report (see p. 140), baseline test scores were not controlled for in the impact analyses. Furthermore, the report released in February 2003 does not acknowledge the likelihood that

findings were biased because of these baseline differences. In fact, these baseline scores were selectively omitted from the report (see Table III.I on p. 55 and Table A.8, p. 127). In our judgment, the first year analyses and any subsequent analyses are un-interpretable given these substantial baseline differences on key measures.

The First Year Report states that the 21st CLC evaluation is the "most rigorous examination to date of school-based after-school programs" (p. xi) and that the analyses revealed that the programs had "limited academic impact" (p. xii). In light of the substantial baseline differences that were ignored in subsequent analyses, these statements are not justified.

As you know, the First Year Report was used as the justification for a recommendation to cut the 21st Century CLC budget by 40%. If funding decisions are to be based on scientific evidence, it is critical that all relevant data be included within reports.

II. Problems in the Elementary School Evaluation

We also have substantive concerns about the elementary school evaluation: these relate to the small sample size, premature release of the findings, as well as the fact that the programs were not representative of the total universe of such programs.

Sample Size. The sample size for the elementary schools is so small that it raises questions about the validity of the report conclusions. The elementary school evaluation was conducted in 18 elementary sites operated by seven 21st CCLC grantees. To place these numbers in context, there were over 6,800 sites in the program operated by more than 1,400 grantees in 2001-2002. The grantees included in the sample represented one-half of one percent of this number. The elementary grade sample included a total of 403 children in 21st CCLC centers (the treatment group) and 226 on waiting lists (the control group) (p. 100), the equivalent of less than one percent of children in the program in 2001-2002. On key variables, the sample size was considerably smaller. For example, in the case of the standardized reading test, the sample size was 278 for the treatment group and 148 for the control group. A valid sample, one that would yield generalizable results, would be considerably larger, certainly more than double this size.

It was recognized that the size of the elementary school sample had to be augmented and that this was achieved by randomly assigning another 1,600 students in seven additional elementary grantees in the fall of 2001. Given the small sample size in the first cohort, used to generate the data in the released report, and the plan to conduct analyses on the whole sample later in 2003, we do not understand the justification for presenting the preliminary findings in the February report as being conclusive.

Mismatch Between Focus of Sampled Programs and Outcomes. The decision to present preliminary data was particularly problematic because in accordance with federal program guidelines that existed at that time, four of the 18 programs (22%) had only an incidental focus on academic and developmental experiences for children. These four programs focused on serving adults in the school's community, and children attended the center only when they accompanied their parent or grandparent. It is not clear why or how these adult-focused programs would be expected to impact child outcomes.

It also is important to note that federal guidance has changed since that time, casting further doubt on the connection between the evaluation of a program that existed in the first year and programs now being funded.

Inconsistencies Between the Executive Summary and the Full Report. A number of conclusions reported in the January 2003 Executive Summary warrant further explanation. The data reported in the body of the report and in earlier documents appear to be at odds with parts of the Executive Summary.

In particular, the Executive Summary states that there was limited academic impact of the 21st CCLC program. It reports:

> "At the elementary school level, reading test scores and grades in most subject were not higher for program participants than for similar students not attending the program." (p. xii).

It is unclear why the statistically significant findings in social studies/history or the overall pattern of higher grades for treatment students are not included in the Executive Summary.

The report found that:

> "Centers increased grades in social studies significantly (the effect size is 30 percent), but while grades in other subjects generally appeared higher for treatment students, the differences were not significant." (p. 96). In interpreting these findings, it is important to recognize the effect of the small sample size in diminishing the levels of statistical significance.

Need to Report Mathematics Data. As was the case with the report of the middle school evaluation, it appears that key data were not included about elementary programs. Baseline and follow-up standardized test scores in mathematics were reported in the Sept. 12 report (pp. 93 and 97), but omitted from the report that was released in February (pp. 93 & 97). These data suggest that the performance of the treatment group improved whereas the performance of the control group decreased across the school year.

Need to Examine Changes Related to Student Safety. The Executive Summary reported "No improvements in Safety and Behavior" and stated that "Programs did not increase students' feeling of safety after school." The data on student safety seem discrepant with this. At baseline, treatment students were reported to be "somewhat less likely than control students to feel safe walking in their neighborhoods" (p. 93). Within the treatment group, 72.3 percent of students felt safe walking in their neighborhood, while 78.5 percent of control group students felt safe doing so. The end-of-year outcome data indicate that the children in the program had an increase in their sense of safety relative to their peers: 74.3% of students in the program reported feeling very safe after school as did 75.5% of the control group. It is particularly noteworthy that among children who had participated in the 21st CCLC programs, only 1.7% said that they felt "not at

all safe" after school while 5.2% of control group students indicated feeling this way during the after school hours (p. 100 of the report).

