Jane Butler Kahle

Evaluation of Systemic Reform: Learning while Doing

One state's effort to evaluate the outcomes of its systemic reform elucidated several issues critical to most large-scale efforts in education

Introduction

"The theory of systemic reform rests on some assumptions that should be examined and tested. First, systemic reform seeks greater coherence, an alignment of policies, but the education system is fragmented by design—fifty states, fifteen thousand districts, countless other agencies impacting the schools—and this fragmentation is intended to permit variation. The agencies of government responsible for the schools are divided from each other by the federal structure and by the separation of powers. They are further divided by powerful traditions of local control and parental rights. On top of that within any given jurisdiction there are a variety of stakeholders each with their own views about standards, assessment, and locus of authority" (Corcoran, 1997, p. 64).

Beginning in 1990 with the National Science Foundation's (NSF) Statewide Systemic Initiative (SSI) program, science leaders at the state, district, and local levels increasingly have faced accountability issues. Today both private and public funding agencies recommend that proposed projects be based upon scientific research; that is, research that meets the criteria delineated in the National Research Council's book, *Scientific Research in Education* (Shavelson & Towne, 2002). Because evaluation that provides scientific-based research is the crux of the *No Child Left Behind (NCLB)* Act (Paige, Hickok, & Newman, 2000), it remains a top priority for science teachers and supervisors. This paper describes one state's efforts to evaluate its SSI, including issues that emerged during the evaluation as well as the lessons learned from the reform and its evaluation. Those lessons are particularly pertinent today as science teachers and district supervisors seek to meet the requirements of NCLB.

One of the goals of Ohio's SSI was to narrow any achievement gaps between identifiable subgroups of students, e.g., between boys and girls, between African American and European American students, and/or between students from different economic backgrounds.

Beginning in 1994, Ohio's Statewide Systemic Initiative (*Discovery*) began to evaluate the outcomes of its reform.[1] Ohio, similar to most SSI states, focused its reform on professional development. However, it differed from others in that it offered long-term (six week), content institutes that were taught by inquiry[2] and provided several ways for teacher participants to receive support during the academic year (follow-up meetings, electronic networks, classroom visits from master teachers and scientists, etc.).

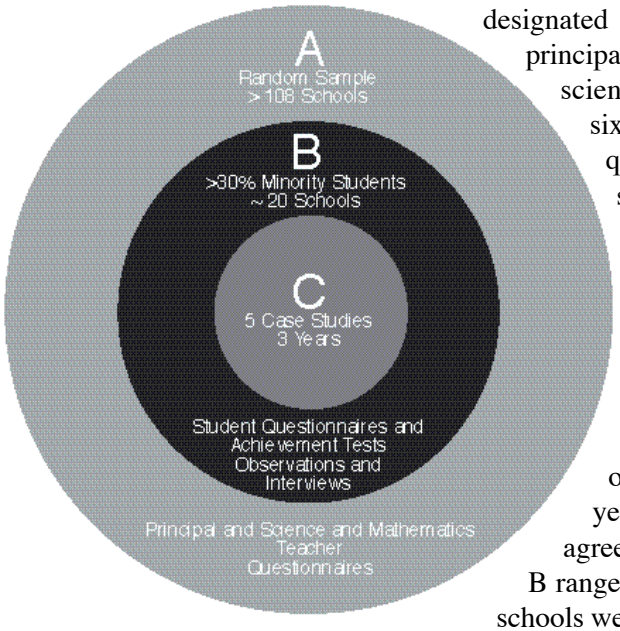
I'll describe the types of issues we faced as we attempted to evaluate systemic reform concurrently with doing it. The evaluation spans a period of seven years and was supported by both the SSI and a subsequent research project, funded by the NSF. Although the descriptions are specific to Ohio's reform, the issues faced as well as the lessons learned may be generalized to any large-scale reform effort.

What Issues Emerged?

Initially, we needed to know if the type of professional development we offered was indeed changing teaching practices. A series of studies,

comparing teaching practices in SSI and non-SSI classrooms indicated that teachers who had participated in the institutes and follow-up activities used inquiry methods (extended questioning, open-ended laboratories, and student-generated hypotheses) more frequently than non-SSI teachers did. As the reform progressed, we sought to find out if achievement differences by subgroups of students had narrowed or disappeared.

