The Need for Comprehensive Evaluation in Science Education

DAVID D. KUMAR
Florida Atlantic University

JAMES W. ALTSCHULD
The Ohio State University

Abstract
This article argues for policies regarding the implementation of evaluation as an integral part of science education reform. Approaches to evaluation should be made comprehensive enough to include multiple methods and, through their use, to gain in-depth information on large-scale science education programs and projects. There must be a sustained commitment from stakeholders such as the National Science Foundation to support the training of specialists in the evaluation of science education. Specialists who are able to arrive at independent conclusions that are meaningful and beneficial for science teaching and learning are needed. Evaluation recommendations posited by the authors should be put through careful analysis and feasibility testing before moving to widespread adoption.

INTRODUCTION
There has been an outpouring of reform efforts that claim to improve science education in the US, notably since the Sputnik era. However, the state of science education in the US is not encouraging. Millions of federal tax dollars spent over the decades have hardly produced any significant improvement in public education. While most such efforts receive support from federal sources, mechanisms in place to fully evaluate the impact of them in public education are sparse.

In fact, very little attention is paid to systematically evaluating large-scale tax-funded programs in science education and to determine why they fail to produce improvement in science teaching and learning in American schools. Instead, “criticizing education has become a political tactic” (Kliszuz, 2002, p. 167). Most public attention is misguided in the form of scathing attacks on public schools and colleges of education, and the condition of science education remains unchanged.

According to a national survey (Weiss, Banilower, McMahon, & Smith, 2001), the condition of science education in K-12 is not encouraging. Pertinent findings of the survey follow. On an average, 23 minutes per day is spent teaching science in K-3, and 31 minutes in fourth- to sixth-grade levels. The survey also found that only 18% of elementary teachers feel very well qualified to teach physical science, and 25% to teach earth science. About half the teachers in K-4 lack preparation in science and science education. Additionally, 33% of physical science teachers and 60% of earth science teachers in grades 7–12, and 36% of physics teachers and 42% of earth science teachers in grades 9–12 lack sufficient preparation in
the respective sciences. After all the bells and whistles about the national science standards, and millions of tax dollars spent on their development, and on the accompanying propaganda and implementation, about 63% of teachers in grades 9–12 are “not at all familiar” with the standards.

This finding was supported in a somewhat different and perhaps oblique way in a 1997 set of case studies carried out by Cullen and colleagues (Cullen et al., 1997). They identified, through a national nomination process, four public school sites that were cited for exemplary work in revamping their science education curricula and instructional processes. In onsite interviews the research team observed that the overall nature of what was transpiring at the district level was almost divorced from the classroom, where change realistically has to occur especially if impact on teaching methods and student achievement and understanding is desired. Teachers were unfamiliar with or had only “passing knowledge” of what was happening at the district level and were almost uninvolved in the new initiatives. This became apparent to the researchers during interviews when they had to dramatically alter interview protocols not only conceptually but also in regard to even the specific terms and words that were used for the teacher interviewees. The teachers simply did not resonate with the changes that were supposedly being pushed by the district. Thus, the new initiatives were very questionable regarding their meaning and lasting effects on the delivery of science education in the four exemplars. And it is to be stressed that the exemplars were selected through an intensive cross-nomination process as being representative of the best such endeavors in the US.

Similarly, Anderson (2002) reported that the evaluation of various new programs in science education funded by The National Science Foundation (NSF) have been characterized by poor implementation and have not led to substantial improvement in instruction and achievement. Further she stated that even after five years of engagement and investment in the new programs, not one single school district in the county had been able to “go-to-scale”, that is, to achieve a full-scale implementation of their science education reform efforts.

