JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 45, NO. 8, PP. 922–939 (2008)

Standardized Test Outcomes for Students Engaged in Inquiry-Based Science Curricula in the Context of Urban Reform

Robert Geier,[1] Phyllis C. Blumenfeld,[1] Ronald W. Marx,[2] Joseph S. Krajcik,[1] Barry Fishman,[1] Elliot Soloway,[1] Juanita Clay-Chambers[3]

[1]*University of Michigan, 610 East University, Ann Arbor, Michigan 48109*
[2]*University of Arizona, Tucson, Arizona*
[3]*Detroit Public Schools, Detroit, Michigan*

Received 29 March 2007; Accepted 29 October 2007

Abstract: Considerable effort has been made over the past decade to address the needs of learners in large urban districts through scaleable reform initiatives. We examine the effects of a multifaceted scaling reform that focuses on supporting standards based science teaching in urban middle schools. The effort was one component of a systemic reform effort in the Detroit Public Schools, and was centered on highly specified and developed project-based inquiry science units supported by aligned professional development and learning technologies. Two cohorts of 7th and 8th graders that participated in the project units are compared with the remainder of the district population, using results from the high-stakes state standardized test in science. Both the initial and scaled up cohorts show increases in science content understanding and process skills over their peers, and significantly higher pass rates on the statewide test. The relative gains occur up to a year and a half after participation in the curriculum, and show little attenuation with in the second cohort when scaling occurred and the number of teachers involved increased. The effect of participation in units at different grade levels is independent and cumulative, with higher levels of participation associated with similarly higher achievement scores. Examination of results by gender reveals that the curriculum effort succeeds in reducing the gender gap in achievement experienced by urban African-American boys. These findings demonstrate that standards-based, inquiry science curriculum can lead to standardized achievement test gains in historically underserved urban students, when the curriculum is highly specified, developed, and aligned with professional development and administrative support. © 2008 Wiley Periodicals, Inc. J Res Sci Teach 45: 922–939, 2008
Keywords: general science; program evaluation; urban education; middle school science

New standards based 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 (Edelson, Gordin, & Pea, 1999; Hancock, Kaput, & Goldsmith, 1992; Marx, Blumenfeld, Krajcik, & Soloway, 1997). Inquiry learning challenges students too. It requires them to collaborate with peers, construct usable knowledge by linking new and old ideas, relate new science content to their lives in and outside of school, and self-regulate across the weeks that an inquiry project might unfold (Blumenfeld et al., 1991; Krajcik et al., 1998).

Research into science achievement among diverse learners in urban environments remains a significant priority within the science education community (Fraser-Abder, Atwater, & Lee, 2006). In particular, there is a lack of credible research on effective science instruction and curricula for diverse student populations (Lee & Luykx, 2004). Implementing standards-based instructional practice in diverse urban school systems presents a particular set of challenges for educators and their partners in reform efforts. These challenges

Contract grant sponsor: REPP; Contract grant numbers: REC-9720383, REC-9725927, REC-9876150; Contract grant sponsor: USI; Contract grant number: ESR-9453665.
Correspondence to: R. Geier; E-mail: bobgeier@umich.edu
DOI 10.1002/tea.20248
Published online 29 August 2008 in Wiley InterScience (www.interscience.wiley.com).

include lack of resources, high levels of poverty, low student achievement, below grade level English proficiency, high student mobility, attendance problems, and difficulty recruiting and retaining highly qualified teachers (Hannaway & Kimball, 2001; Kahle, Meece, & Scantlebury, 2000; Lynch, 2000; Tobin, Roth, & Zimmerman, 2001). Successful implementation of instructional innovation may also be challenged by the cultural, policy and management environment of urban districts (Blumenfeld, Krajcik, Fishman, Marx, & Soloway, 2000). Inquiry instruction must also incorporate materials that engage and are congruent with the culturally relevant knowledge and beliefs of students from diverse backgrounds (Lee & Luykx, 2005; Moje, Collazo, Carrillo, & Marx, 2001).

Despite these challenges, a number of researchers have shown that in highly resourced settings where small numbers of teachers and students can be supported directly by the curriculum designers, inquiry instruction in urban classrooms can be successful (e.g., Bouillion & Gomez, 2001; Fortus, Dershimer, Krajcik, Marx, & Mamlok-Naaman, 2004; Warren, Ballenger, Ogonowski, Rosebury, & Hudcourt-Barnes, 2001). Cuevas, Lee, Hart, and Deaktor (2005) showed that inquiry instruction can yield greater increases in achievement for low-achieving, low-SES at risk students in particular. The problem that remains, however, is to demonstrate that these efforts and results can be expanded beyond the small-scale development environments in which they were created to new sites with less concentrated resources and teacher support, allowing the reform to progress successfully throughout an urban system. Elmore (1996) notes that efforts to advance inquiry and other reform pedagogy typically fail to engage large numbers of teachers in systems that deliver education to most children.

Assessment

The resources required to bring standards-based instructional reforms to wide implementation bring with them demands for ongoing program evaluation that further challenge reformers. State and federal funding agencies are subject to public scrutiny, and Americans ''want numbers when they look at students, schools, state education systems, and how America's students compare to those of other countries'' (Barton, 1999, p. 4). Reform efforts must be able to demonstrate not only that the reform is changing instruction but that changing instruction leads to better prepared, more capable students in a wide array of communities. To achieve wide implementation of an instructional reform, it is necessary throughout the scaling process to ''show that measurable achievement gains have been realized by the intended audience, especially historically disadvantaged students'' (Bruckerhoff, 1997, p. i). Absent such evidence, reform efforts can be compromised, particularly in light of the mandated reorganization penalties associated with the Elementary and Secondary Education ''No Child Left Behind'' statute (Corcoran & Christman, 2002; Marx & Harris, 2006; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Thomas, Woods, Hillman, & Ross, 2002).

Several studies have examined the impacts of inquiry science interventions on measured achievement as they scale up to enactment across multiple schools involving thousands of students. Lee, Buxton, Lewis, and LeRoy (2006) measure the impact of an urban instructional intervention in grades 3–5 using matched pre- and post-tests, and find substantial learning gains. Their data also suggests a cumulative effect of the intervention as students participate over several years. Lynch, Kuipers, Pyke, and Szesze (2005) describe positive achievement effects from a standards-based curriculum unit enacted in 10 urban middle schools. Students who participated in the curriculum unit showed moderate (Cohen's $d = 0.35$) positive effects on a curriculum-aligned test over a same-district comparison group that did not participate. By contrast, Pine et al. (2006) found no significant differences in achievement in investigative tasks between 5th grade students in a hands-on inquiry science curriculum and those receiving a textbook-based curriculum.

For both economic and political reasons, statewide standardized testing has emerged in the current policy environment as the *de facto* means to evaluate educational impacts on students from reform efforts. These tests based on state standards have been referred to as *distal* measures, to distinguish them from *close* and *proximal* assessments used in the aforementioned studies, which are embedded in the curriculum or which examine curricular concepts in a new context (Ruiz-Primo et al., 2002). Distal standardized tests are frequently politically charged, with results used to determine scholarships or graduation for students, funding for schools and programs, or mandated reorganizations and staff replacements. These tests are ''high stakes'' for students, schools, and reformers alike, and demonstrating increased achievement on such distal

instruments thus far has been a weakness in the literature on the effects of inquiry science curricula and the scaling of instructional reform.