Figure 1. Nested Research Design



One of the goals of Ohio's SSI was to narrow any achievement gaps between identifiable subgroups of students, e.g., between boys and girls, between African American and European American students, and/or between students from different economic backgrounds. And, eventually, we sought answers as to why the reform worked in some schools and not in others and what components were replicable across sites.

Our first challenge was to develop a research design that allowed us to obtain reliable data across a variety of schools over time. We needed to assess multiple components in a complex system and to compare responses and achievement scores from cohorts of students, teachers, and principals. Therefore, a nested, three-tier design, shown in Figure 1, was used. In 1994, 150 schools were randomly selected to participate in an assessment of Ohio's SSI, and in each of the following five years over 100 participated. At these schools, designated in Figure 1 as Level A, principals and all mathematics and science teachers (for grades six through nine) completed questionnaires focusing on standards-based teaching of, parental involvement with, and administrative support for science and mathematics education.

Level B consisted of a subset of the original random sample of schools. Across the years, the number of schools agreeing to participate in Level B ranged from 12 to 16. Level B schools were selected using specific demographic factors that would enable us to assess changes in teaching and learning among Ohio's high-risk students.[3] The following criteria were used to select Level B schools: they were part of the statewide random sample; they enrolled approximately 30% African American students; they had at least one teacher who had participated in the SSI's professional development programs; and they had high proportions of their students eligible for free or reduced-price lunch. At Level B schools, students completed a questionnaire that included items parallel to those on the teacher and principal questionnaires as well as a science or a mathematics achievement test that was designed to measure students' problem solving abilities and conceptual understanding. In addition, from 1994 through 1999, three or four day site visits were made to Level B schools.

Our first challenge was to develop a research design that allowed us to obtain reliable data across a variety of schools over time.

Level C involved intensive, multiple year case studies of five schools, spread across Ohio and representing urban centers, small towns, urban fringe, and suburbia. At this level, extensive school and classroom observations were conducted. Students, teachers, administrators, parents, and community leaders were interviewed. Observations and interviews were made serially and contingently in the sense that decisions about whom to involve and how to involve them were dependent upon what had been learned at other levels of the study and at other schools in Level C.

We found that the nested research design produced fairly quick information at the survey (state) level to guide Ohio's continued reform, and at the school and district levels it provided ways (observation and interviews) to validate the survey data. Further, the case studies elucidated how systemic reform affected schools at different stages of readiness.

Other major issues faced were: using self-report data, blurring of the distinction between SSI and non-SSI teachers as the reform

progressed, comparing scores for cohorts of students across a five year time span, and collecting meaningful achievement data in an economical way. Indeed, one lesson learned is that the quality of the data must be weighed against the cost of the data.

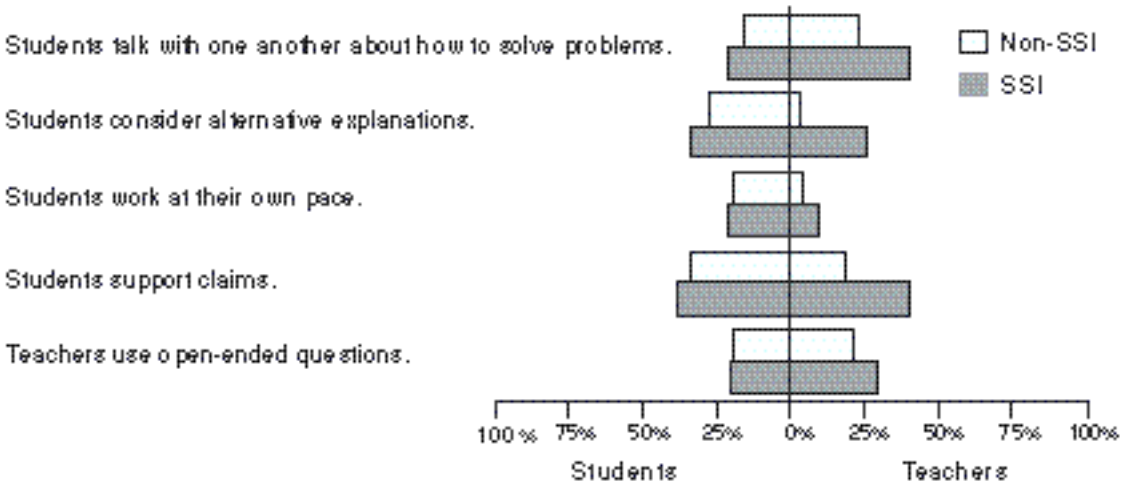
In order to evaluate the impact of the SSI's professional development on teaching practices a set of articulated questionnaires for principals, teachers, and students were developed. At the classroom level, we hoped to diminish any self-reporting bias by having teachers and students respond to similar items. For example, as shown in Figure 2, both groups were asked to estimate the frequency that teachers used open-ended questions on a five-point Likert scale that ranged from *Very Often* to *Almost Never*. By comparing the percent of responses to the left and right of the center (zero) bar, one can see how well student responses supported those of their teachers.