Often, science teaching remains disconnected from real-world applications. For example, 81% of the emphasis is on learning basic scientific concepts, but learning the science, technology, and society relationship, and business-related applications of science, are emphasized less than 30% in grades 9–12 (Weiss et al., 2001). Furthermore, despite the recent surge in technology applications in education, surprisingly, only 6% of the students in grades K–4, 11% in grades 5–8, and 16% in grades 9–12 “use computers as a tool” in science classrooms. These findings are startling and, along with the writings of other authors given above, they point to the complexity of dealing with change in the social milieu of multifaceted educational settings and systems.

The startling results of Weiss and colleagues (Weiss et al., 2001) also raise serious questions. What are the impediments to improving science
education in American schools? Have any of the federally funded large-scale projects such as the Statewide Systemic Initiatives had any impact on science education? Are adequate evaluation measures available and are they being used to monitor and measure the impact of large-scale tax-funded science education projects? Furthermore, how should we evaluate schools—what are the key variables that should we focus on? How does information collected at what is usually a high level of the system, filter down to the classroom, and with what effect on practice? What have we learned from the evaluation of reforms that will guide us for future efforts and their evaluations? Do we have a cadre of trained and experienced evaluators who are up to the evaluation task and who can move beyond the facade of test scores to help us see what is really taking place in science education?

According to Anderson “it is the role of evaluation to provide credible evidence to inform funding decisions and to generate knowledge that will advance the practice of restructuring the K-12 education systems” (p. 78). When considering large-scale systemic initiatives, the role of evaluation becomes even more critical. As Chubin (cited in Anderson, 2002), said “measuring systemic reform must entail more than measuring the performance of the system’s components. Thus, the implication is that just as systemic change demands new strategies, systemic evaluation demands new metrics and innovative methods of assessing system capacity, infrastructure and performance” (p. 78).

SOME OBSERVATIONS ABOUT EVALUATION IN SCIENCE EDUCATION

In a recent publication, “What Does the Future Have in Store for the Evaluation of Science and Technology Education,” Altschuld and Kumar (2002) analyzed the chapters of a book they edited on the evaluation of science and technology education. The authors arrived at a number of conclusions relevant to exploring how to evaluate science education programs. First, there is a pressing need for in-depth research on evaluation in science and technology education. For example, we know more about TIMSS (Third International Mathematics and Science Study) than we do about the goings-on in science classrooms and in science education programs. It is not just knowledge but how the process of scientific thinking is transmitted to our students, how they become inculcated into the scientific method, and how the excitement and joy of science is transferred to them.

Second, there is an emerging demand for more understanding of the context of a program, in addition to the outcomes of a program. If a context is not supportive of change, if policies are not there to foster and reinforce change, if resources in the form of time and training besides finances are not provided, if the environment does not afford the opportunity to try out ideas and to learn from failures, and if other aspects of
a conducive, open atmosphere are not present, the probability of institutionalizing successful new programs will be extremely low. (Surprisingly, a number of these points were explicated by Bhola more than thirty-five years ago.) Following this line of logic, we might even have to go so far, in a very pessimistic sense, to axiomatically conclude that the culture of educational systems is not change-oriented and that education will limp along and advance in small, disjointed, incremental steps. If this is true we will continue to be dismayed by the inexorably slow pace of progress.

A third conclusion is that comprehensive approaches to evaluation of science education programs are lacking, instead, disparate studies are used to draw evaluative conclusions (Kumar & Altschuld, 2002). We have not yet done thorough meta-analytic studies across our evaluations to generate sound principles for developing science and technology education programs and analogously, in turn, for evaluating them.

A fourth conclusion deals with the fact that some areas have barely received any attention by evaluators. As an illustration, Cannon (2002) indicates that there were approximately 17,000 online courses that students could take as of the year 2000. Yet what do we know about these courses? What evaluation evidence is available other than possibly the reduced cost of delivery and the increased time demands of instructors? Do we have data and information about how students process the concepts contained in them? To what extent are the knowledge and concepts fostered in these programs learned and retained? Do evaluation techniques borrowed from regular classroom situations fit with interactive and other computer (Internet and web-based) capabilities, and so forth.

