# Inquiry-Based Science in the Middle Grades: Assessment of Learning in Urban Systemic Reform

Ronald W. Marx,[1] Phyllis C. Blumenfeld,[2] Joseph S. Krajcik,[2] Barry Fishman,[2] Elliot Soloway,[2] Robert Geier,[2] Revital Tali Tal[3]

[1] *University of Arizona, 1430 East 2nd Street, P.O. Box 210069, Tucson, Arizona 85721*

[2] *University of Michigan, 610 East University, Ann Arbor, Michigan 48109*

[3] *Department of Education in Technology and Science, Technion, Israel Institute of Technology, Haifa 32000, Israel*

*Received 2 January 2003; Accepted 26 April 2004*

Abstract: Science education standards established by American Association for the Advancement of Science (AAAS) and the National Research Council (NRC) urge less emphasis on memorizing scientific facts and more emphasis on students investigating the everyday world and developing deep understanding from their inquiries. These approaches to instruction challenge teachers and students, particularly urban students who often have additional challenges related to poverty. We report data on student learning spanning 3 years from a science education reform collaboration with the Detroit Public Schools. Data were collected from nearly 8,000 students who participated in inquiry-based and technology-infused curriculum units that were collaboratively developed by district personnel and staff from the University of Michigan as part of a larger, district-wide systemic reform effort in science education. The results show statistically significant increases on curriculum-based test scores for each year of participation. Moreover, the strength of the effects grew over the years, as evidenced by increasing effect size estimates across the years. The findings indicate that students who historically are low achievers in science can succeed in standards-based, inquiry science when curriculum is carefully developed and aligned with professional development and district policies. Additional longitudinal research on the development of student understanding over multiple inquiry projects, the progress of teacher enactment over time, and the effect of changes in the policy and administrative environment would further contribute to the intellectual and practical tools necessary to implement meaningful standards-based systemic reform in science. © 2004 Wiley Periodicals, Inc. J Res Sci Teach 41: 1063–1080, 2004

Contract grant sponsor: Spencer Foundation; Contract grant sponsor: National Science Foundation; Contract grant numbers: REPP (REC-9720383 REC-9725927 REC-9876150), USI (ESR-9453665).

*Correspondence to*: R.W. Marx; E-mail: ronmarx@email.arizona.edu

DOI 10.1002/tea.20039

Published online 1 November 2004 in Wiley InterScience (www.interscience.wiley.com).

© 2004 Wiley Periodicals, Inc.

Science education standards set forth by the American Association for the Advancement of Science (1993) and the National Research Council (1996) urge less emphasis on memorizing decontextualized scientific facts and more emphasis on students investigating the everyday world and developing deep understanding from their inquiries. By emphasizing scientific inquiry, the standards challenge the education and science communities to transform the very heart of students' experiences in science classrooms. In support of the standards, new approaches to science instruction feature inquiry as essential for student learning. These approaches assume that students need to find solutions to real problems by asking and refining questions; designing and conducting investigations; gathering and analyzing information and data; making interpretations, creating explanations, and drawing conclusions; and reporting findings (Krajcik, Blumenfeld, Marx, & Soloway, 2000; Linn, Clark, & Slotta, 2003; Songer, Lee, & McDonald, 2002).

These new approaches to instruction present challenges to both teachers and students. For teachers using instructional methods based on recitation and direct instruction, inquiry teaching challenges them to develop new content knowledge, pedagogical techniques, approaches to assessment, and classroom management (Blumenfeld, Krajcik, Marx, & Soloway, 1994; Edelson, Gordin, & Pea, 1999; Marx, Blumenfeld, Krajcik, & Soloway, 1997). Students are equally challenged, as these innovations change how they interact in classrooms. Inquiry learning requires them to collaborate with peers, think deeply about complex concepts, relate new science content to their lives inside and outside school, and self-regulate their behavior and thinking across the weeks that an inquiry project might unfold (Blumenfeld, Soloway, Marx, Krajcik, Guzdial, & Palincsar, 1991; Krajcik, Blumenthal, Marx, Bass, Fredricks, & Soloway, 1998; Roth, 1995). For students who are already performing below grade level with respect to literacy and mathematics (as is the case in many urban settings), learning through inquiry can be even more daunting, especially when there are also discontinuities in scientific and cultural ways of knowing (Lee, 2002).

Our core belief is that the promise of science education reforms ought to be achievable by all students, as is clearly claimed by the AAAS's position presented in *Science for All Americans* (Rutherford & Ahlgren, 1990). A number of researchers have shown that in highly resourced settings, inquiry instruction in urban classrooms can be successful, when it includes materials that leverage the culturally relevant knowledge and beliefs held by students from diverse backgrounds (Bouillion & Gomez, 2001; Moje, Collazo, Carrillo, & Marx, 2001; Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001). The challenge remains, however, to move these successes from the design environments in which they were created to the wider rough and tumble of neighborhood schools where teachers can enact them successfully and students can benefit from the opportunity.

Over the past decade, systemic reform has been advanced as a comprehensive and systematic approach to school improvement (Fuhrman, 2001; Smith & O'Day, 1991) designed to increase student learning through careful programming and alignment of curriculum and instruction, assessment, and professional development. Systemic reform in science often takes place in large urban systems that present a particular set of challenges for educators and their partners in reform (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000). These challenges include, for example, low student achievement, high student mobility, and difficulty recruiting and retaining highly qualified teachers (Berends, Kirby, Naftel, & McKelvey, 2001; Hannaway & Kimball, 2001).

Assessments of the effects of these efforts often focus on teacher reports of their instruction and student performance on either nationally normed tests or mandated state assessments because the efforts are national and statewide and therefore are not tied to specific curriculum (Garet, Porter, Desimone, Birman, & Yoon, 2001; Kahle, Meece, & Scantlebury, 2000; Supovitz, Mahyer, & Kahle, 2000; Supovitz & Turner, 2000). Both forms of assessment are rather distant from the content of particular district or school programs (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002) and might lack the sensitivity to detect important learning. Evaluations using these assessments might fail to detect effects of the systemic reform effort because of a failure of alignment of curriculum and assessment (Porter & Smithson, 2001). This problem of alignment of curriculum and instruction with assessment is present in many studies that evaluate systemic reform programs that span many school districts, such as state systemic reform efforts, because it is difficult to align test content to the variations in how different districts approach particular standards (Cohen, 1995).