A number of studies have reported on the effects of systemic reform in science education using distal state tests aggregated at the school or state level (e.g., Berends, Kirby, Naftel, & McKelvey, 2001; Kim et al., 2001; Ross et al., 2001). Such studies have enhanced our knowledge of the impacts of reform efforts. Aggregated average test scores, however, may be contaminated by factors other than school performance, including student mobility and family and community influences, making it difficult to isolate the effects of the reform effort (Meyer, 1999). Information from disaggregated student data is also necessary for establishing success at reaching historically underserved students, and providing information on how well the reform effort is addressing issues of educational equity (Kahle, 1998; Rodriguez, 2001). Reform efforts which have access to student-level data can examine not only whether a reform ''lifts all boats,'' but whether the reform also helps address ''gaps'' in achievement outcomes between subpopulations.

Unfortunately, student-level state testing data can be difficult to obtain, may prove problematic for technical reasons such as database compatibility, and may not be useful when a reform extends across state boundaries. These challenges may cause researchers to construct their own tests to enable detailed investigations at the student level (Kahle et al., 2000; Klein et al., 2000; Upton & Supovitz, 1996), or to use work-arounds like comparing treatment student responses on individual NAEP or TIMMS items with the students from those broader studies as a proxy for a quasi-experimental control (Lee et al., 2006). The lack of student-level distal standardized test data to demonstrate achievement gains from standards-based inquiry science curricula remains a weakness in the literature.

Urban Partnership

Our group has been engaged in an ongoing partnership effort between the University of Michigan and the Detroit Public Schools, which has been financed in part by the National Science Foundation's Urban Systemic Initiative (USI, 1994–1999) and Urban Systemic Program (USP, 1999–2004). USP efforts in Detroit have provided opportunities for summer workshops, technology resources in classrooms, developing teacher mentors and learning communities, and other supports important to the multi-faceted nature of our intervention.

The partnership efforts have been just one aspect of systemic reform in Detroit. Under consistent leadership during USI and USP funding, the district maintained a well-defined vision for science and mathematics education based on constructivist learning theories. The vision helped drive a pre-K-12 articulation of curriculum objectives consistent with the Michigan Framework and standards. On an administrative level, the district reorganized around ''constellations'' of high school feeder and specialty schools to provide articulation of curriculum across grades and a tiered system of professional development. Constellations were supported by science and mathematics resource centers and district specialists, and helped provide an environment to encourage learning communities. Detroit Public Schools have also partnered with several area universities to provide professional development and special undergraduate and graduate programs focused on teaching in urban schools. The progress of Detroit's efforts has been highlighted in reporting on the Urban Systemic Programs more generally (Kim et al., 2001), including growth on distal achievement test measures in a number of areas.

As a science education group working in an urban reform partnership, we have had a unique opportunity to examine the effects of standards-based reforms in science instruction using project-based inquiry curriculum on a broad scale. Because curriculum in the U.S. is largely a function of local districts, implementation of reform efforts and assessment of student achievement in statewide or multi-state reforms has naturally focused on pedagogy and classroom processes, rather than on curriculum materials (Clune, 1998). However, efforts within a single large district can support standards-based instruction with curriculum materials that are aligned with both professional development and local and state standards. Curriculum, instruction, and professional development can also be supported by learning technologies embedded in the aligned effort. We refer to an effort where standards, policy, curriculum, instruction, professional development, assessment, and learning technologies are coherent and integrated as ''highly aligned.''

The Center for Learning Technologies in Urban Schools (LeTUS) at the University of Michigan developed project-based science curriculum units for middle school students in Detroit which were aligned

with DPS curriculum standards. As Cohen and Ball (1999) recommend, the curricula we designed are highly specified (the theoretical principles and methods are clearly defined) and developed (materials for teachers and learners are available and usable). Professional development is tailored to helping teachers enact the curriculum (Fishman, Marx, Best, & Tal, 2003) and changes in policy and management structures are sought to support the reform (Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004). Instructional technologies support both student learning and the ongoing development of teachers (Krajcik, Blumenfeld, Marx, & Soloway, 2000). Working within a single large district in support of a curricular effort also allows us to link assessment to the instructional effort at several levels of proximity, ranging from close assessments tied to the curriculum to distal standardized measures. Addressing the impact of the reform on a full range of measures is necessary for developing a complete understanding of the impacts of reform (Klein et al., 2000).

We have previously reported on student outcomes from curriculum-specific pre- and post-test ''close'' assessments aligned to the curriculum (Marx et al., 2004). These results in the first 3 years of our partnership effort were promising, with learning gains showing generally large effect sizes that were robust across years and increases in scale. In this article, we examine the impact of participation in project-based curriculum units on distal state standardized achievement measures which are not directly tied to the curriculum but which are aligned in terms of outcome goals. Our core research question is whether urban student participation in project-based inquiry science curricula leads to demonstrably higher student achievement on statewide assessments over and above general district-wide efforts at reform. We believe addressing this question is vital to the ongoing support of standards-based instructional reforms in science.

Background

The Center for Learning Technologies in Urban Schools (LeTUS) is an NSF-funded collaborative project including the Detroit and Chicago Public Schools, the University of Michigan, and Northwestern University. 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.

Setting

The Detroit Public Schools (DPS) serves a large urban community in the industrial Midwest. As the largest school district in the state of Michigan, it has a student population of approximately 160,000 students and employs a professional staff of nearly 10,000. The student body is 91% African American, 5% Latino, and 4% representing a diverse ethnic mix including recent immigrants. Like many urban communities, a high percentage of the school-aged population comes from families with limited economic means, with 69% of the students eligible for federal free and reduced-price lunch. Dropout rates and achievement scores are persistent challenges, as are student mobility and absenteeism. The Detroit Public Schools have also been subject to significant external pressures and changes in administrative leadership. During the period of work reported here, the state dissolved the elected school board in a contentious takeover and the district was led by three different superintendents (now referred to as chief executives). New principals were appointed at 80% of our participant schools. While the uncertainty makes it difficult to retain a focus on science efforts (Borthwick, Stirling, Nauman, & Cook, 2003; Fishman, 2005), the partnership effort was sustained by tenacious efforts of key DPS leaders and University of Michigan collaborators.

Curriculum

The inquiry science curriculum enacted in Detroit was based on learning outcomes identified by state and national standards, and carefully aligned with the curriculum framework of the Detroit Public Schools.

The curriculum was designed as a series of 8- to 10-week units, incorporating inquiry investigations contextualized by driving questions. Student learning was scaffolded by embedded technology, and students created artifacts to demonstrate understanding and serve as a means for discussion and feedback (Singer et al., 2000). Each unit was revised annually based on feedback from teachers and analyses of student test results. Core content remained essentially unchanged, but the annual revisions allowed the designers to address issues of clarity and practical instructional concerns raised by participant teachers.

Curriculum units examined in this study include:

What Is the Quality of Air in My Community? (Amati, Singer, & Carrillo, 1999). This unit focuses on factors that affect air quality to build understanding of 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.

What Is the Water Like in My River? (Singer, Rivet, Schneider et al., 2000). In the context of exploring local ecology, learners construct an integrated understanding of science concepts such as watersheds, erosion and deposition, and chemistry concepts such as pH and dissolved oxygen.

Why Do I Need to Wear a Helmet When I Ride My Bike? (Schneider & Krajcik, 2002) While exploring the nature of collisions in a real-world context, learners develop an understanding of force, velocity, acceleration and Newton's Laws. Learners also develop strategies for interpreting and visualizing physical phenomena graphically.