Even a seemingly simple evaluation task related to science instruction via technology may not be so simple. Altschuld (1994) questioned the placement of imbedded quizzes and questions in Computer Assisted Instruction (CAI) in science education. Do we understand learning enough to know when and where to probe for developing understanding, and when and where to provide feedback so that it is attended to and perceived as useful? Could premature testing actually interfere with learning, particularly for those students who slowly construct and generate mental models and images of what they are studying? These slower more cogitating learners may, indeed, be interrupted and potentially harmed by incorrect placement of test items even if they are included for formative purposes.

Among a host of other problems that Cannon (2002) suggested is that of shovelware. Shovelware would be best thought of as just taking what an instructor routinely does in his or her classroom and transferring or putting it on the web or Internet without much in the way of modification. In other words, shoveling over what is done in traditional instruction without much extra thought or focus on how to take advantage of the technology. Overall, it seems that the evaluation of science education
as delivered through burgeoning technological capabilities is to a large extent missing in current evaluation thinking.

The fifth conclusion relates to the training of evaluators for science education and the creation of what is called evaluation capacity. In 1995, Altschuld and Kumar lamented the rather limited application of evaluation models to science education programs. Moreover, their perception at that time was that the overall literature base of evaluation in the area was not abundant. In the mid-1990s, recognizing these deficiencies as well as others, the NSF funded four universities to increase the ranks of individuals skilled in the evaluations of science and mathematics education programs. The intent was to recruit, if possible, people with a background in science and mathematics (or science and mathematics education) and then to prepare them for the subtleties of evaluation in the two disciplines. This proved to be difficult for a number of reasons, but most prominently because few evaluation positions are solely in mathematics and science education. The training programs have now been phased out with what would be thought of as moderate success at best.

Now the NSF has shifted its focus to the idea of building evaluation capacity in educational endeavors. So many of its prior projects have not been evaluated well or have placed evaluation at such a low priority that the agency has limited evidence of their impacts and effects. Many project directors who come from a discipline-based science have not had much exposure to program evaluation or view it as useful only for obtaining outcome data. Given this situation, the NSF recently has funded a well-known evaluation contractor to develop a web site containing brief illustrative cases of good evaluations to guide project staff. It also has funded a small number of universities to play a role in enhancing or building a much stronger capacity for evaluating programs and projects.

It is obvious that comprehensive evaluations must be undertaken for large-scale projects, and such evaluations could collectively lead to major policy implications. In the next section we will describe several such evaluations and then derive implications of them in relation to science education policy.

COMPREHENSIVE EVALUATION

These evaluations come from three distinct evaluations of the NSF-funded interactive media science teacher education project, “Improving Science Education: A Collaborative Approach to the Preparation of Elementary School Teachers.” The aim of the project, which was conducted under the auspices of Vanderbilt University, was to develop video-based cases involving effective and ineffective teaching strategies suitable for use in the science (and mathematics) methods courses taught at the university. It involved science faculty from the college of arts and sciences, science and math educators from the college of education, and consultant
teachers from grades 4 to 7 in the local schools. (For specific details of the interactive media project see Barron et al., 1993).

Three different groups of evaluators used distinctly unique evaluation methods and perspectives to study the project. They included: a traditional evaluation by the project staff, another traditional but not the same evaluation by the Office of Technology Assessment (U.S. Congress, 1995), and a somewhat nontraditional context-based evaluation by Kumar and Altschuld (1999). The availability of three evaluations presents an unusual opportunity to closely scrutinize the degree to which they arrived at similar or disparate conclusions and to see the extent to which they were complementary in nature, and, finally, to examine the richness of information they offer for gaining a comprehensive picture of the Vanderbilt interactive media project. Such information is conducive to in-depth project analysis and informed decision-making, which is central to developing effective policies for science education.

The three evaluations are presented in Table 1 and their outcomes are in Table 2 with associated implications for policy.