Inquiry Tests were developed by two task forces composed of university mathematics or science faculty, members of Ohio's SSI academic leadership teams, and other Ohio teachers. Each task force identified eighth grade public release items from the National Assessment of Educational Progress' (NAEP) 1990 and 1992 tests. Items were selected specifically to measure student ability in problem solving and conceptual understanding. Beginning in 1995, students of SSI and non-SSI "match" teachers took the tests.[4] Although the SSI focused on grades six through nine, only seventh and eighth graders were tested because of the age appropriateness of the test. Using the Cronbach Alpha Test of Internal Consistency, reliabilities were established at .86 in mathematics and .94 in science. Achievement data collected in 1995 served as base-line data, and all test scores in subsequent years were calibrated on the 1995 scale.

One of the most perplexing issues faced in the evaluation of systemic reform was the gradual loss of any control group.

A basic methodological issue faced by all long-term reforms is valid comparisons across different student cohorts in different educational settings. Item Response Theory (IRT) was used to address that issue. For example, IRT was used to refine and revise the achievement tests. Because cohorts of students were assessed, it was important to develop a bank of items, called anchor items, that were common across the years. These items, which anchored the test from year to year, were continually monitored in three ways: (1) evaluation of misfit statistics, (2) analysis of differential item functioning (DIF), and (3) consideration of external issues that might cause items to drift. Items showing any of those characteristics were removed as anchor items. Anchor items allowed us to compare responses on questionnaires and student test scores across the years, although the questionnaires and tests were modified as the reform progressed in order to retain sensitivity to its changing nature (e.g., implementation of the

Figure 2. Students and teachers responding "Very Often" to use effective classroom practices in mathematics and science



The quality of data had to be carefully weighed against the cost of obtaining the data.

Ohio Proficiency Test in science, wider use of standards-based teaching, wider dissemination of the state model curriculum, better alignment of state and district policies). (For a detailed description of instrument development and revision, please see Scantlebury, Boone, Kahle, & Fraser, 2001.)

Although carefully designed paper and pencil achievement tests provided a very useful measure of student learning, we were sensitive to the limitations of using only one measure. In 1998, we explored the use of performance assessments by implementing performance tasks from the Third International Mathematics and Science Study (TIMSS) in selected schools. In addition, multiple-choice versions of selected TIMSS' tasks were added to the Inquiry Test in science. Analysis of the data suggested that paper and pencil tasks alone inadequately measured student learning, particularly for urban, African American students (Harmon, 1991; Kelly, 2001). However, expense as well as difficulties in both delivery and scoring caused us to drop performance testing from the evaluation.

One of the most perplexing issues faced in the evaluation of systemic reform was the gradual loss of any control group. If a reform is systemic (and working), participants infect their colleagues with their enthusiasm and ideas. Our review of approximately 90 teacher portfolios as well as our monitoring of the SSI's electronic "teacher lounge" suggested wide sharing of inquiry-based lessons, alternative assessments, and other teaching materials. Because there was clear blurring of the two groups (SSI and non-SSI teachers) in some schools, we moved to comparisons that involved the percent of SSI teachers in a school.

Ohio's systemic initiative was based on equity, and our evaluation focused on equity issues, particularly at Level C. We used the *Equity Metric* (Kahle, 1998) as one way to interpret findings both within one site and across sites. It proved to be an effective model for analyzing why the reform works in some schools or districts and not in others (Hewson, Kahle, Scantlebury, & Davies, 2001). Further, it helped to elucidate what aspects of the systemic reforms could be replicated across sites (Kahle and Kelly, 2001).

We were sensitive to the issue of causality, because we could not directly relate outcomes to treatment (the SSI's professional development institutes and follow-up activities). As a consequence, we used multiple sources of data and looked for similar trends. Although attribution could not be established, common trends suggested more than a chance phenomenon.

What Did We Learn?