In urban systemic programs dealing with only one district, reformers can focus their attention on issues of a particular program of curriculum and instruction. Assessment can be more closely linked to the specifics of the reform efforts, what Ruiz-Primo et al. (2002) refer to ‘‘close’’ outcome measures. Evaluations of urban systemic programs need to include nationally normed and state-mandated measures, but they also can include assessments that are carefully matched to the district's curriculum goals and daily activities of teachers and students, thereby tracking program effects over time on potentially sensitive measures of outcome. A thorough understanding of systemic reform requires evaluations that address the full range of measures, from very close to distal (Klein, Hamilton, McCaffrey, Stecher, Robyn, & Burroughs, 2000).

This article reports findings from the use of curriculum materials collaboratively developed by the University of Michigan and Detroit Public Schools. The curriculum materials span the middle school years (grades 6–8) and are aligned with the Detroit curriculum framework and serve the district's urban systemic reform program in science. These materials along with professional development form one significant component of a larger urban systemic reform effort in the district. The data we present are pretest–posttest gain scores based on ‘‘close’’ measures of learning from four curriculum units. Three years of outcomes are included to examine whether outcomes improved over time as the reform scaled. Our goal is to demonstrate student achievement that can be attained when the focus of an urban reform is highly specified (the principles and methods are clearly defined) and developed (materials for teachers and learners are available and usable) (Blumenfeld et al., 2000; Cohen & Ball, 1999). We present data across 3 years to show gains over time as the reform effort scales.

## Background

This effort is part of the Center for Learning Technologies in Urban Schools (LeTUS), an NSF-funded collaborative project including the Detroit and Chicago Public Schools, the University of Michigan, and Northwestern University. The LeTUS effort partners university-based research and development organizations with the NSF-funded urban systemic reform programs in the school districts. LeTUS creates inquiry curriculum supported by technology that is aligned to national and local standards and provides professional development to help teachers enact the curriculum materials.

The LeTUS mission is to form collaborations among the four participating organizations to create capacity in the districts to succeed in their science reform programs. We feature new learning technologies, but focus on a range of systemic issues that are needed for success: curriculum design, development and enactment; teacher professional development; and creating and sustaining policy and management structures that support reform. The work is highly collaborative, with teachers, administrators, and researchers working together on the full range of the Center's activities. Curriculum materials are designed by collaborative teams and are revised yearly based on teachers' experiences in enactment and student outcome data. Teacher professional development began as an effort led by university researchers, but increasingly has become an effort jointly constructed by teachers and researchers and largely conducted by teachers. The work reported in this article is based on the University of Michigan/Detroit Public Schools collaboration.

### *Curriculum Design*

Curriculum design that addresses the new standards encompasses several critical dimensions. These include identifying the learning outcomes based on national standards; contextualizing the inquiry through driving questions; structuring activities and benchmark lessons to prepare students for investigations; using learning technologies to scaffold student learning; and creating artifacts that demonstrate student understanding and serve as the basis for discussion, feedback and revision (Marx, 2003; Singer, Marx, Krajcik, & Clay-Chambers, 2000a). The curriculum projects used in this work were all derived from the application of these critical dimensions and were developed collaboratively with school personnel.

### *Learning Technologies*

Our approach makes extensive use of software tools that feature modeling, visualization, and information searching. The technologies are used in order to support the learning goals of the curriculum, rather than building curriculum to capitalize on the technology's affordances. Each tool has been designed to take into consideration the unique characteristics of novice learners. They have specially designed supports that help students to complete inquiry tasks they normally would not be able to complete. The tools we use expand the range of questions that students can investigate, the types of data and information that can be collected, and the forms of data representations that can be displayed to aid interpretation. The tools are used across several curriculum projects and years, so that students become familiar with them and can benefit from repeated use (Krajcik et al., 2000; Soloway, Guzdial, & Hay, 1994).

### *Professional Development*

Teachers cannot simply move to inquiry approaches to instruction from recitation and direct instruction. They need to learn many new ideas about students, learning, curriculum, pedagogy, and assessment (Wilson & Berne, 1999). In the collaboration between the University of Michigan and Detroit Public Schools, the conception of professional development is rooted in a theoretical frame called CERA (Marx, Blumenfeld, Krajcik, & Soloway, 1998), which stands for *C*ollaborative construction of understanding; *E*nactment of new practices in classrooms; *R*eflection on practice; and *A*daptation of materials and practices. The goal of our professional development is to prepare teachers to enact the curriculum appropriately for its underlying theoretical basis while adapting it to classroom circumstances (Fishman, Marx, Best, & Tal, 2003). Professional development is conducted in summer institutes, monthly work sessions, teacher discussion groups and with some classroom support. Over the years of this effort, Detroit personnel have taken over much of the responsibility for organizing and conducting professional development (Fishman, Fogleman, Kubitskey, Peek-Brown, & Marx, 2003).

## Methods

The data come from 3 years of curriculum enactment in the sixth, seventh, and eighth grades in Detroit Public Schools. There is one curriculum project in the sixth grade (''How can I build big things?''), two in seventh grade (''What is the quality of air in my community?'' and ''What is the water like in my river?''), and one in the eighth grade (''Why do I have to wear a helmet when I ride my bike?'').

The research design uses pretests and posttests for all students in the classrooms that used the curriculum projects. This design enables us to examine a number of issues that are essential to understanding how systemic reform impacts students over the years it takes to gain traction in the schools. We are able to examine improvements in gain scores across years on the same curriculum and in cognitive level of assessment items addressed by each unit.

Throughout the 3 years of this work, we continued to revise our curriculum materials and professional development and work with schools to create conditions to improve access to technology and increase support for the innovation. Consequently, the gains in student learning are the joint result of factors associated with the curriculum materials, teaching, and school. This research design does not allow us to compare these results with those that might have obtained from other innovations or other instructional approaches in use in Detroit. The design does allow us to assemble evidence for the type of change that might result from a component of a broader, multifaceted urban systemic reform using inquiry-based curriculum supported by technology.

### *Setting*

The Detroit Public Schools is a large system serving about 165,000 students from a diverse urban community and employing about 10,000 teachers and other education professionals. Like most large American cities, students often come from poor families (about half of Detroit's children and youth live in families that are at or below the poverty line), are largely minority, and tend to be mobile. Dropout rates are high and students' test scores are low compared with performance of students across the state. We do not wish to promulgate a ''deficit'' model to characterize students who participated in this work. Indeed, students in Detroit have ''funds of knowledge'' (Moll, Amanti, Neff, & Gonzalez, 1992) that they bring to bear on their classroom learning as well as a range of family and community support mechanisms (Moje et al., 2001). Like many urban systems, there are frequent changes in leadership. During the period of the work reported here, there were three different superintendents (now called chief executive officers) and new principals were appointed in 11 of the 14 participating schools. The resulting uncertainty makes it very difficult to sustain attention to science reform, develop capability and build system capacity.