III. Level of Student Participation and the Impact on Outcomes

The 21st CCLC evaluation reports that students attended the after-school programs in both the elementary- and middle-grades samples for relatively limited amounts of time. Students enrolled in the elementary-grades program sample participated an average of 58 days per year (out of a typical 180-day school year). Students in the middle-grades sample attended 32 days per year on average. These attendance facts, along with other information presented in the evaluation's first-year report, suggest that, while students may have been engaged in beneficial experiences after school, they did not experience enough "dosage" for these experiences to result in measurable differences in outcome measures.

With respect to dosage, the third-year interim report of the evaluation of The After-School Corporation (TASC) programs found that students typically required exposure to after-school services over two years and for a minimum of 60 days per year and an after-school attendance rate of at least 60 percent to demonstrate improvements in educational performance.[1] Studies of LA's BEST[2] and of California's After-School Learning and Safe Neighborhoods Partnerships Program,[3] while non-experimental in design, have also suggested dosage effects, with increased effects most notable when students participate for over 150 days, or for more than one year.

In the case of the 21st Century programs included in the evaluation, which were measured over a period of only one year, relatively few students in elementary and secondary samples participated for 60 days or more. Other studies suggest that the overall low dosage among the students included in the sample may be the primary reason for the absence of statistically significant results on outcome measures.

IV. Treatment Contamination in the Comparison and Control Groups

At a meeting of the scientific advisory board, evaluators from Mathematica Policy Research, and staff from the U.S. Department of Education held in January 2002, we discussed the serious challenge to treatment fidelity that occurs if comparison students receive services that are comparable to those received by the treatment group. As discussed at the January meeting, the implementation data indicated that some site were funding some after-school programs with 21st Century dollars and other after-school programs with state and local dollars. In essence, the programs differed only in their source of funding. This critical problem was not addressed in the

---

[1] M.E. Welsh, C.A. Russell, I. Williams, E.R. Reisner, & R.N. White. *Promoting learning and school attendance through after-school programs: Student-level changes in educational performance across TASC's first three years*. Washington, DC: Policy Studies Associates, Inc., October 31, 2002.

[2] Huang, D., Gribbons, B., Kim, K., Lee, C. and Baker, E. (2000). A decade of results: *The impact of the LA's BEST after school enrichment program on subsequent student achievement and performance.* Los Angeles, CA: UCLA Center for the Study of Evaluation.

[3] Bissell, J. et al. (2001). *Preliminary Report: Evaluation of California's After School Learning and Safe Neighborhoods Partnerships Program: 1999-2000.* Sacramento, CA: California Department of Education.

report. Given this fact, the absence of statistically different effects for the evaluated 21st CLC programs is hardly surprising.

V. Violations of Accepted Principles of Scientific Research in Education

A 2003 report from the National Academy of Sciences identified 6 principles that are recognized as essential for scientific research in education (Committee on Scientific Principles for Education Research, 2003).

One of these principles is replication and generalization:

> "Scientific inquiry emphasizes checking and validating individual findings and results. Since all studies rely on a limited set of observations, a key question is how individual findings generalize to broader populations and settings. Ultimately, scientific knowledge advances when findings are reproduced in a range of times and places and when findings are integrated and synthesized." (p. 4)

A second principle is professional scrutiny and critique:

> "Scientific studies do not contribute to the larger body of knowledge until they are widely disseminated and subjected to professional scrutiny by peers." (p. 5)

The use of evaluation findings from the first year as a basis for policy recommendations is a violation of both of these standards. It is inappropriate in this particular case and misleads policy makers into thinking that it is appropriate to judge the outcomes of social interventions after a single year. It reinforces the fallacious assumption that interventions such as the 21st Century Community Learning Centers are capable of achieving rapid results with far fewer resources in terms of money, time and staff capacity than the schools attended by the youth they serve.

Good evaluation practice recognizes the arc of program implementation, particularly when the program being evaluated is new in both funding and form. The coupling of the release of the First Year Report and its role as the foundation for the recommendation to cut the program by 40% failed to recognize these standards and principles. When evaluations fall short of these standards, they weaken not only the value of the information in the particular case at hand, but also the quality of the public policy informed by the evaluation and the confidence we can place in program evaluations. It is damaging for both sound public policy and for evaluation as a tool for guiding policy making to subject new program effort to being held accountable for achieving results in an unrealistically short period of time.

Actions such as this may undermine the very important movement currently underway to set higher standards for conducting research in education, a change that is long overdue and one that we endorse. Setting the bar higher should be a great benefit to improving the quality of education received by children across the nation. Unfortunately, that promise will be extinguished long before its promise is achieved if policy is made on the basis of research that fails to meet the quality standards that are essential to sound decision-making.