> The underlying assumption … is that systemic reform is a proven strategy and that we know how to do it, and therefore the only important question is "are they doing it right?" (Corcoran, 1997).

As the above quote suggests, when NSF initiated the SSI program, systemic reform was not a proven strategy for improving science and mathematics education. Therefore, questions concerning efficacy, efficiency, and effectiveness needed to be addressed by evaluations. Yet, those evaluations faced many unresolved issues in research and design as well

Figure 3. Comparison of seventh and eighth grade mathematics scores of students taught by SSI teachers

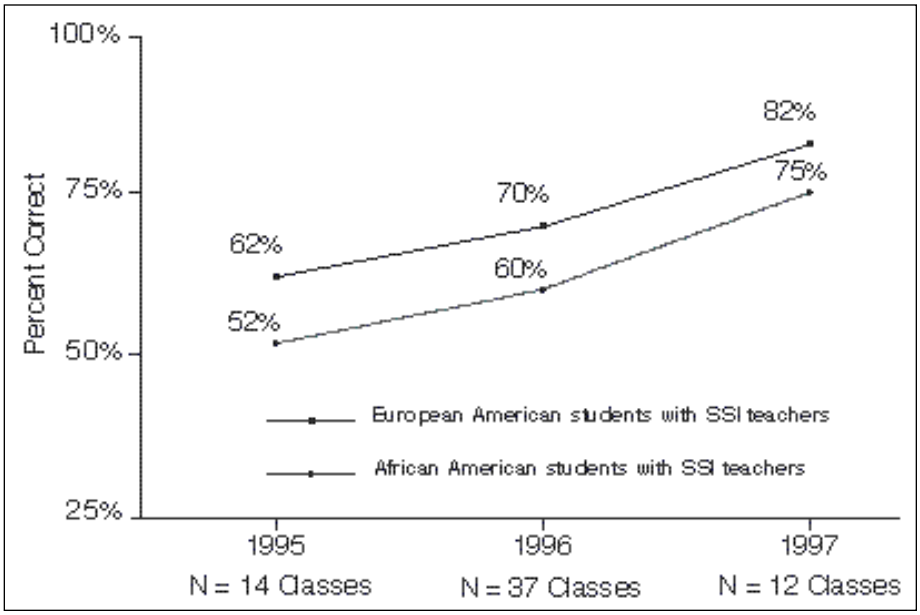
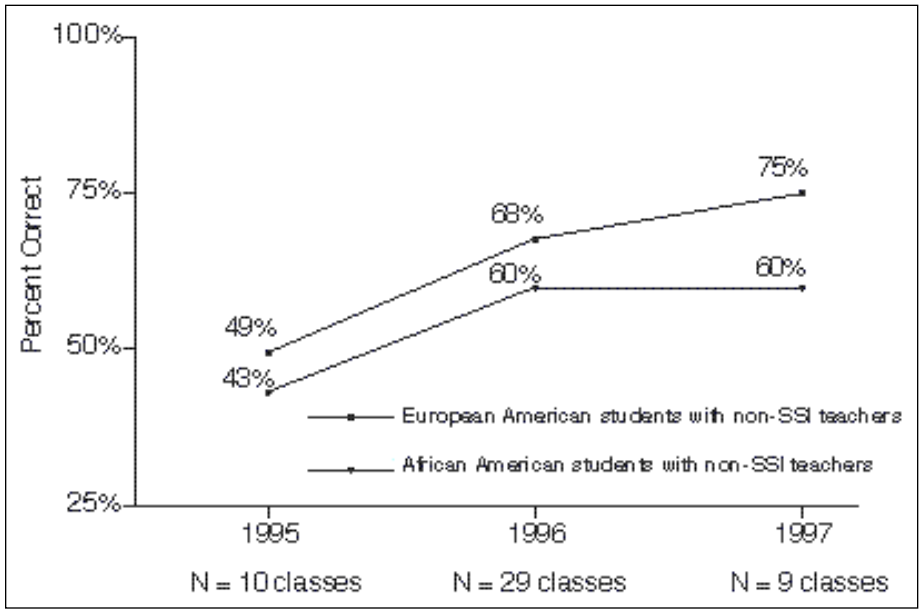


Figure 4. Comparison of seventh and eighth grade mathematics scores of students taught by non-SSI teachers



as in analyses and interpretation. For example, simply obtaining reliable and valid data across a state, particularly in Ohio's largest urban districts, was complex. The quality of data had to be carefully weighed against the cost of obtaining the data. Therefore, data on teaching practices, student attitudes, and principal support of science and mathematics education were collected by questionnaires. However, those data were confirmed by visits to Level B schools. In addition, because data collected with questionnaires are subject to self-report bias, principals, teachers, and students responded to similar items. Responses of different groups to the same items could be compared, increasing confidence in the findings.