The 14 schools that were involved in this work represent the broad range of schools and neighborhoods in the city, ranging from inner city schools serving communities with high poverty to schools in more suburban, and somewhat more affluent neighborhoods. Across the district, 91% of the students are African American, 4% Latino, 4% white, and 1% Asian. This distribution characterizes the sample in this study. Because of mutually agreed-upon procedures established with the school district, we did not collect and do not report data based on the racial or ethnic identities of the students.

### *Participants*

The teachers who participated in this work were faculty members at the schools participating in LeTUS. University researchers collaborated with senior district administrators in the selection of schools, which were invited to participate based on several informal criteria. First, the district required that the teachers had the capacity to engage in the professional development and innovative instructional program. Our goal with this criterion was that we would not have a large number of teachers working out-of-field or who had major skill deficiencies in fundamental areas of teaching. Second, we wanted participating schools to have a sufficient computing infrastructure so students would have access to technology when it was required in the curriculum. Third, there needed to be a supportive administration in the schools so when problems and difficulties arose they could be resolved in a timely and efficient manner. Fourth, the district administration wished to ensure a broad program of equity across schools so innovative programs were not concentrated in some schools to the exclusion of others.

The procedure for selecting schools began with a discussion at the central office, identifying the schools that met the above criteria. The associate superintendent then sent an invitation to the principal and/or assistant principal, the science unit head, and the technology coordinator (if the school had such a person) to a meeting with University of Michigan researchers and her. In the first year, 10 schools agreed to participate (two declined); in the second four schools were added to the group, and the same 14 schools continued in the third year.

In most schools, one to three regular faculty participated, based on interest or because they were selected by their school administration. In general, the LeTUS teachers were comparable in most respects to the general Detroit Public Schools teacher population. The LeTUS teacher pool had a slightly lower percentage of uncertified teachers working under a special teaching license (LeTUS, 9%; DPS, 12%), but a higher percentage of teachers teaching outside of their certification area (LeTUS, 30%; DPS, 14%), although several of this group had extensive professional development experiences in science education. The most significant difference is that the LeTUS teachers have less teaching experience than the general DPS population. The teachers were highly experienced (about 11 years, about 7.6 in science), although less than the average DPS teacher (about 16 years of experience).

Despite our intention that the schools have adequate technology available, the schools had a range of computer technology for classroom use, reliability and capacity of machines varied considerably, maintenance was not always timely and Internet access was inconsistent (see Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004). The technology professionals in the schools were primarily teachers who taught computer courses; they were not responsible for maintenance of the computers, peripherals, and networks in the building. The LeTUS project provided support personnel to help the schools maintain equipment and systems to support the curriculum, although we were not able to maintain all schools at an optimum level of functioning. Teachers had access to learning technologies either in their classrooms or in labs that were available to other classrooms in the school. In most cases, if there were computers in the classrooms, they were in ratios of about 2–4 students per machine. As is the case in most schools, the computers ranged in age and power. Because the computer labs were in high demand, they were often not available for science classes at the times when they were needed for the curriculum or they needed to be reconfigured to run the software used in the units.

Table 1 presents the number of teachers, classrooms, and students engaged in the curriculum units over each of the 3 years. In most cases, teachers taught the curriculum to several of their classes. Information is shown for the seventh- and eighth-grade units (air, water, and helmets) for 3 years; the sixth-grade project, "How can I build big things?," was developed after the others and was used only in 1999–2000 and 2000–2001.

### *Curriculum Materials*

Four projects were designed to engage students in inquiry-based learning activities supported by embedded learning technologies. The projects, which take about 8–10 weeks to complete, were designed to align with science education standards identified in Benchmarks for Science

Table 1
*Number of teachers, classrooms, and students*

| Project | Grade | Year | Teachers | Classrooms | Students |
|---|---|---|---|---|---|
| Air | 7 | 1998–1999 | 10 | 31 | 627 |
| | | 1999–2000 | 8 | 33 | 900 |
| | | 2000–2001 | 14 | 40 | 1203 |
| Water | 7 | 1998–1999 | 11 | 33 | 615 |
| | | 1999–2000 | 12 | 35 | 1091 |
| | | 2000–2001 | 19 | 58 | 1201 |
| Helmets | 8 | 1998–1999 | 3 | 6 | 110 |
| | | 1999–2000 | 8 | 25 | 750 |
| | | 2000–2001 | 11 | 26 | 800 |
| Big Things | 6 | 1999–2000 | 2 | 7 | 210 |
| | | 2000–2001 | 4 | 14 | 490 |

Literacy (American Association for the Advancement of Science, 1993) and the National Science Education Standards (National Research Council, 1996). In addition, they were carefully aligned with the middle school science curriculum framework for Detroit Public Schools. These curricula: (a) use driving questions related to students daily lives, (b) embed learning technologies; (c) engage students in inquiry, (d) contain activities to build skills and background knowledge to prepare students for investigations; and (d) contextualize the learning experiences (see Singer et al., 2000a):

- *How Can I Build Big Things?* This sixth-grade project (Rivet & Krajcik, 2002) enables the learner to develop understanding of simple machines, mechanical advantage, and the relationship among balanced and unbalanced forces in using building structures to contextualize. Students examine built structures such as their school building or other community sites, and investigate the machines that were used in their construction. Learners use technology for interpreting and visualizing physical phenomena graphically. The project integrates the use of microcomputer based labs such as force probes to compare how unbalanced forces result in motion.
- *What Is the Quality of Air in My Community?* In this seventh-grade project (Amati, Singer, & Carrillo, 1999), the learner develops an understanding of factors that affect air quality with a focus on the particulate nature of matter and chemical and physical properties. Learners examine different sources of pollution in their neighborhood and use archived data to compare air quality in Detroit with that of other cities. Through the use of Model-It (Jackson, Krajcik, & Soloway, 2000), a dynamic modeling tool, students model their emerging understanding of factors that affect the quality of the air in their community. Students also use eChem software (Wu, Krajcik, & Soloway, 2001) to visualize and compare molecules in air.
- *What Is the Water Like in My River?* In the context of learning about water ecology (Singer, Rivet, Schneider, Krajcik, Amati, & Marx, 2000b), seventh-grade learners construct an integrated understanding of science concepts such as watersheds, erosion and deposition, and chemistry concepts such as pH and dissolved oxygen. Students use microcomputer based labs (i.e., pH, dissolved oxygen, and temperature probes) to collect and visualize real-time data as they conduct water quality testing. Students use Model-It to represent their understandings of the watershed, erosion and deposition, runoff, and the impact of these factors on water quality.
- *Why Do I Need to Wear a Helmet When I Ride My Bike?* In this eighth-grade project (Schneider, Krajcik, & Blumenfeld, 2002) students focus on the investigation of the

> physics of collisions. Learners develop an integrated understanding of force, velocity, acceleration and Newton's first law in the context of being pitched off their bike, getting injured, and learning how helmets work. Learners also develop strategies for interpreting and visualizing physical phenomena graphically. The project integrates the use of microcomputer-based labs such as the use of motion probes to explore the relationship in distance and time graphs.