Professional Development

The professional development effort was carefully aligned with the project-based inquiry model of the curriculum units. The goal of the professional development effort was to prepare teachers to enact the curriculum in a manner consistent with its underlying theoretical basis while adapting it to classroom circumstances (Krajcik, Blumenfeld, Marx, & Soloway, 1994). Professional development activities included week-long summer institutes, monthly Saturday workshops, teacher discussion groups, online resources, and limited classroom support by graduate students and peer teachers. Workshops and summer institutes made use of data on student learning outcomes, content and pedagogical content concerns and teacher enactment difficulties among other topics, in a process that (Fishman et al., 2003). Teachers new to the program started with a summer institute, and generally participated in the workshops introducing a unit, then selected from among the other professional development opportunities. In addition, each unit contains educative materials for teachers that focus on content, student understanding and potential enactment problems (Schneider & Krajcik, 2002). Professional development and educative curriculum features were continuously redesigned in response to teacher evaluation of PD, student performance, and observations of classroom teaching (Fishman et al., 2003).

Learning Technologies

Approaching standards-based instructional reform through curricula facilitates embedding technologies to support learning goals. Software tools are incorporated into each curriculum unit which are designed to take into consideration the unique characteristics of novice learners via embedded supports. Tools used included Model-It, a graphical multivariate modeling package, eChem, a molecular visualization package and data probes for collecting and analyzing experimental data from student investigations. The technology tools embedded in the curriculum help expand the range of questions that students can investigate, the types of data and information that can be collected, and the types 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 (Jackson, Krajcik, & Soloway, 2000; Krajcik et al., 2000; Wu, Krajcik, & Soloway, 2001).

Methods

Participants

During 3 years of implementation ending with the 2000–2001 school year, 37 teachers in 18 schools participated in our curriculum efforts, involving approximately 5,000 students. Students participated in the two 7th grade and/or the one 8th grade LeTUS curriculum units. Teachers and schools were selected based on

several criteria which initially included adequate technology infrastructure, supportive administration, and equity among schools in access to innovative programs (Marx et al., 2004). In most schools, between one and three teachers selected in a method determined by their building administrator participated in the curriculum with some or all of their classes. These classes generally represented less than half of the school's eligible student body. Participating teachers averaged 11 years of teaching, of which 8 were primarily teaching science; approximately 70% report having a science credential. These figures were slightly lower than district statistics overall. School technology access varied between buildings and over time (Fishman et al., 2004).

Each year of the study involved a gradual scale-up of the effort within the Detroit system, as new school sites and additional teachers were added to the participant pool, building on the initial group of teachers who were developing increasing proficiency in enacting the curriculum (Table 1). Teachers who were added each year received the same summer workshop and in-service support, though responsibility for professional development was shifting from university to district staff. Classroom-level technology assistance provided by university personnel also diminished significantly as the effort scaled.

Measures

The Michigan Educational Assessment Program (MEAP) tests are a statewide standardized assessment aligned with the state objectives for science achievement. MEAP tests are considered high-stakes instruments in Michigan. The aggregate results for schools affect the statutory accreditation of the school and potentially the state funding allocation both directly and through enrollment pressures. The MEAP examines student achievement in reading, writing, mathematics, and science, with a recent addition of social studies. Science tests are given in the 5th grade for elementary school, the 8th grade in middle school, and in the junior year of high school. The passing rate for each subject on the MEAP test is widely publicized and the scores result in considerable political pressure for schools and districts.

The MEAP has been subject to revisions between years, and the threshold scores for the ''passing'' proficiency categories vary with restructuring and political pressures. During the period we have been working in the schools, responsibility for MEAP development and administration has been shifted back and forth between the State Departments of Education and Treasury in response to the influence and priorities of different state governors. For the years in question, the MEAP test was not vertically equated. While not an ideal research instrument, the MEAP is in many ways typical of the general development of high stakes state standardized testing, and the political pressures which influence such instruments. Since distal assessments like the MEAP often determine the fate of reform efforts, they must not be too readily dismissed because of their analytic or structural shortcomings.

The middle school MEAP science tests are administered in late January of the 8th grade year, several months to a year after the students' exposure to various LeTUS curricular units. The test provides information on student understanding in three science content areas (Life, Physical, and Earth) and two areas of science process skills (''Constructing'' and ''Reflecting''), using a combination of approximately 54 multiple choice and 6 free response questions. LeTUS units addressed content in all 5 areas. These five areas are combined to provide an overall science score, as well as a three-level student proficiency category that divides students into

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	1,203
Water	7	1998–1999	11	33	615
		1999–2000	12	35	1,091
		2000–2001	19	58	1,201
Helmets	8	1998–1999	3	6	110
		1999–2000	8	25	750
		2000–2001	11	26	800

two passing categories (''proficient'' and ''novice'') and one non-passing category (''not yet novice''). Cutoff points for the passing categories were adjusted yearly.

Analysis

Because the between-year test variability does not allow for reliable multi-year analysis, we divide the study by cohort group for our analysis. This division also allows us to examine any change in effects from scaling and ongoing development efforts in LeTUS, and provides us an opportunity to replicate initial findings on a second group at increased scale. We obtained MEAP results for the entire Detroit school district for the January 2000 and January 2001 administrations of the 8th grade MEAP science test, corresponding to the first two cohorts of students participating in LeTUS curricula over 3 years of enactment. Cohort I students may have participated in up to three LeTUS units: 7th grade Air Quality in the fall of 1998, 7th grade Water in the spring of 1999, and 8th grade Helmets in the fall of 1999. These students then participated in the administration of the MEAP test in January of 2000. Similarly, Cohort II students may have participated in the same three units the following years, culminating in MEAP administration in January of 2001.

The method of analysis involves a pooled comparison of students who participated in LeTUS curriculum with students in the Detroit Public Schools System who did not. It is important to note that this pooled comparison does not constitute a true control study since many of these students are experiencing other science improvement efforts as part of the NSF supported Urban Systemic Program and other local and state efforts. We are aware of at least 6 other significant instructional interventions in the Detroit system. The comparison is thus between the LeTUS intervention and a combined pool of students receiving no intervention or other interventions in science.

For the purposes of the initial analysis, we chose the most conservative criterion for distinguishing groups. Students were considered to have participated in LeTUS if they completed at least one LeTUS unit at any time during 7th or 8th grade year. We are thus examining whether participating in at least one project unit of 8–10 weeks has an impact on student understanding that persists for up to 1 year after the student completes the project unit. We also examine if the performance exceeds the impact on student understanding that results from regular instruction and other interventions in the remainder of the DPS population.

In determining inclusion in the LeTUS participant sample, ''completion'' of a unit was defined to include only those students for whom we had both start of unit (pre-test) and end-of-unit (posttest or student survey) data. Detroit Public Schools, like many urban districts, has a moderately high amount of student mobility and absenteeism, averaging approximately 20% in our curriculum enactments. This definition of completion therefore excludes students who may have participated in part or all of the curriculum but who missed a testing date, placing them in the district-wide comparison group. This gives us a treatment sample of 760 students in Cohort I and 1,043 in Cohort II, with a comparison group of 8,900 and 8,662 respectively.

In addition to the overall comparison, we conduct similar analyses disaggregating the effect of LeTUS participation by gender and by grade level. Achievement results for boys are of particular interest because of the at-risk nature of this population in Detroit and other urban districts. Because of the highly charged nature of racial achievement reporting and 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, including data on English as a second language learners. Given that the student population of Detroit exceeds 91% African American, we do not feel there would be sufficiently representative samples to allow for a meaningful comparison in any event.

Addressing Potential Sample Bias

While the nature of the ongoing reform effort and the lack of detailed demographic data for non-participant students and teachers made constructing a matched sample impossible, we examined the sample pool in detail to ascertain potential sources of treatment or sampling bias. We address the potential influence of student absenteeism and attrition, biases in student selection due to tracking, and possible school site and participant teacher selection factors.