To address the issue of causality, we analyzed achievement data in multiple ways. In one analysis, data were examined by student racial group; two independent studies looked at gender differences, while a third analysis assessed possible bias related to teacher characteristics. First, a comparison of 610 science students in matched science classes indicated that both African American girls and boys in classes taught by SSI teachers scored 9% higher on the science test than did their peers in matched classes. In addition, European American girls in SSI classes scored 10% higher and European American boys scored 5% higher than their peers in non-SSI classes (Damnjanovic, 1998). Achievement also was analyzed at the class level using only classes that had at least 25% of their students in a minority group (either 25% African American or 25% European American students). Although many classes did not fit that profile, we had a representative sample (comparable numbers of classes taught by SSI and non-SSI teachers) for three years in mathematics. One hundred and eight classes, enrolling over 3000 students, in ten schools were involved. As Figure 3 shows, the achievement gap in mathematics in classes taught by SSI teachers narrowed from 10.4 percentage points in 1995 to 7.5 in 1997. On the other hand, according to Figure 4, it widened from 7.3 percentage points in 1995 to 15.1 in 1997 in classes whose teachers had not participated in the SSI's professional development.

To address the issue of causality, we analyzed achievement data in multiple ways.

In addition, two independent analyses established that gender gaps in both mathematics and science decreased both across and within racial groups (Damnjanovic, 1998; Goodell, 1998). Another analysis compared the predicted scores of students whose teachers had completed the SSI professional development to those of students whose teachers had applied to participate but had not yet done so. That is, all teachers were volunteers.[5] The positive effect of the SSI's professional development was suggested by higher scores (from 2% to 7%) on both the mathematics and science tests of students (N = 2374) whose teachers had completed SSI's sustained professional development, compared to those who had not (Supovitz, 1996). Because all teachers were volunteers, this analysis controlled for the "volunteer" effect.

Other analyses examined the impact of the number of SSI teachers in a school or district. In 1998 and 1999, we were able to obtain Ohio Proficiency Test (OPT) mean scores in science and mathematics for eighth grade students in several large urban districts. Schools were clustered, depending upon the percentage

of SSI teachers in their faculty. In total, nearly 12,000 students were involved. The percent of minority students in the schools ranged from 77 to 79%, while between 67 and 71% of students in the schools qualified for free or reduced-price lunch. As shown in Figure 5, if over 51% of the teachers in a school had participated in the SSI's professional development activities, it was designated as *High* participation. Clearly, the percent of teachers involved in the professional development was a factor in students passing the science proficiency test. One explanation is that students in schools with *High* percentages of SSI teachers were more likely than students in *Medium* or *Low* participating schools to have had several SSI teachers. That explanation was verified by the class and school visitations that occurred at Levels B and C.