The curriculum materials were revised annually based on feedback from teachers and analyses of student test data. In all cases, however, core content remained essentially the same. Revisions often addressed clarity of tasks for students, revision of activities in order to focus them more clearly on intended outcomes, and attention to instructional issues to make the curriculum more teachable.

### *Measures*

To assess student understanding of the curriculum content and science process skills, we developed written assessment instruments that were administered to each student participating in the curriculum projects (for sample items, see the Appendix). Each unit included a range of artifacts that could measure student learning. In other articles, we have reported the results of analyses of some of these artifacts (Hug & Krajcik, 2002; Moje et al., 2004). In the present study, we concentrate on those aspects of learning from the units that are assessed by achievement measures we constructed. We recognized that these achievement measures represent only a range of the student learning that is possible and even desirable from participation in the units. However, these measures do reflect performance on tests that are commonly used to assess the impact of curriculum materials.

The assessments consisted of a combination of multiple choice and free response items, which were further classified as either curriculum *content knowledge* or *science process* skill items. Content and process items were categorized into one of three cognitive levels (Anderson & Krathwohl, 2001): *lower* (recalling information; understanding simple and complex information); *middle* (drawing or understanding simple relationships; applying knowledge to new or different situations; shifting between representations such as verbal to graphic; scientific processes such as identifying hypotheses, procedures, results, or conclusions); and *higher* (describing or analyzing data from charts and graphs; framing hypotheses; drawing conclusions; defining or isolating variables given in a scenario; applying investigation skills; and using concepts to explain phenomena). Content validity was ensured by creating items based on a matrix of topics that reflected the relative importance of the content and processes in the curriculum materials. Our approach to construct validity follows Cronbach's (1971) conception. Each year, the measures were changed slightly in order to correct inappropriately worded items and to ensure that the tests remained closely aligned with the curriculum materials, which were also modified slightly each year. A large core set of items was retained across all 3 years.

Total score reliability (Cronbach's alpha) for each of the test instruments fell in the range 0.63 to 0.78, with the exception of the Helmets unit (alpha ~0.5 over 3 years). Subscale reliabilities fell within the range of 0.30–0.69, again with the exception of some Helmets subscales. As a relatively small number of items were contributing simultaneously to several constructs, we considered somewhat weak statistical scale reliabilities to be acceptable when coupled with strong theoretical content validity.

The curriculum development teams (including science educators, content specialists, educational psychologists, and classroom teachers) constructed the tests. We analyzed all questions on all tests according to the scheme described above with teams of 3–5 raters (senior researchers and graduate students working in LeTUS), achieving at least 95% accuracy in categorizing items on each test. Disagreements were settled by consensus. Yearly, we used rubrics to score each open-ended question using a 10% sample of actual item responses for each test (pre and post). Two to four scorers (undergraduate science students and graduate science and science education students) scored the open-ended items after reaching 95% agreement. Again, disagreements were settled by consensus.

### *Data Collection*

The tests were administered to all the students participating in a curriculum project at the start of the first week of the curriculum (pretest), and again at the conclusion of the last week of the curriculum as implemented by the individual teachers (posttest). The same tests were used both as pre- and posttests. Test administration time allotted was one class period.

Test administration in urban curriculum projects involving many schools and teachers poses considerable challenges, including students not returning informed consent forms, and high absenteeism and mobility. Additional logistical problems were raised by the fact that teachers finished the curriculum at different times. During the first year, blizzards with accompanying school closures and student absences resulted in considerable attrition. For each curriculum enactment, analysis of achievement in science content and process skills consisted of the set of students for whom we were able to obtain matched pretest–posttest pair; the attrition rate between pretest and posttest was relatively consistent at 20% across curricula and years. All subsequent analyses were conducted on a paired-sample basis using this sample only. Checks showed some differences between this group and students without posttests.[1]

## Results

This research concerns the extent to which the students learned from participating in the curriculum projects. Recall that the tests had two component scores, science content and process, and a total score that was the sum of the two components. For the first set of analyses, we aggregated across teachers and computed within-subject *t*-tests for the two component scores and the total score. The results of these analyses are reported in Table 2.

All the analyses, with the exception of the process score on the 1998–1999 water unit, showed statistically reliable gains. The effect sizes for these gains were stronger for content scores than process scores. Moreover, the weighted average effect sizes for total, content and process scores (calculated across curriculum projects) grew stronger across the 3 years (Figure 1). Specifically, Figure 1 shows that for all three scores, the effect sizes were more robust in the second and third years. For the content scores, the average student on the posttest in 1998–1999 achieved a percentile score of 73 compared with the pretest distribution, but in 1999–2000 the posttest percentile score was 88 and in 2000–2001 it was 91. For the process scores, the average student in 1998–1999 achieved a percentile score of 62 on the posttest, but in 1999–2000 the percentile score was 69 and in 2000–2001 it was 72.

All the tests were constructed to have items at three cognitive levels: low, medium, and high. We analyzed gains from pre- to posttests for all the projects on these three item types (Table 3). All the gains were statistically reliable at $p < .001$ with two exceptions: the low items for the 1998–1999 Air project were not reliably higher at posttest, and the high items on the Water project in 1998–1999 were significant at $p < .01$. Figure 2 shows the weighted average effect sizes for these item types for each year, aggregated across projects. It is clear from these data that there was a steady increase in effect sizes for medium and high items across the 3 years, but the low items had a more dramatic increase from 1998–1999 to 1999–2000 and remained steady for the third year.