Given student mobility and absenteeism, we considered the possibility that students dropped out during participation in a LeTUS unit or showed a higher absenteeism rate which excluded them from the LeTUS sample by not being present for a pre- or post-test. This might result in classifying more academically able

students as LeTUS participants. While it was unlikely that this group was sufficiently large to affect the results, we repeated the analysis for each cohort. In the replication, we included all students who had any contact with a LeTUS unit for any length of time, so as to eliminate any attrition effect. This change had no effect on reported results.

While schools were chosen for participation in the study by school district officials who were very mindful of equity issues in providing opportunities for high-quality instruction in all schools in the district, it is important to consider the possibility of unintended selection bias in choosing participant schools. In our first effort to rule out this bias, we repeated the analysis using only LeTUS participant schools. The comparison for each cohort was therefore between LeTUS participant students and non-participants within the same school. As with the analysis of student attrition, the results of this comparison were not substantially different, although a slight increase in effect size (+0.05 SD) was noted for Cohort I.

Within each participant school, it is possible that various issues of student tracking or scheduling may have allowed more academically capable students to participate in LeTUS units. To rule out student selection bias we made use of the 7th grade MEAP administration in mathematics and reading as a proxy for prior academic achievement. Our analysis found a slight positive bias in students 7th grade MEAP scores favoring students who completed at least one LeTUS unit. The largest effect size was 0.03 standard deviations, which is not sufficient to account for our results. This small effect is complicated by the fact that the 7th grade MEAP is administered in January, following completion of the fall 7th grade LeTUS Air Quality unit. Mathematical topics like graphing and data interpretation and reading skills like identifying arguments are covered in the Air Quality unit. The small increase may therefore also be a treatment effect. Given the very small effect size and the additional possibility that the scores were influenced by the fall semester Air Quality unit, we do not find evidence to suggest substantial sampling bias in student selection.

The issue of teacher selection bias in our sample is more problematic. Teachers who choose to participate in reform efforts might tend to be mavericks or show greater levels of commitment to self-improvement, bringing a broader repertoire of techniques to the classroom. Since the nature of our intervention was targeted at teacher practice and measures of prior teacher performance are not available, we have no analytic method to rule out this possibility. However, the nature of the teacher selection process did not lend itself to such a bias since participants were chosen on the basis of schools. Any selection effect would have to be present across multiple unrelated sites with different administrations. Within schools, participant teachers were directed to participate by building and district administrators based on scheduling, balancing professional development opportunities among staff, and other considerations. We characterize this as a ''non-volunteer'' sample (Fishman et al., 2003). In addition, pre- to post-test gains on our curriculum specific tests showed high teacher variability in outcomes, suggesting that if there were a selection bias it would likely be relatively small. Our teacher pool was comparable to the general Detroit teacher population in highest degree earned but was lower than the DPS average in years of teaching experience, with a slightly higher percentage teaching outside of their certification area than the district norm (Marx et al., 2004).

Results

Presenting our findings below, we begin by examining the overall impact of participation in one or more LeTUS curriculum units on 8th grade science MEAP score. This identifies whether participation has any general effect on our distal achievement measure, over and above general Detroit Public School population. Next, we disaggregate the results for each cohort by level of participation, examining whether there is an increased effect for students who participate in LeTUS units in both 7th and 8th grade. We conclude by examining results separated by student gender to determine if there is any differential impact on girls or boys. Each analysis is performed on the first cohort and then replicated on the second, scaled-up cohort.

Overall Achievement

Our first analysis compares the students who completed at least one LeTUS unit with the general pool of DPS students who did not participate in LeTUS. The analysis is replicated for each cohort of students. The total raw MEAP science score is examined, along with the raw score totals for each of the 5 content and process area subscores, using a standard comparison of means. Table 2 shows the results and the relevant effect sizes for the Cohort I and II students. All differences for both cohorts were significant at $p < 0.001$.

Table 2
LeTUS participants compared to DPS population by cohort

	Cohort I			Cohort II		
	LeTUS mean	DPS mean	Eff. Size (std.)	LeTUS mean	DPS mean	Eff. size (std.)
Sample size	760	8,900		1,043	8,662	
Total score	389.16	340.40	0.44*	360.05	320.03	0.37*
Content areas						
Life science	75.09	67.73	0.28*	97.40	85.99	0.32*
Physical science	83.80	73.67	0.36*	52.28	49.31	0.17*
Earth science	66.01	53.15	0.53*	89.55	77.67	0.34*
Process areas						
Constructing	77.25	67.34	0.34*	53.49	47.73	0.29*
Reflecting	87.00	78.52	0.33*	67.33	59.33	0.32*

*Indicates statistical significance (independent samples Student's *t*) at $p < 0.001$.

In Cohort I, students who completed at least one LeTUS unit during 7th or 8th grade significantly outperformed their DPS peers on their overall MEAP science score. Moreover, the difference was not confined to one area of the test. Higher scores were achieved in all three science content areas (Earth, physical, and life science) and both science process skill groups (constructing and reflecting) measured by the science MEAP. The effect sizes are respectable, with participation in at least one unit corresponding to a 14% improvement in total score when compared with the remaining DPS population. This amounts to a standardized effect size of 0.44 standard deviations.[1]

Cohort II represents a substantial upward scaling of the intervention to include more school sites and instructors, with consequently less individualized support for teachers in the classroom. As with the Cohort I group, the second cohort demonstrates significantly higher achievement in all five content and process categories measured by the MEAP science assessment compared with the other students in the district. This amounts to a 13% difference and a standardized effect size of 0.37 standard deviations. Overall and individual category effect sizes are slightly but not appreciably attenuated with the scale-up, with the exception of the physical science content category. The reduced effect size for physical science corresponded to a significant curriculum revision in our 8th grade Helmets unit, with which some teachers reported difficulty. The Helmets unit was particularly focused on physical science concepts.

Pass Rates

The MEAP passing categories are the ''high stakes'' numbers for students and school districts. Two passing categories (''proficient'' and ''novice'') and one non-passing category (''not yet novice'') are defined with cutoffs determined yearly. Because of the threshold nature of determining passing categories, the gains from exposure to LeTUS curriculum are more dramatic when one examines MEAP passing rates for students. Figure 1 compares MEAP passing rates to the general DPS passing rate for both cohorts. Participating in at least one LeTUS unit is associated with a 19% increase in passing rate in Cohort I and a 14% increase for Cohort II. The differences are statistically reliable (Chi Square 117.8 and 103.1, respectively; df = 9,660, 9,704; $p < 0.001$). These gains in MEAP passing rate have very important positive implications for the schools and the district, beyond the learning gains for the individual students.

Grade Level Contributions

Each participant school in the study had both LeTUS participant teachers and other science teachers who were not enacting LeTUS curriculum. Students in either of the middle school cohorts therefore had four possible patterns of exposure to the curricula. They may have participated in LeTUS units in both 7th and 8th grade, they may have only had the opportunity to participate in the two 7th grade units, they may have only participated in the 8th grade unit, or they may not have had access to any of the LeTUS curriculum.