### A Traditional Evaluation Conducted by Internal Staff

The project staff at Vanderbilt evaluated the interactive video project through traditional evaluation techniques (Table 1). Summarized highlights of their work will be given here. The sample included students enrolled in science methods courses with a practicum component, and students in the elementary student teaching experience approximately one or two semesters after the methods course. Students enrolled in the science methods course before implementing the interactive videos formed a baseline group and individuals in the science methods course that used the interactive videodiscs formed the video group. Classroom observation ratings (focused on Teaching Competency and Student

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Onsite observations</td>
<td>Review of project publications and software</td>
<td>Document review Observations onsite and in schools</td>
</tr>
<tr>
<td>Follow-up interviews</td>
<td>Observation of science methods class</td>
<td>Semistructured interviews</td>
</tr>
<tr>
<td>Tests</td>
<td>Interview of project investigators</td>
<td></td>
</tr>
</tbody>
</table>
Behavior) of students in the teaching practicum and student teaching groups, post-observation interviews, and pre- and post-tests were used to gather data. Course examination scores and the percentage of class time devoted to various instructional activities were part of the data. Descriptive and inferential statistical procedures were used to analyze data.

The findings can be classified into teaching competencies, student behavior, and activities (Table 2). Examples of significant outcomes under teaching competencies were “selecting materials/learning experiences which stimulate student curiosity and support their investigation,” “appropriate sequencing of content and pedagogy,” and “monitoring understanding.” In the area of student behavior, “student involvement in lesson,” “students’ interest in lesson” and in other activities, “student activities (discovery or inquiry),” and “individual seatwork” were found to be some of the significant effects of the interactive media project.

**Traditional Evaluation Conducted by an External Group**

As part of a periodic effort to evaluate federally funded projects, the Office of Technology Assessment (OTA) (U.S. Congress, 1995) evaluated the

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant gain in teaching competency</td>
<td>Strong theoretical foundation</td>
<td>Strong administrative support</td>
</tr>
<tr>
<td>Significant gain in student behavior</td>
<td>Strong research and development efforts by faculty</td>
<td>Permeation into science departments</td>
</tr>
<tr>
<td>Significant gain in other activities (e.g., individual seatwork)</td>
<td>Access to a “rich array of resources” by faculty</td>
<td>Student perception on campus</td>
</tr>
<tr>
<td></td>
<td>Availability of funds from various sources</td>
<td>Student perception in schools</td>
</tr>
<tr>
<td></td>
<td>Faculty incentives to get involved</td>
<td>Strong technical support</td>
</tr>
<tr>
<td></td>
<td>“Expensive infrastructure”</td>
<td>Cross-fertilization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conducive organizational climate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Critical mass of interest in the project</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Video vs. real classroom experience</td>
</tr>
</tbody>
</table>
same project along with several others. The overall effort included various forms of surveys. In the case of the Vanderbilt project, the evaluation consisted of a review of project publications and software, observation of science methods classes, and interviews of project investigators (Table 1). See *Teachers and Technology: Making the Connection* (U.S. Congress, 1995) for more details.

Results of the OTA evaluation (Table 2) indicated that the project could be characterized as having a solid theoretical foundation in “Anchored Instruction,” showing strong evidence of research and development efforts by faculty, and possessing access to a “rich array of resources” and availability of funds from various sources. The findings also raised concerns about the adaptability of the project to other diverse institutional settings.

**Context Evaluation by an External Group**

The external evaluation reported by Kumar and Altschuld (1999) involved the use of an alternative conceptualization and approach predicated on a context evaluation model (Altschuld & Kumar, 1995). The model builds upon the lifecycle of a product or innovation from initial idea through development, tryout, and eventual implementation. The model then focuses on every part of the life cycle as embedded in a complex milieu of environmental (e.g., social, economic, and organizational) factors. Evaluations following the premises of the model would stress organizational environment, evidence that the environment was conducive to change and experimentation, the provision of human, financial, and technical resources to enable users to adopt new innovations, administrative commitment to promote programs, and other similar factors. If these factors are compared to the previously described evaluations, it should be apparent that they are somewhat orthogonal to what was looked at in more traditional ways. The evaluations are being conducted from appreciably different worldviews. They looked at different types of variables and a host of different methods were employed for data collection, analysis, and interpretation.