Table 2
*Summary of learning gains for total content and process scores*

| Project | Year | Component | Maximum Score | *N* | Pretest Mean (SD) | Posttest Mean (SD) | Effect Size |
|---|---|---|---|---|---|---|---|
| Air | 1998–1999 | Total | 34 | 389 | 8.79 (4.34) | 11.19 (5.11) | 0.55*** |
| | | Content | 18 | 389 | 4.72 (2.10) | 5.66 (2.37) | 0.45*** |
| | | Process | 16 | 389 | 4.07 (3.15) | 5.52 (3.56) | 0.46*** |
| | 1999–2000 | Total | 24 | 587 | 6.51 (3.04) | 10.33 (4.65) | 1.25*** |
| | | Content | 16 | 587 | 3.69 (1.86) | 6.24 (3.02) | 1.37*** |
| | | Process | 8 | 587 | 2.83 (1.92) | 4.09 (2.24) | 0.66*** |
| | 2000–2001 | Total | 24 | 860 | 6.46 (2.62) | 11.27 (4.86) | 1.84*** |
| | | Content | 16 | 860 | 4.11 (1.75) | 7.51 (3.33) | 1.94*** |
| | | Process | 8 | 860 | 2.35 (1.55) | 3.76 (2.06) | 0.91*** |
| Water | 1998–1999 | Total | 33 | 312 | 8.38 (4.10) | 10.00 (5.05) | 0.40*** |
| | | Content | 15 | 312 | 4.75 (1.96) | 6.01 (2.47) | 0.64*** |
| | | Process | 18 | 312 | 3.63 (3.06) | 3.99 (3.64) | 0.12 |
| | 1999–2000 | Total | 24 | 755 | 8.31 (3.11) | 11.56 (4.03) | 1.05*** |
| | | Content | 16 | 755 | 5.64 (2.04) | 8.15 (2.80) | 1.23*** |
| | | Process | 8 | 755 | 2.67 (1.71) | 3.41 (1.85) | 0.43*** |
| | 2000–2001 | Total | 24 | 754 | 9.24 (3.37) | 11.59 (4.16) | 0.70*** |
| | | Content | 16 | 754 | 6.24 (2.18) | 8.23 (2.75) | 0.91*** |
| | | Process | 8 | 754 | 3.00 (1.83) | 3.36 (1.99) | 0.20*** |
| Helmets | 1998–1999 | Total | 53 | 78 | 13.29 (4.39) | 17.82 (7.26) | 1.03*** |
| | | Content | 42 | 78 | 7.69 (2.72) | 11.46 (5.44) | 1.39*** |
| | | Process | 11 | 78 | 5.66 (2.38) | 6.44 (2.56) | 0.33* |
| | 1999–2000 | Total | 21 | 529 | 5.97 (2.06) | 7.67 (2.73) | 0.83*** |
| | | Content | 16 | 529 | 4.13 (1.63) | 5.50 (2.19) | 0.84*** |
| | | Process | 5 | 529 | 1.84 (1.11) | 2.18 (1.12) | 0.31*** |
| | 2000–2001 | Total | 24 | 413 | 6.69 (2.56) | 8.83 (3.28) | 0.84*** |
| | | Content | 16 | 413 | 3.85 (1.71) | 5.27 (2.21) | 0.83*** |
| | | Process | 8 | 413 | 2.84 (1.54) | 3.56 (1.85) | 0.47*** |
| Big Things | 1999–2000 | Total | 24 | 179 | 9.78 (3.67) | 14.78 (5.19) | 1.36*** |
| | | Content | 16 | 179 | 7.03 (2.56) | 10.51 (3.31) | 1.36*** |
| | | Process | 8 | 179 | 2.74 (1.55) | 4.26 (2.23) | 0.98*** |
| | 2000–2001 | Total | 24 | 299 | 7.57 (3.36) | 12.34 (3.99) | 1.42*** |
| | | Content | 16 | 299 | 5.19 (2.37) | 8.49 (2.89) | 1.39*** |
| | | Process | 8 | 299 | 2.37 (1.59) | 3.85 (1.76) | 0.93*** |

*$p < .05$.
***$p < .001$.

For the low scores, the average student on the posttest in 1998–1999 achieved a percentile score of 62 compared with the pretest distribution, but both in 1999–2000 and 2000–2001 the posttest percentile score was 83. For the medium items, the average student in 1998–1999 achieved a percentile score of 64 on the posttest, but in 1999–2000 the percentile score was 71 and in 2000–2001 it was 77. For the high items, the average student in 1998–1999 achieved a percentile score of 68 on the posttest, but in 1999–2000 the percentile score was 79 and in 2000–2001 it was 82.

## Discussion

The fundamental question investigated in this work was whether an inquiry-based and technology-infused curriculum could help students in an underperforming urban school district

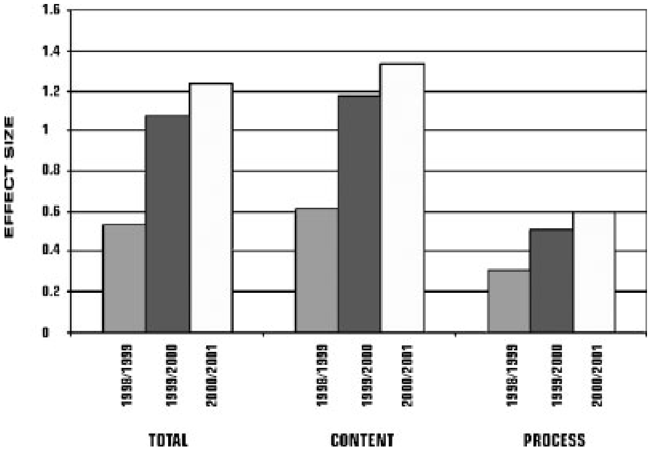


*Figure 1.* Weighted average effect sizes for total, content, and process scores across years.

learn important science content that addresses national standards. Our data show the answer to be in the affirmative. It is noteworthy that the impact of the innovation, as measured by the increasing effect size statistics over the years, continued to grow as scaling occurred. We were able to accomplish these results because of the work was embedded in a systemic reform context. The curriculum units were collaboratively developed to align with the district's evolving curriculum framework. Assessments were used that were aligned with curriculum materials and the district's curriculum framework. A professional development program was designed to engage teachers in the intensive learning needed for them to change their practices to support standards-based, inquiry instruction. Senior district leaders worked with researchers to realign district policies to enable standards-based instruction, thereby providing direction and momentum for the science reform initiative within the district.

It is noteworthy that these findings represent success under very real circumstances of urban schooling. The conditions under which we attained these gains were not always ideal, as of course they rarely are (Blumenfeld et al., 2000; Fishman et al., 2004). There were many competing priorities for teachers and schools; these curriculum units were nested in a range of issues that demanded attention. For example, teachers' efforts and time were also devoted to several other science initiatives and professional development opportunities as part of the systemic reform effort. In addition, there was continued teacher mobility, with some participating teachers leaving the district and others transferring to other schools. Despite attention to maintenance of computer and network technology, difficulties persisted. In the classrooms, time for curriculum enactment competed with the pressing need for teachers and students to prepare for highly visible and politically important testing programs. And of course, there was continued change in the superintendent's office and among the school principals. We hypothesize that were we able to overcome these challenges, student learning gains would be even more impressive.