Because 7th grade teachers and students participated in both 7th grade units, it was not possible to disaggregate effects by individual unit. Instead, we further examined the effect of participation in LeTUS

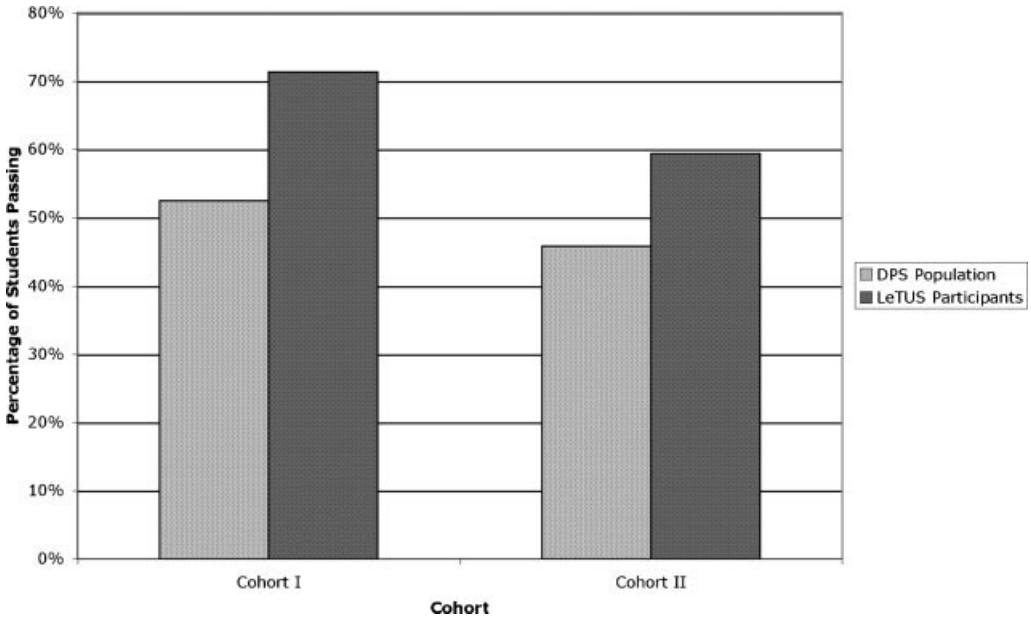

Figure 1. Science MEAP passing rates by LeTUS Participation.

units by disaggregating the contributions to student achievement from participation in LeTUS units at each grade level. Seventh and 8th grade participation were dummy coded and examined along with the first level interaction term using a standard ANOVA model. The results for each cohort are illustrated in Figure 2.

In Cohort II, higher MEAP score performance was associated with both 7th and 8th grade participation independently ($F = 91.7$, 17.5, df = 9,705, $p < 0.001$; interaction $F = 0.15$). Participation in LeTUS 7th grade units was associated with a 37 point greater raw MEAP score compared with DPS peers, while participation in the one 8th grade LeTUS unit was reflected in a 23 point MEAP score difference. As there was no significant grade level interaction, positive differences by grade level were essentially additive. Students who participated in LeTUS in both 7th and 8th grade showed the largest benefit, with a 66 point higher MEAP score.

The results for Cohort I are less clear. In Cohort I, better MEAP score performance was associated only with 8th grade LeTUS participation ($F = 186$, df = 9,660, $p < 0.001$). There was no independent effect for 7th grade participation, nor was there any cross-grade interaction. MEAP score actually declined slightly in absolute points when 7th grade participants were compared with their DPS peers. The 7th graders in the first cohort were the ''pioneer'' group, the first and smallest set of students to participate in any of the projects. There is some indication that selection of this group may have experienced some negative bias, as comparison of their 7th grade mathematics MEAP scores yielded slightly lower scores ($t = 1.74$, df = 9,219, $p < 0.1$). It is unlikely that this slight selection bias is sufficient to account for the difference. We suspect that the weakness is associated with the pilot year of enactment. This is consistent with the results of our curriculum-specific tests, where the first year of 7th grade enactment showed the weakest pre- to post-test gains (Marx et al., 2004).

Gender Differences

The Detroit Public School district-wide MEAP results show statistically significant gender differences between girls and boys achievement on the science assessment. Boys scored 17 points below girls in 2000 ($t = 7.45$, df = 9,544, $p < 0.001$, ES = 0.15) and 11 points below girls in 2001 ($t = 5.01$, df = 9,631, $p < 0.001$, ES = 0.10). These follow the trend that has been widely reported in urban schools with

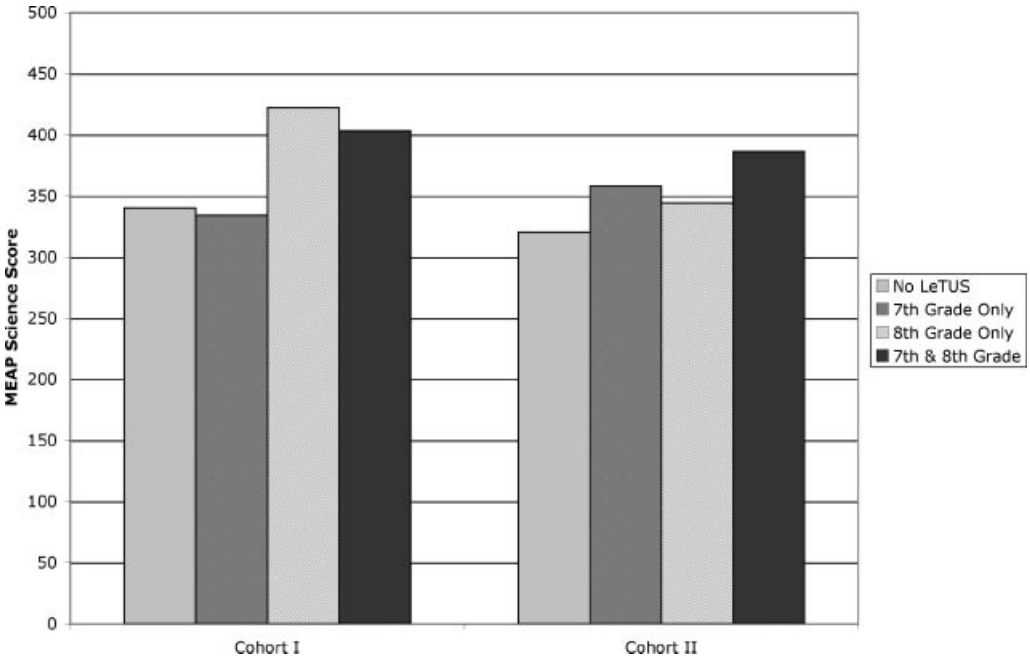

Figure 2. MEAP science scores by grade level participation in LeTUS units.

predominantly African-American populations, where the boys under perform in comparison to the girls (Graham, 2000; Lynch, 2000; Weaver-Hightower, 2003; Witherspoon, Speight, & Thomas, 1997).

Our data, however, suggest that for the population of students in LeTUS program schools, participation in at least one LeTUS unit has an attenuating effect on this gender difference in achievement. Figure 3 shows the interaction of gender with participation in LeTUS curriculum by cohort. In both cohorts there is an apparent reduction in boy-girl achievement differences on MEAP associated with LeTUS participation. The interaction term is marginal for the first cohort of students ($F = 1.90$, df $= 9{,}546$, $p < 0.17$), but reaches conventional statistical reliability for Cohort II ($F = 4.59$, df $= 9{,}633$, $p < 0.05$).

As with the overall score results, the trend toward reducing the performance gap between boys and girls is more dramatic when we examine the MEAP proficiency categories considered ''passing.'' Figure 4 shows MEAP passing categories for each cohort. In each case, there is a substantial difference between students who

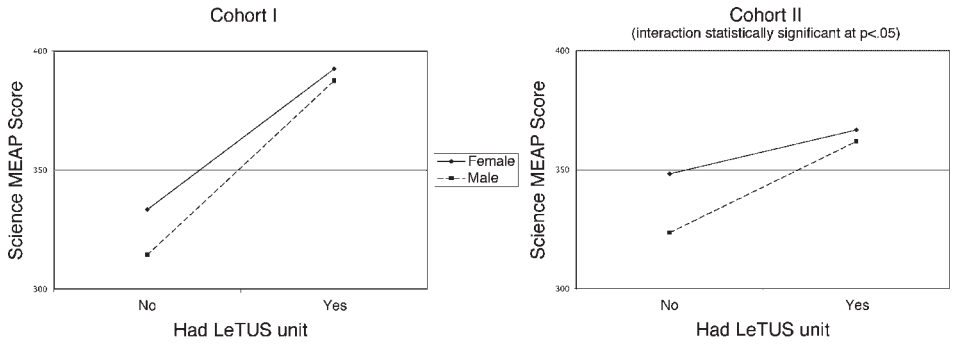

Figure 3. Gender interaction with LeTUS participation.