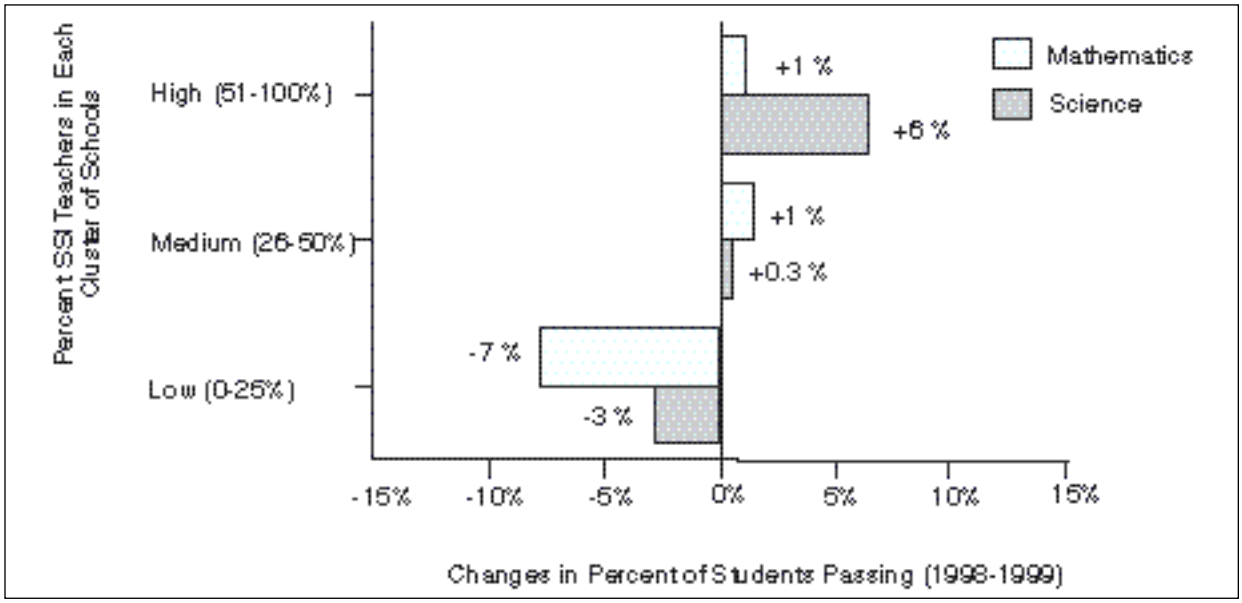
Clearly, the percent of teachers involved in the professional development was a factor in students passing the science proficiency test.

Another set of analyses involved Hierarchical Linear Modeling (HLM) procedures; they were used to more closely examine teacher influences on students' achievement and attitudes in science. HLM is a relatively new statistical technique that can examine variables measured at different levels (individual, teacher, and school), as well as variables measured on different scales (e.g., categorical and continuous variables). It is considered the most appropriate procedure for dealing with hierarchical data structures in which individuals are nested within other organizational contexts, such as instructional groups, classrooms, or schools (Bryk & Raudenbush, 1992). Because of the SSI's focus on teacher professional development and equity, our analyses were directed toward explaining variations in African-American students' achievement and attitude scores in relation to teacher differences. We found that approximately 16% of the variance in achievement scores on the Inquiry Test in science and 10% of the variation in attitude scores on the student questionnaire could be attributed to between-teacher differences (Kahle, Meece, & Scantlebury, 2000).

What Lessons Were Learned?

Our efforts to evaluate the outcomes of one state's systemic reform elucidated several issues critical to most large-scale efforts in education. Our success (or failure) in addressing those issues produced lessons learned that are applicable to similar efforts. The first lesson learned was that a nested, multi-layer research design enabled us to collect appropriate data at several

Figure 5. Changes in percent of urban students passing the Ohio Proficiency Tests in mathematics and science by percent of SSI teachers in school



levels (student, teacher, classroom, and school). Further, it allowed us to balance relatively inexpensive data collection techniques (survey) with intensive (and expensive) ones like the case studies.

The initial lesson learned was that the evaluation design had to address all parts of the system and that it needed to include various types of research techniques.

How to responsibly address causality also was an important lesson learned. Because multiple factors and at least ten years are involved in any systemic reform, it is impossible to attribute change to any one factor or condition. To address the thorny issue of causality, especially in reporting student achievement data, we performed multiple analyses, using different controls and techniques, and we looked for patterns of change. A third lesson learned was the value of statistical techniques (IRT) that allowed us to conduct the evaluation across many years and in many sites using cohorts of students, rather than a longitudinal sample. We experimented with performance items only to find that the value added was not equal to the costs incurred—a fourth lesson learned. And, the fifth lesson was the value of interpreting quantitative data through the lens of qualitative data. That lesson is particularly pertinent today when one considers reports emanating from the quantitative data required by the NCLB legislation. Recently we learned that only 52 schools in the country have been identified as dangerous. Los Angeles, Chicago, Miami, Detroit, Cleveland, San Diego, Baltimore, and Washington, D.C. have no violent schools; and New York City only has two (Schouten & Toppo, 2003). Clearly, the criteria used to identify a dangerous school varied in 52 different state surveys, and in no state were the quantitative findings verified by qualitative observations.

The initial lesson learned was that the evaluation design had to address all parts of the system and that it needed to include various types of research techniques.

Due to the evaluation's findings of positive outcomes, the Ohio reform of science and mathematics education has been maintained with state funding. The reform continues to face new situations and issues due to changing conditions in the state (the Ninth Grade Ohio Proficiency Tests are being replaced by Tenth Grade Ohio Graduation Tests, and new science and mathematics standards have been adopted). However, both the issues faced and the lessons learned continue to inform the reform as well as its evaluation.