This study contributes to the creation of a body of empirical data on urban reform in science. It helps build a corpus of knowledge about the range of achievement gains that are possible over time under day-to-day conditions when this type of innovation is enacted in an urban setting undergoing systemic reform in science. Much of the research deals with initiatives that examine how self-reports of participation in professional development relate to standards-based teaching (Supovitz & Turner, 2000). Other studies investigate the extent to which teacher participation in professional development influences students' performance on state tests (Kahle et al., 2000). Such efforts

Table 3
*Summary of learning gains for high, medium, and low scores*

| Project | Year | Component | Maximum Score | *N* | Pretest Mean (SD) | Posttest Mean (SD) | Effect Size |
|---|---|---|---|---|---|---|---|
| Air | 1998–1999 | High | 13 | 389 | 2.40 (2.67) | 3.99 (3.09) | 0.60*** |
| | | Medium | 18 | 389 | 5.36 (2.15) | 5.99 (2.37) | 0.29*** |
| | | Low | 3 | 389 | 1.21 (0.88) | 1.27 (0.88) | 0.07 |
| | 1999–2000 | High | 6 | 587 | 1.07 (1.58) | 2.21 (1.92) | 0.72*** |
| | | Medium | 10 | 587 | 2.96 (1.53) | 4.02 (1.99) | 0.69*** |
| | | Low | 8 | 587 | 2.51 (1.30) | 4.12 (1.92) | 1.24*** |
| | 2000–2001 | High | 6 | 860 | 0.72 (1.06) | 2.14 (1.70) | 1.34*** |
| | | Medium | 10 | 860 | 3.26 (1.47) | 4.71 (2.14) | 0.99*** |
| | | Low | 8 | 860 | 2.49 (1.39) | 4.42 (2.14) | 1.39*** |
| Water | 1998–1999 | High | 16 | 312 | 2.76 (2.81) | 3.21 (3.23) | 0.16** |
| | | Medium | 10 | 312 | 3.23 (1.63) | 3.67 (1.74) | 0.27*** |
| | | Low | 7 | 312 | 2.39 (1.32) | 3.12 (1.56) | 0.55*** |
| | 1999–2000 | High | 6 | 755 | 0.49 (0.87) | 1.26 (1.54) | 0.87*** |
| | | Medium | 9 | 755 | 3.70 (1.58) | 4.67 (1.64) | 0.61*** |
| | | Low | 9 | 755 | 4.12 (1.76) | 5.64 (1.94) | 0.86*** |
| | 2000–2001 | High | 6 | 754 | 0.65 (1.00) | 1.23 (1.45) | 0.58*** |
| | | Medium | 9 | 754 | 4.00 (1.61) | 4.68 (1.84) | 0.42*** |
| | | Low | 9 | 754 | 4.59 (1.85) | 5.69 (2.00) | 0.59*** |
| Helmets | 1998–1999 | High | 18 | 78 | 1.23 (1.17) | 2.32 (2.17) | 0.93*** |
| | | Medium | 19 | 78 | 5.79 (1.83) | 7.91 (2.81) | 1.15*** |
| | | Low | 16 | 78 | 6.27 (2.73) | 7.59 (3.29) | 0.48*** |
| | 1999–2000 | High | 4 | 529 | 0.23 (0.50) | 0.62 (0.81) | 0.78*** |
| | | Medium | 9 | 529 | 3.37 (1.37) | 3.73 (1.56) | 0.26*** |
| | | Low | 8 | 529 | 2.37 (1.27) | 3.33 (1.40) | 0.75*** |
| | 2000–2001 | High | 6 | 413 | 1.14 (1.05) | 1.50 (1.36) | 0.34*** |
| | | Medium | 10 | 413 | 3.12 (1.54) | 4.30 (1.71) | 0.77*** |
| | | Low | 8 | 413 | 2.44 (1.33) | 3.03 (1.46) | 0.44*** |
| Big Things | 1999–2000 | High | 6 | 179 | 0.71 (1.05) | 1.58 (1.41) | 0.83*** |
| | | Medium | 10 | 179 | 4.14 (1.77) | 5.34 (1.86) | 0.68*** |
| | | Low | 8 | 179 | 2.42 (1.26) | 3.99 (1.49) | 1.25*** |
| | 2000–2001 | High | 6 | 299 | 1.15 (1.27) | 2.73 (1.72) | 1.25*** |
| | | Medium | 10 | 299 | 3.96 (1.95) | 5.34 (1.93) | 0.71*** |
| | | Low | 8 | 299 | 2.46 (1.51) | 4.28 (1.64) | 1.21*** |

$^{**}p < .01$.
$^{***}p < .001$.

rely on tests that are distal measures that may not be well aligned with district or state standards. In addition, often such studies do not report change across years. Our approach to assessment that includes close and proximal measures (Ruiz-Primo et al., 2002) enabled us to track changes over 3 years, thus producing evidence that reform can pay off. Such careful and long-term assessment provides information about where work needs to be done and where to focus additional resources. For example, gains on science content items were more robust than gains on process items, which suggests areas needing attention in the curriculum materials and professional development.

Our findings show what can be accomplished with a highly specified and developed multifaceted effort. We were able to show increasing gains on student achievement, but because

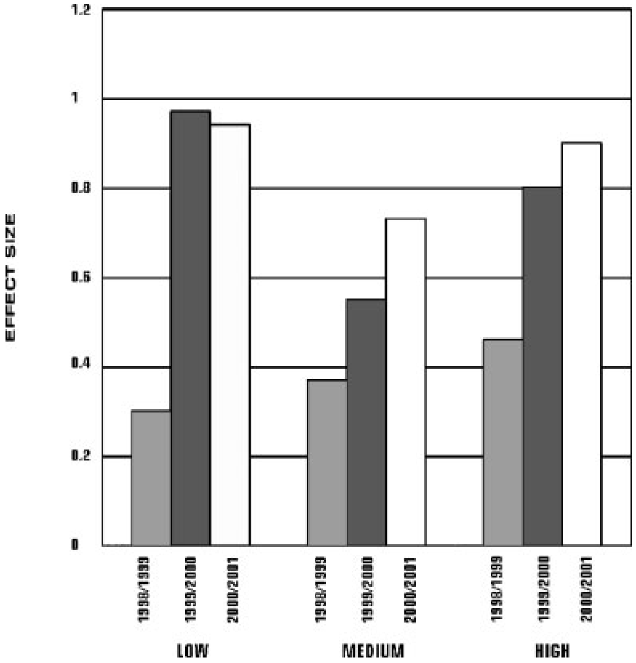


*Figure 2.* Weighted average effect size for low, medium, and high cognitive items across years.

systemic reform efforts are complex and integrated, it is difficult, if not impossible to tease apart the effects of the components and attribute causality to one or another (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). Teachers who taught these units over the years of this study became more experienced with them. Some of the students participated in more than one unit, potentially benefiting from a possible cumulative effect on their learning. Professional development changed as teachers attained more experience with the LeTUS effort. Our findings suggest that future research needs to track the various components of the innovation over time, addressing questions such as: What are the effects of professional development over time? What are the cumulative effects of students participating in several units across the middle school years? How do teachers change in their enactment of the materials? Over time, how does policy change support or interfere with reform efforts?