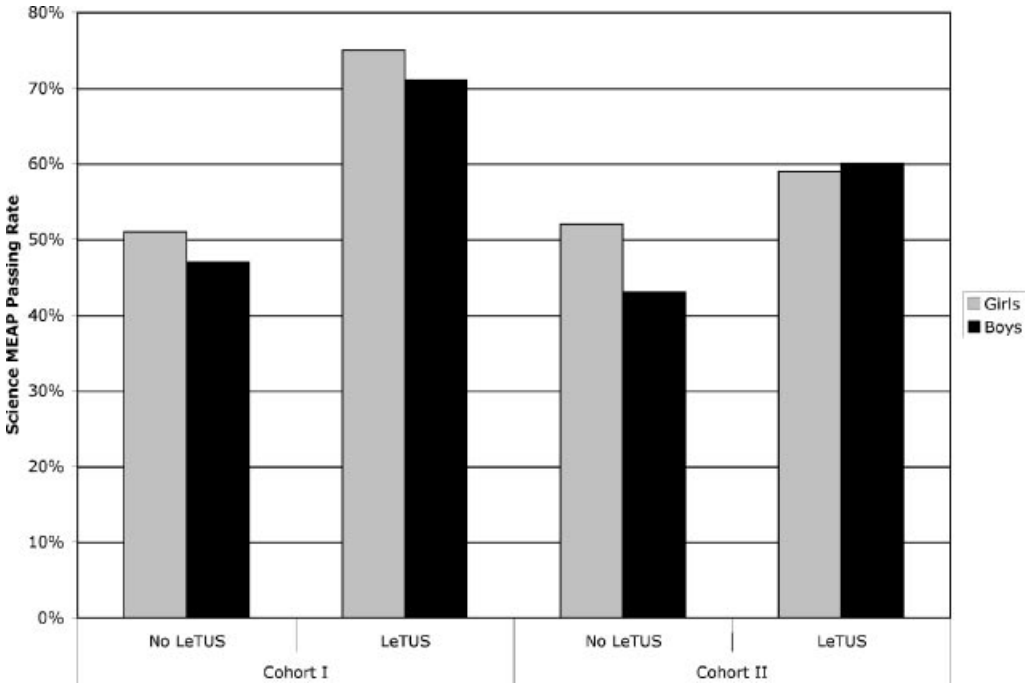

Figure 4. Science MEAP passing rates by gender.

had completed at least one LeTUS unit and those who did not. While both boys and girls benefit from participation, the gender gap between passing rates is attenuated for LeTUS participants in Cohort I. In Cohort II we see that the boys involved in at least one LeTUS unit ''caught up'' and showed no significant difference from the girls. It is noteworthy that for both cohorts, the reduction in the gender gap is achieved by the relatively greater gains by boys rather than attenuated gains by girls.

This is a very positive finding. Standards-based instruction including pervasive technology and project-based units appear to engage at-risk urban male learners, narrowing and closing the gender gap in achievement with their female peers.

Discussion

In an era of tight school budgets and increased pressure for public school accountability, one of the major problems facing researchers engaged in reform is the ability to demonstrate effects from the reform on distal ''high stakes'' achievement tests. Do students really show improved understanding as a result of standards-based practice, at a level sufficient to justify the long-term social and financial commitment required to sustain a reform effort?

Individual Student Achievement

A number of other groups have pursued this question with student-level distal achievement data in conjunction with standards-based reforms. While some of the larger statewide or multi-state studies have found no effect from standards-based teaching after controlling for other factors (e.g., Berends, Chun, Schuyler, Stocly, & Briggs, 2002), other researchers have been able to identify gains associated with the reform efforts and standards-based instruction, but often with relatively small effect sizes between 0.05 and 0.15 (Borman, Hewes, Overman, & Brown, 2002; Kahle, Meece, & Damnjanovic, 1999; Klein et al., 2000). Most of these studies, however, involved statewide or multi-state efforts where researchers did not have information on nor control over the curriculum materials in use. When the nature of the reform allows for the

implementation of specified and well-developed curricula to support other aspects of the reform effort, larger effect sizes have been achieved (Borman et al., 2002; Lynch et al., 2005, Ross et al., 2001).

Our results support and add to this work. The partnership between the Detriot Public Schools and LeTUS allowed us to use curriculum materials as a basis for aligning instruction, professional development, learning technologies, and administrative support within an urban reform framework. This approach of building a highly specified and highly developed inquiry-based curriculum, while harder to adopt across jurisdictional boundaries and therefore less common in the literature, shows strong effect sizes on distal measures of achievement. In the context of the educational challenges of a large urban district, the effects we find from this aligned instructional effort are in many ways quite heartening. The strength of the effect, its replication in two cohorts, and its persistence through a substantive increase in scale strongly supports the conclusion that well-aligned standards-based reform efforts of this type can positively impact urban student results on distal standardized achievement tests.

While the project units were a core component of our effort, they were embedded in the broader context of the ongoing partnership and Detroit's urban systemic initiatives. Attempting to impute the causal contribution from specific elements of a multi-faceted reform effort such as ours is very difficult, if not impossible (Cobb, Confrey, DiSessa, Lehrer, & Schauble, 2003). Teachers' experience of policy and administrative support, professional development, and enactment of inquiry instruction through LeTUS are all intertwined with student experiences of inquiry, technology use, collaboration, and other features of the effort. Revision of the curriculum and professional development programs in response to feedback adds an additional layer of complexity. We do not claim nor do our data support a conclusion that inquiry science units alone will enhance achievement. Rather, the results indicate that an effort incorporating and aligning the best practices in curriculum, professional development, and learning technology in the context of a systemic reform can achieve substantive results on politically important measures.

Work by others engaged in reform efforts suggests that several years are required before the effects of the reform can be discerned (Borman et al., 2002; Fuhrman, 2001). Our more limited results from the first year of enactment with Cohort I are consistent with this literature. This is further supported by the progressive increases in pre- to post-test effect sizes between years in the curriculum-specific tests that we reported previously (see Marx et al., 2004). In the initial implementation, teachers are undergoing an intense period of learning new content and new pedagogy within a short space of time. Students, too, must adapt to new expectations and teaching styles, which may increase the management burden placed on teachers. At the same time, the first year of enactment poses additional technical challenges. Learning technology may be unfamiliar, and computer support infrastructure may not be in place so that the technology is not functional when needed. In our effort, the first years of implementation also provided significant feedback for curriculum revisions, enabling designers to improve curriculum materials in support of classroom enactment. Taken together, these factors and our results support prior work in maintaining that even highly aligned multi-component efforts at implementing standards require several years of enactment experience before student achievement results show consistent improvement on distal measures. In the unstable administrative environment common in many urban districts, it takes considerable fortitude on the part of key district leaders and collaborators to sustain the efforts until results are achieved. We were fortunate to have been able to sustain the effort by maintaining a strong partnership based on mutual respect and common goals. We believe this collaboration was crucial to the effort (Fishman, 2005).