Furthermore, a critical feature of the evaluation that Kumar and Altschuld described was to not just evaluate the context of the Vanderbilt project, but to do so in a period of time after its products had been developed and implemented. Often, this type of follow-up evaluation, though essential to long-term use, is not part of most evaluations. It should be noted that the project ended approximately two years before this evaluation took place, and thus the evaluation carefully investigated the interaction of the project and its environment, and the press for or against continued implementation. The evaluation included document review, development of interview protocols, sample selection, onsite observations, interviews (Table 1), and analyses of interview data into initial data categories and emerging explanatory themes.
The evaluation produced an array of findings (Table 2). Examples include “strong administrative support,” “student perception on campus,” “student perception in schools,” “osmosis/permeation” across units and disciplines at the university, “cross-fertilization” of ideas, and the notion that a “critical mass” of individuals is often necessary for change. The evaluation also revealed a gulf in the student experience between learning from the videos about hands-on science and actually doing hands-on science in real classrooms.

Clearly, one can see from examining the two tables that each evaluation helped to expand the understanding of what transpired and what the impact of the project was from both a short- and long-term view. On the other hand, if we did not have the three evaluations to compare and contrast, we would have a radically diminished and limited picture of what this project was about, how it was implemented, what it produced overtly and covertly, and, most importantly, we might develop substantially less-meaningful science education policies solely from one such evaluative study. This is why we chose to stress the concept of comprehensive, multifaceted evaluation as vital for policy considerations in the field. Without it we simply do not have enough reference points and enough insight for the policy job and the development of sound and thoughtful policies for the field.

With our assumptions and views now apparent, let us turn to the policies on evaluation that emerge from our vantage point from having done one of the three studies just portrayed. Some of the policy recommendations proposed would be costly so they should be considered only for large, major, and well-funded projects that have been carefully thought through and have sufficient time for implementation and getting rid of the inevitable “bugs” that all programs contain.

POLICY RECOMMENDATIONS

1. The NSF (and other concerned funding agencies) should continue investing in and encouraging what we would term comprehensive evaluations of major science education reform projects. Although the investment would be costly and time-consuming, it is a *sine qua non* condition. It is patently obvious that major reforms deal with the nuances and complexities of multiple subsystems within the larger overall system of education. One single model or method for evaluating will not suffice. Moreover, the unpredictable nature of the context and how it affects program implementation and success varies from site to site and will not be revealed through one evaluation lens or filter. The evaluations must use both qualitative and quantitative methodologies and the insights they collectively, rather than individually, produce. Lessons learned from such evaluations should be collated and studied via funded projects in order to better
inform change and reform efforts. To some extent this has already occurred (Anderson, 2002) but much more is needed. We would gain understandings of the critical variables that are key to successful programs and that should be emphasized in evaluation.

2. The NSF should continue to sponsor the training of evaluators in general as it now does (for example, the Evaluator’s Institute and the training offered through Western Michigan University, and the American Educational Research Association Grants Program) as well as the training of specialists in the evaluation of science education. Evaluating the development of scientific thinking in young minds requires knowledge of process, recognition of misperceptions about science principles, understanding of how data is interpreted, and other dimensions with which generalists in program evaluation often tend to be unfamiliar. Ideally individuals trained and experienced in science or science education would be recruited for additional graduate work in evaluation (qualitative and quantitative methods, general and specific models of evaluation, needs assessment, measurement, statistics, research strategies and techniques, program logic models, organizational development, change theory and so forth).