In conclusion, our findings show that reform programs that address the range of elements needed for coherence can succeed in urban settings. A combination of carefully designed curriculum materials, learning technologies that are embedded in the materials and serve the needs of learners, high quality professional development, and policies that support reform are necessary. But equally important, collaboration among partners is fundamental. In the case of LeTUS, the collaboration is premised on mutual benefit; the district was able to move toward its goals of science reform and the researchers were able to contribute to the field's growing corpus of literature on science curriculum, learning technology, and teacher professional development. The quality of the collaboration enabled the participants to be tenacious and to persist toward shared goals. It also enabled timely solutions to problems that emerged along the way. When coupled with feedback about student learning, design teams were able to modify the units to make them more usable and still focused on standards. Thus, the district's capability was enhanced by the growing skill and knowledge of teachers, and curriculum materials that were more tailored to the district's context.

## Acknowledgments

All the opinions expressed in this work are those of the authors.

## Note

[1] We examined the pretest results for which we were unable to gather a matching posttest, and compared this pool of absent or otherwise attritted students with the analysis sample's pretest scores using pooled-sample *t*-tests. For the 1998–1999 school year implementations, there were no significant differences or trends between the analysis sample and the dropped sample on total pretest score. For the 1999–2000 and 2000–2001 curriculum implementations, analysis of the missing data did indicate some differences between the analysis sample and the students for whom no posttest was collected. In general, and not surprisingly, the analysis sample slightly outperformed the dropped students on the pretest measures, suggesting that the students included in the results were better prepared academically than those who were dropped. The effect sizes, however, were relatively small, averaging only .25. The difference between the analysis sample and the dropped sample does not directly affect the paired-sample analysis, but it may have an indirect effect in that for our sample, better prepared students tend to show slightly lower gain scores ($r = -.18$, $p < .001$). This suggests that our student attrition may cause a slight reduction in overall gain scores possible without attrition. However, the small effect sizes indicated for the difference in the missing data lead us to be reasonably confident in the validity of our estimates.

## Appendix: Sample Test Items

### *Low Cognitive Items*

From “What is the quality of air in my community?” (seventh grade):
Which substance occurs in the largest amount in “clean” air?

A. Nitrogen
B. Oxygen
C. Carbon dioxide
D. Sulfur dioxide

From “What is the water like in my river?” (seventh grade):
When water flows in a river, soil may be removed from the sides and transported down stream where it settles to the bottom. This process of settling to the bottom is called:

A. ground water
B. run-off
C. erosion
D. deposition

### *Medium Cognitive Items*

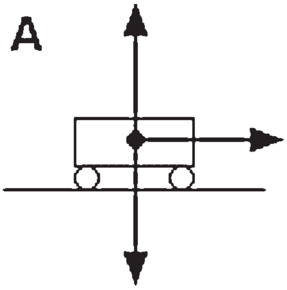
From “What is the quality of air in my community?” (seventh grade):
When you find water droplets on the grass in the morning and it hasn't rained, it is an example of:

A. sublimation
B. freezing
C. boiling
D. condensation

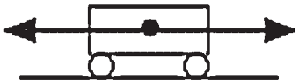
From “Why do I have to wear a helmet when I ride my bike?” (eighth grade):

A car is parked on the street in front of your house. The street is level. Which of the following diagrams shows all the forces acting on the car?

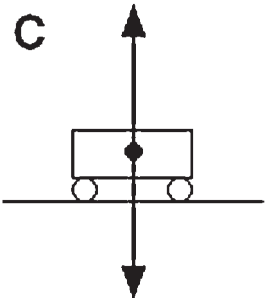
A



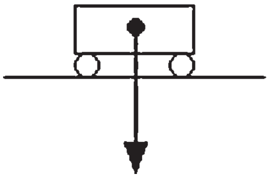
B



C



D



*High Cognitive Items*

From “How can I build big things?” (sixth grade):
Explain why it is easy to use a screwdriver to open a can of paint. Use the terms machine, force, and distance in your response:

From “Why do I have to wear a helmet when I ride my bike?” (eighth grade):
It is less dangerous to jump from a 5-foot-high wall onto very loose sand than onto concrete pavement. You are more likely to be injured when landing on the concrete.

Describe how the speed, stopping time, and force compare for landing on concrete and landing in sand. Use this information to describe why it is safer to land in the sand.

## References

Amati, K., Singer, J., & Carrillo, R. (1999, April). What affects the quality of air in my community? Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

American Association for the Advancement of Science. (1993). Benchmarks for science literacy, Project 2061. New York: Oxford University Press.

Anderson, L.W. & Krathwohl, D.R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Berends, M., Kirby, S.N., Naftel, S., & McKelvey, C. (2001). Implementation and performance in New American Schools. Santa Monica, CA: RAND.

Blumenfeld, P., Fishman, B., Krajcik, J.S., Marx, R.W., & Soloway, E. (2000). Creating usable innovations in systemic reform: Scaling-up technology-embedded project-based science in urban schools. Educational Psychologist, 35, 149–164.

Blumenfeld, P., Krajcik, J.S., Marx, R.W., & Soloway, E. (1994). Lessons learned: How collaboration helped middle grade science teachers learn project-based instruction. The Elementary School Journal, 94, 539–551.

Blumenfeld, P., Soloway, E., Marx, R.W., Krajcik, J.S., Guzdial, M., & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. Educational Psychologist, 26, 369–398.

Bouillion, L. & Gomez, L. (2001). Connecting school and community with science learning: Real world problems and school-community partnerships as contextual scaffolds. Journal of Research in Science Teaching, 38, 878–898.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. Educational Researcher, 32, 9–13.

Cohen, D.K. (1995). What is the system in systemic reform? Educational Researcher, 24, 11–17, 31.

Cohen, D.K. & Ball, D.L. (1999). Instruction, capacity, and improvement (CPRE Research Report Series No. RR-043). Philadelphia, PA: University of Pennsylvania Consortium for Policy Research in Education. Retrieved March 30, 2003 (http://www.cpre.org/publications/rr43.pdf).

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education.

Edelson, D.C., Gordin, D.N., & Pea, R.D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. Journal of the Learning Sciences, 8, 391–450.

Fishman, B., Fogleman, J., Kubitskey, B., Peek-Brown, D., & Marx, R. (2003, March). Taking charge of innovations: Fostering teacher leadership in professional development to sustain reform. Presented at the Annual Meeting of the National Association of Research on Science Teaching, Philadelphia, PA. Retrieved November 11, 2003 (http://www.umich.edu/~fishman/papers/FishmanFoglemanNARST2003.pdf).