An additional finding from our work is that student experience with project-based inquiry units has an independent and cumulative effect on distal achievement scores. Repeated exposure to standards-based instruction supported by highly specified and developed curriculum leads to additional learning gains, at least when science inquiry is pursued at multiple grade levels. This supports work by Lee et al. (2006), who also found cumulative effects from continuing inquiry science instruction across grade levels on proximal curriculum-aligned assessments. This indicates that the more inquiry science instruction we are able to provide to students during their schooling, the larger the learning growth we will expect to see in achievement.

Our results show some differences in science content effect sizes between Cohorts I and II. Each LeTUS curriculum unit is built around a driving question for student inquiry. The driving question in turn leads students to different science content areas as the unit progresses. While each unit incorporates science process skills and content from several MEAP areas, there is a content emphasis to units. The Water Quality unit, for

example, has the most earth science content, while the Helmets unit focuses more heavily on physical science. In Cohort II, we speculate that the drop in effect size for the Earth Science content area may be associated with increased difficulty completing the late spring Water Quality unit as a result of scheduling and special programs at the end of the school year. Teachers reported considerable difficulty finishing the unit. We mention previously that the similar effect size decline in Physical Science corresponded with a major revision to the Helmets unit, which some teachers found more difficult to teach. While the results still demonstrate positive effects in all content areas, these observations suggest that student learning can be disrupted by scheduling or enactment difficulties.

Gender Equity

The research literature has frequently identified a pattern of lower measured achievement in urban minority boys when compared with their female peers (Graham, 2000; Lynch, 2000; Weaver-Hightower, 2003; Witherspoon et al., 1997). These are generally attributed to a number of contextual factors that limit or discourage learning for boys (Graham, 2000; Graham, Taylor, & Hudley, 1998; Harry & Anderson, 1994; Rascoe & Atwater, 2005). Examination of the 8th grade science MEAP scores in the Detroit population demonstrates a similar gap in achievement by gender in our comparison pool, with boys scoring significantly lower than girls. Participation in at least one LeTUS unit resulted in attenuation of this gender equity issue. The strength of the effect in our results is substantial; by the second cohort of students, LeTUS participation had a sufficiently large effect to essentially eliminate the boy-girl difference in science achievement, allowing this often at-risk group to catch up with their peers. This finding is congruent with previous work by Kahle et al. (2000), who similarly demonstrated greater gender equity in classes taught by teachers who had participated in extensive professional development to encourage standards-based instruction.

Kahle et al. (2000) also identified a strong positive relationship between student attitudes and achievement for urban minority males. Work by Graham et al. (1998) shows that this group experiences a substantial decline in attitudes toward academic achievement during middle school. Taken together, the studies suggest that the gap in achievement outcomes may be due in part to the decline in student attitudes experienced by urban male students. We have seen in some of our work on motivation that participation in our science inquiry projects attenuates the loss of science motivation that occurs as our urban Detroit population progresses through middle school (Blumenfeld, Soloway, Krajcik, & Marx, 2002). Additionally, our survey data from LeTUS work indicate that boys in our curriculum show substantially higher levels of technology interest than girls. Technology interest in boys may contribute to the use of more effective learning strategies through the use of technology in LeTUS units, and in turn improve achievement (Blumenfeld, Middleton, Geier, & Marx, 2004). We speculate that in contrast to traditional instruction, the variety of experiences provided by LeTUS curricula, such as technology use, inquiry, peer collaboration and interaction with phenomena are more appealing and therefore result in more positive attitudes and greater participation in learning. This effect may be stronger for students who are not engaged by traditional instruction, including urban boys.

Ongoing Work

When considering the results of this study, it is important to be mindful of its methodological limitations. While we have made efforts to rule out potential issues with selection bias at several levels, the nature and limits of our data do not allow us to rule out all such influences, particularly with regard to individual teacher levels of commitment and pedagogical ability. As the LeTUS efforts in Detroit continue to expand into more schools and classrooms, such selection bias issues will become increasingly moot. One primary area of interest for future work therefore is following the progress of additional cohorts of students as the effort continues to scale up in the district, and as responsibility for managing and supporting the effort is transferred to district staff. Demonstrating the ability to sustain improvements in student achievement scores on distal measures for large number of urban students while fading outside support is necessary to justify the considerable investment educational leaders must make to enact and sustain standards-based science reform efforts.

Scaling up of the student population engaged in inquiry science is only one dimension of our scaling effort, however. The results of this study suggest that it is also important to scale up the amount of exposure

each student has to standards-based inquiry projects, by expanding the number of different curriculum units available as part of the middle school program. To that end, as we are scaling up total student numbers we are also adding additional curriculum units with aligned professional development and learning technologies. If the effect of participating in inquiry projects is in fact cumulative, we would hope to see additional achievement gains for students who experience a greater number of units during their middle school years.

We wish to give special acknowledgement to the faculty, staff, and administration of the Detroit Public Schools, our partners in this effort, and to Deborah Peak-Brown in her role as DPS liaison to the LeTUS effort contributed thousands of hours to the effort to promote, scale-up, and document the curriculum efforts of the partnership. Funding for these projects was made possible by the Spencer Foundation in a grant for ''Technologies to Enable Inquiry: The Influences on Student Learning and Motivation'' and by the National Science Foundation under the following programs: REPP (REC-9720383 REC-9725927 REC-9876150) and USI (ESR-9453665).

Note

[1]A raw score difference between LeTUS students and the comparison pool of 10 points corresponds to 1 multiple choice item or half credit on a free response question. As an example, the approximately 50 point difference in the Cohort I results corresponds to answering 5 additional multiple choice questions correctly (out of 54) or receiving full credit for 2.5 additional free response questions (out of 6).

References

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

Barton, P.E. (1999). Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education. New Jersey: Policy Information Center, Mailstop 04-R, Educational Testing Service, Rosedale Road, Princeton, NJ 08541-0001; e-mail: pic@ets.org; World Wide Web: http://www.ets.org/research.

Berends, M., Chun, J., Schuyler, G., Stockly, S., & Briggs, R.J. (2002). Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district. Santa Monica: RAND.

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

Blumenfeld, P., Fishman, B.J., Krajcik, J., 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(3), 149–164.

Blumenfeld, P., Middleton, M., Geier, R., & Marx, R. (2004). Connecting standards based instruction, motivation, and achievement in urban middle school science classrooms. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Blumenfeld, P., Soloway, E., Krajcik, J., & Marx, R. (2002). Technologies to Enable Inquiry: The Influences on Student Learning and Motivation: Spencer Foundation.

Blumenfeld, P.C., 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(3&4), 369–398.

Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S., & Center for Research on the Education of Students Placed At Risk Baltimore MD. (2002). Comprehensive School Reform and Student Achievement: A Meta-Analysis (No. CRESPAR-R-59). Maryland: Publications Department, CRESPAR/Johns Hopkins University, 3003 N. Charles Street, Suite 200, Baltimore, MD 21218. For full text: http://www.csos.jhu.edu.

Borthwick, A., Stirling, T., Nauman, A.D., & Cook, D.L. (2003). Achieving successful school-university collaboration. Urban Education, 38(3), 330–371.

Bouillion, L.M., & Gomez, L.M. (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.

Bruckerhoff, C. (1997). Lessons learned in the evaluation of statewide systemic initiatives. Curriculum Research and Evaluation, I. Chaplin, CT.

Clune, W. (1998). Research monograph No. 16: Toward a theory of systemic reform: The case of nine NSF statewide systemic initiatives. Madison, WI: National Institute for Science Education.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. Educational Researcher, 32(1), 9–13.

Cohen, D.K., & Ball, D.L. (1999). Instruction, capacity, and improvement (CPRE Research Report Series RR-43). Philadelphia, PA: Consortium for Policy Research in Education.