References

Bryk, A. S., & Raudenbush, S. (1992). *Hierarchical Linear Model*. Thousand Oaks, CA: Sage Publishing.

Corcoran, T. B. (1997). The role of evaluation in systemic reform. In W. H. Clune, S. B. Millar, S. A. Raizen, N. L. Webb, D.C. Bowcock, E. D. Britton, R. L. Gunter, & R. Mesquita (Eds.), *Research on systemic reform: What have we learned? What do we need to know? Synthesis of the second annual NISE forum, Volume 1: Analysis* (Workshop Report No. 4, pp. 64-68). University of Wisconsin-Madison, National Institute for Science Education.

Damnjanovic, A. (1998). Ohio Statewide Systemic Initiative (SSI) factors associated with urban middle school science achievement: Differences by student sex and race. *Journal of Women and Minorities in Science and Engineering, 4*, 217-233.

Goodell, J. E. (1998). *Equity and reform in mathematics education*. Unpublished doctoral dissertation, Curtin University of Technology, Perth, Western Australia.

Harmon, M. E. (1991). Fairness in testing: Are science education assessments biased? In G. Kulm & S. M. Malcolm (Eds.), *Science assessment in the service of reform* (pp. 31-54). Washington, DC: American Association for the Advancement of Science.

Hewson, P., & Kahle, J. B., Scantlebury, K., & Davies, D. (2001). Equitable science education in urban middle schools: Do reform efforts make a difference? *Journal of Research in Science Teaching 38*, 1130-44.

Kahle, J. B. (1998). Equitable systemic reform in science and mathematics: Assessing progress. *Journal of Women and Minorities in Science and Engineering, 4*, 91-112.

Kahle, J. B., & Kelly, M. K. (2001). Equity in reform: Case studies of five middle schools involved in systemic reform. *Journal of Women and Minorities in Science and Engineering, 7*, 79-96.

Kahle, J.B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching, 37*, 1019-1041.

Kelly, M. K. (2001). Moving toward equitable, systemic science education reform: The synergy among science education and school-level reforms in an urban middle school. *Dissertation Abstract International 62* (06), 2017A.

McDermott, L. C., Shaffer, P. S., & Rosenquist, M. L. (1996). *Physics by inquiry*. New York, NY: John Wiley and Sons, Inc.

Paige, R., Hickok, E., & Neuman, S. B. (2000, September). *No Child Left Behind: A desktop reference*. Jessup, MD: Education Publications Center, U.S. Department of Education.

Scantlebury, K., Boone, W., Kahle, J. B., & Fraser, B. J. (2001). Design, validation and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching*, *38*, 646-662.

Schouten, F., & Toppo, G. (2003, September). Many schools turn heads to danger. Albany, NY: *Times Union*. Retrieved September 22, 2003 from <http://www.timesunion.com/AspStories/story.asp?storyID=171851>.

Shavelson, R.J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Supovitz, J. (1996, February). *Ohio's Project Discovery 1995 Discovery test student results*. Chapel Hill, NC: Horizon Research, Inc.

Endnotes

1. A member of the first cohort of SSI states, all of Ohio's eligible cities (Columbus, Cleveland, and Cincinnati) received USI awards, and five Ohio counties were part of the Appalachian RSI.
2. The initial institutes were based on the University of Washington's *Physics by Inquiry* curriculum (McDermott, Shaffer & Rosenquist, 1996). Later, content-based, inquiry courses were developed in mathematics and life science.
3. Ohio's public school population is a little more than 1.8 million, with African Americans constituting its largest racial/ethnic group (17%). Over a half million students are eligible to receive free or reduced-price lunch.
4. A "Match" teacher taught similar classes as an SSI teacher in the same school; that is, ninth grade, general biology teachers who had and who had not participated in the SSI's professional development would be "matched," by experience, gender (if possible), and type of license.
5. Evaluations of the systemic initiatives have debated the "volunteer effect;" that is, the difficulty of reaching beyond the teachers who volunteer for professional development. The concern is that "volunteer" teachers, as a group, may differ substantively from the "non-volunteer" teacher group.

Jane Butler Kahle is Condit Professor of Science Education, Miami University, Oxford, OH 45056. E-mail: <kahlejb@muohio.edu>.