Fishman, B., Marx, R., Best, S., & Tal, R. (2003). Linking teacher and student learning to improve professional development in systemic reform. Teaching and Teacher Education, 19, 643–658.

Fishman, B., Marx, R., Blumenfeld, P., Krajcik, J.S., & Soloway, E. (2004). Creating a framework for research on systemic technology innovations. Journal of the Learning Sciences, 13, 43–76.

Fuhrman, S.H. (Ed.). (2001). From the capitol to the classroom: Standards-based reform in the states. 100th Yearbook of the National Society for the Study of Education (Part II). Chicago, IL: University of Chicago Press.

Garet, M.S., Porter, A.C., Desimone, L., Birman, B.F., & Yoon, K.S. (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, 38, 915–945.

Hannaway, J. & Kimball, K. (2001). Big isn't always bad: School district size, poverty, and standards-based reform. In S.H. Fuhrman (Ed.), From the capitol to the classroom: Standards-based reform in the states. 100th Yearbook of the National Society for the Study of Education (Part II) (pp. 99–123). Chicago, IL: University of Chicago Press.

Hug, B. & Krajcik, J.S. (2002). Students' scientific practices using a scaffolded inquiry sequence. In P. Bell, R. Stevens, & T. Satwicz (Eds.), International Conference of the Learning Sciences (ICLS) (pp. 167–174). Mahwah, NJ: Erlbaum.

Jackson, S., Krajcik, J., & Soloway, E. (2000). Model-It: A design retrospective. In M. Jacobson & R. Kozma (Eds.), Advanced designs for the technologies of learning: Innovations in science and mathematics education (pp. 77–116). Hillsdale, NJ: Erlbaum.

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.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000). Teaching practices and student achievement: Report of the first year findings of the "Mosaic" study of systemic initiatives in mathematics and science. Santa Monica, CA: RAND.

Krajcik, J.S., Blumenfeld, P., Marx, R.W., Bass, K., Fredricks, J., & Soloway, E. (1998). First attempts at inquiry strategies in middle school, project-based science classrooms. Journal of the Learning Sciences, 7, 313–350.

Krajcik, J.S., Blumenfeld, P., Marx, R.W., & Soloway, E. (2000). Instructional, curricular, and technological supports for inquiry in science classrooms. In J. Minstrell & E.H.v. Zee (Eds.), Inquiring into inquiry learning and teaching in science (pp. 283–315). Washington, DC: American Association for the Advancement of Science.

Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. Secada (Ed.), Review of Research in Education (Vol. 26, pp. 23–69). Washington, DC: American Educational Research Association.

Linn, M.C., Clark, D., & Slotta, J.D. (2003). WISE design for knowledge integration. Science Education, 87, 517–538.

Marx, R.W. (2003, April). Partnerships for urban systemic reform: The effects of inquiry curriculum developed by the Center for Learning Technologies in Urban Schools. Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Marx, R.W., Blumenfeld, P., Krajcik, J.S., & Soloway, E. (1997). Enacting project-based science. Elementary School Journal, 97, 341–358.

Marx, R.W., Blumenfeld, P., Krajcik, J.S., & Soloway, E. (1998). New technologies for teacher professional development. Teaching and Teacher Education, 14, 33–52.

Moje, E., Collazo, T., Carrillo, R., & Marx, R.W. (2001). "Maestro, what is 'quality'?": Language, literacy, and discourse in project-based science. Journal of Research in Science Teaching, 38, 469–498.

Moje, E., Peek-Brown, D., Sutherland, L., Marx, R., Blumenfeld, P., & Krajcik, J.S. (2004). Explaining explanations: Developing scientific literacy in middle-school project-based reforms. In D. Strickland & D.E. Alvermann (Eds.), Bridging the gap: Improving literacy for preadolescent and adolescent learners in grades 4–12 (pp. 227–251). New York: Teachers College Press.

Moll, L.C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. Theory Into Practice, 31, 132–141.

National Research Council. (1996). The national science education standards. Washington, DC: National Academy Press.

Porter, A.C. & Smithson, J.L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S.H. Fuhrman (Ed.), From the capitol to the classroom: Standards-based reform in the states. 100th Yearbook of the National Society for the Study of Education (Part II) (pp. 60–80). Chicago, IL: University of Chicago Press.

Rivet, A. & Krajcik, J.S. (2002, April). Project-based science curricula: Achieving national standards in urban systemic reform. Presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Roth, W.M. (1995). Authentic school science. Dordrecht, The Netherlands: Kluwer Academic.

Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. Journal of Research in Science Teaching, 39, 369–393.

Rutherford, F.J. & Ahlgren, A. (1990). Science for all Americans. New York: Oxford.

Schneider, R.M., Krajcik, J.S., & Blumenfeld, P. (2002, April). Exploring the role of curriculum materials to support teachers in education reform. Presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Singer, J., Marx, R.W., Krajcik, J.S., & Clay-Chambers, J. (2000a). Constructing extended inquiry projects: Curriculum materials for science education reform. Educational Psychologist, 35, 165–178.

Singer, J., Rivet, A., Schneider, R.M., Krajcik, J.S., Amati, K., & Marx, R.W. (2000b, April). Setting the stage: Engaging students in water quality. Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Smith, M.S. & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), The politics of curriculum and testing (pp. 233–267). New York: Falmer.

Soloway, E., Guzdial, M., & Hay, K. (1994). Learner-centered design: The challenge for HCI in the 21st century. Interactions, 1, 36–48.

Songer, N.B., Lee, H.-S., & McDonald, S. (2002). Research towards an expanded understanding of inquiry science beyond one idealized standard. Science Education, 87, 490–516.

Supovitz, J.A., Mahyer, D.P., & Kahle, J.B. (2000). Promoting inquiry-based instructional practice: The longitudinal impact of professional development in the context of systemic reform. Educational Policy, 14, 331–356.

Supovitz, J.A. & Turner, H.M. (2000). The effects of professional development on science teaching practices and classroom culture. Journal of Research in Science Teaching, 37, 963–980.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A., & Hudicourt-Barnes, J. (2001). Rethinking diversity in teaching science: The logic of everyday sense-making. Journal of Research in Science Teaching, 38, 529–552.

Wilson, S.M. & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. In A. Iran-Nejad & P.D. Pearson (Eds.), Review of research in education (pp. 173–209). Washington, DC: American Educational Research Association.

Wu, H.-K., Krajcik, J., & Soloway, E. (2001). Promoting conceptual understanding of chemical representations: Students' use of a visualization tool in the classroom. Journal of Research in Science Teaching, 38, 821–842.