Corcoran, T., Christman, J.B., Consortium for Policy Research in Education & Philadelphia PA., & Research for Action Inc. Philadelphia PA. (2002). The Limits and Contradictions of Systemic Reform: The Philadelphia Story. Pennsylvania: Consortium for Policy Research in Education, Graduate School of Education, University of Pennsylvania, 3440 Market Street, Suite 560, Philadelphia, PA 19104-3325. Tel: 215-573-0700; Fax: 215-573-7914; Web site: http://www.cpre.org.

Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving science inquiry with elementary students of diverse backgrounds. Journal of Research in Science Teaching, 42(3), 337–357.

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(3&4), 391–450.

Elmore, R.F. (1996). Getting to scale with good educational practice. Harvard Educational Review, 66(1), 1–26.

Fishman, B. (2005). Adapting innovations to particular contexts of use: A collaborative framework. In C. Dede, J. Honan, & L. Peters (Eds.), Scaling up success. San Francisco, CA: Jossey-Bass.

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(6), 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(1), 43–76.

Fortus, D., Dershimer, R.C., Krajcik, J., Marx, R.W., & Mamlok-Naaman, R. (2004). Design-based science and student learning. Journal of Research in Science Teaching, 41(10), 1081–1110.

Fraser-Abder, P., Atwater, M., & Lee, O. (2006). Research in urban science education: An essential journey. Journal of Research in Science Teaching, 43(7), 599–606.

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

Graham, S. (2000). Academic achievement and ethnic minority students: Constraints and opportunities. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Graham, S., Taylor, A., & Hudley, C. (1998). Exploring achievement values among ethnic minority early adolescents. Journal of Educational Psychology, 90, 606–620.

Hancock, C., Kaput, J.J., & Goldsmith, L.T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. Educational Psychologist, 27, 317–364.

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 capital to the classroom: Standards-based reform in the states. 100th Yearbook of the National Society for the Study of Education (Part II). Chicago: University of Chicago Press.

Harry, B., & Anderson, M. (1994). The disproportionate placement of African American males in special education programs: A critique of the process. Journal of Negro Education, 63, 602–619.

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. Hillsdale, NJ: Erlbaum.

Kahle, J.B. (1998). Reaching Equity in Systemic Reform: How Do We Assess Progress and Problems? Research Monograph (NISE-9). Wisconsin: National Institute for Science Education, University of Wisconsin-Madison, 1025 W. Johnson Street, Madison, WI 53706. Tel: 608-263-9250; Fax: 608-262-7428; e-mail: niseinfo@macc.wisc.edu. For full text: http://www.wcer. wisc.edu/nise/Publications.

Kahle, J.B., Meece, J.L., & Damnjanovic, A. (1999). A pocket panorama of Ohio's systemic reform 1999 [Brochure]. Oxford, OH: Miami University.

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(9), 1019–1041.

Kim, J.J., Crasco, L.M., Smith, R.B., Johnson, G., Karantonis, A., Leavitt, D.J. (2001). Academic Excellence for All Urban Students: Their Accomplishment in Science and Mathematics. Urban Systemic Initiatives. Massachusetts.

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

Krajcik, J.S., Blumenfeld, P., Marx, R., & Soloway, E. (1994). A collaborative model for helping middle grade science teachers learn project-based instruction. Elementary School Journal, 94(5), 483–497.

Krajcik, J., Blumenfeld, P., Marx, R., & Soloway, E. (2000). Instructional, curricular, and technological supports for inquiry in science classrooms. In J. Minstell & E. Van Zee (Eds.), Inquiry into inquiry: Science learning and teaching. Washington: American Association for the Advancement of Science Press.

Krajcik, J., Blumenfeld, P., Marx, R.W., Bass, K.M., Fredericks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. Journal of the Learning Sciences, 7(3&4), 313–350.

Lee, O., Buxton, C., Lewis, S., & LeRoy, K. (2006). Science inquiry and student diversity: Enhanced abilities and continuing difficulties after an instructional intervention. Journal of Research in Science Teaching, 43(7), 607–636.

Lee, O., & Luykx, A. (2004). Science education and student diversity: Synthesis and research agenda. A monograph supported by the Center for Research on Education Diversity, and Excellence (CREDE) at the University of California at Santa Cruz and the National Center for Improving Student Learning and Achievement (NCISLA) in Mathematics and Science at the University of Wisconsin in Madison.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in elementary science instruction with nonmainstream students. American Educational Research Journal, 42(3), 411–430.

Lynch, S.J. (2000). Equity and science education reform. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers, 10 Industrial Avenue, Mahwah, NJ 07430.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. Journal of Research in Science Teaching, 42(8), 912–946.

Marx, R., Blumenfeld, P., Krajcik, J., Fishman, B., Soloway, E., Geier, R., & Tal, T. (2004). Inquiry based science in the middle grades: Assessment of learning in urban systemic reform. Journal of Research in Science Education, 41(10), 1063–1080.

Marx, R.W., Blumenfeld, P., Krajcik, J.S., & Soloway, E. (1997). Enacting project-based science. Elementary School Journal, 97(4), 341–358.

Marx, R.W., Blumenfeld, P., Krajcik, J.S., & Soloway, E. (1998). New technologies for teacher professional development. Teaching and Teacher Education, 14(1), 33–52.

Marx, R.W., & Harris, C.J. (2006). No child left behind and science education: Opportunities, challenges, and risks. Elementary School Journal, 106(5), 467–477.

Meyer, R.H. (1999). Value-added indicators. Paper presented at the Evaluation of Systemic Reform in Mathematics and Science: Fourth Annual NISE Forum, Madison, WI.

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(4), 469–498.

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., Phelps, S., Kyle, T., & Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. Journal of Research in Science Teaching, 43(5), 467–484.

Rascoe, B., & Atwater, M.M. (2005). Black males' self-perceptions of academic ability and gifted potential in advanced science classes. Journal of Research in Science Teaching, 42, 888–911.

Rodriguez, A.J. (2001). From gap gazing to promising cases: Moving toward equity in urban education reform. Journal of Research in Science Teaching, 38(10), 1115–1129.

Ross, S.M., Sanders, W.L., Wright, S.P., Stringfield, S., Wang, L.W., & Alberg, M. (2001). Two- and three-year achievement results from the Memphis restructuring initiative. School Effectiveness and School Improvement, 12(2), 323–346.

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(5), 369–393.

Schneider, R.M., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. Journal of Science Teacher Education, 13(3), 221–245.

Singer, J., Marx, R.W., Krajcik, J., & Clay-Chambers, J. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. Educational Psychologist, 35(3), 165–178.

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

Thomas, R., Woods, P., Hillman, S., & Ross, S.M. (2002). The Detroit Public Schools Michigan Department of Education CSRD Grant Funded Comprehensive School Reform Demonstration (CSRD) Models, 1998–1999 through 2000–2001. A Joint Collaborative Preliminary Evaluation. Detroit, MI: ERIC Document Reproduction Service No. ED464981.

Tobin, K., Roth, W.-M., & Zimmermann, A. (2001). Learning to teach science in urban schools. Journal of Research in Science Teaching, 38(8), 941–964.

Upton, J., & Supovitz, J. (1996). Measuring Student Impact in the Context of Statewide Education Reform. Paper presented at the Annual Meeting of the American Educational Research Association, New York NY.

Warren, B., Ballenger, C., Ogonowski, M., Rosebury, 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.

Weaver-Hightower, M. (2003). The ''Boy Turn'' in research on gender and education. Review of Educational Research, 73(4), 471–498.

Witherspoon, K.M., Speight, S.L., & Thomas, A.J. (1997). Racial identity attitudes, school achievement, and academic self-efficacy among African-American high school students. Journal of Black Psychology, 23, 344–357.

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