3. There is a huge caveat here and it is that unfortunately, ideals are not usually realized. It is suspected that the four training programs supported by the NSF in the late 1990s had limited to moderate impact on the need for evaluators of science education programs. One problem might be the evaluation field itself, that is, evaluators seldom will be able to sustain a career by specializing in one narrow area of the profession. So even if one receives extensive, quality preparation for the focused evaluation of science education, they probably will have to apply their skills to other types of programs and projects to survive. If this is true, and we feel that it probably is, then the very premise of focused training and the likelihood of attracting scientists and science educators to it comes into question. (Note that one of the authors who has worked for more than thirty years in evaluation is basing this observation on that work history, and the other author, who participated in evaluation through the AERA (American Educational Research Association) Grants Program, is basing this observation on that training experience.) Thus we move to our fourth policy recommendation.

4. The career pathways for science education evaluators must be carefully thought through or specialized training will quickly fall by the wayside. The NSF and other groups that care about science education and collecting systematic information about what does or does not work may have to create career opportunities that do not now exist to retain the specialized expertise. We suspect that this could
occur, given the total level of funding of science education in the country; there is enough work to begin to think about this issue in this manner. This might be done under the auspices of the NSF but it would have to be carried out with a view toward the independence of the evaluators. (See policy recommendation No. 5)

5. One route through which the above recommendation could be achieved might be by establishing an independent center devoted to the evaluation of science education programs in the US. It would have to be independent to take critical and incisive looks at the nature of what is happening in science education at K-12 levels and in postsecondary education. To preserve that independence and a sense of objectivity (by not being beholden to any one funding source or biased toward the viewpoint of a particular funder), it could be financed by a cross-section of agencies such as the National Science Foundation, the US Department of Education, the National Institutes of Health, and nonprofit groups with an interest in science education. Staffing would include general evaluators, specialists in science education evaluation, scientists, measurement experts, and others who would form vibrant, mixed teams to evaluate programs.

FINAL NOTE

In some respects the formation of policy is a rather simple undertaking. One reviews the literature, examines issues illuminated by it, thinks in depth about the consequences of the issues, and then proposes a framework that sets the direction for future activities. But that framework (those policies) must be of quality, with quality being eventually evident in terms of feasibility of implementation, positive changes, the satisfaction of those benefiting from the policies or providing services, and the like. What this means is that the policies need to undergo a serious dose of reality testing for them to be effective in the real world. The pros and cons of policy decisions have to be carefully analyzed and weighed before any new policy effort is promulgated and adopted.

And so must our ideas be subjected to rigorous reality testing. As an example, the fifth recommendation is one about which much is already known. There are various types of federally funded educational centers (some have existed for an extensive period of time) and their effectiveness and the strengths and weaknesses of their operations probably have been studied. Only after full examination of this information should the policy be given major consideration. All of our thoughts described above are offered with this point of view in mind. In this era of national standards and testing, and increased scrutiny of public education, science educators stand to gain from evaluation valuable insights for improving science teaching and learning. The time for implementation of a more
sustained and comprehensive evaluation system is long overdue in science education.

REFERENCES


ABOUT THE AUTHORS

David D. Kumar is Professor of Science Education at Florida Atlantic University. His research involves evaluation and policy in science and technology education. He is coeditor of the books *Evaluation of Science and Technology Education at the Dawn of a New Millennium* (New York: Kluwer Academic/Plenum Publishers, 2002), and *Science, Technology, and Society: A Sourcebook on Research and Practice* (New York: Kluwer Academic/Plenum Publishers, 2000). He served as a Guest Scholar in the Governmental Studies Program at The Brookings Institution during
the summer of 1996. He is a Fellow of the American Institute of Chemists, and an elected Vice President of the Policy Studies Organization.

James W. Altschuld is a Professor in the School of Educational Policy and Leadership at The Ohio State University. He specializes in educational research and program evaluation. His publications include three coauthored books (two on how needs in society and institutions are assessed and one on the evaluation of science and technology education) and many articles dealing with evaluation methods and issues. He has received international, national, state, and local awards for his work in evaluation and for teaching.