Improving Science Achievement at High-Poverty Urban Middle Schools

ALLEN RUBY
Center for Social Organization of Schools, Johns Hopkins University, Baltimore, MD 21218, USA

Received 14 December 2005; revised 28 April 2006; accepted 12 May 2006

DOI 10.1002/sce.20167
Published online 20 July 2006 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: A large percentage of U.S. students attending high-poverty urban middle schools achieve low levels of science proficiency, posing significant challenges to their success in high school science and to national and local efforts to reform science education. Through its work in Philadelphia schools, the Center for Social Organization of Schools at Johns Hopkins University developed a teacher-support model to address variation in science curricula, lack of materials, and underprepared teachers that combined with initial low levels of proficiency block improvements in science achievement. The model includes a common science curriculum based on NSF-supported materials commercially available, ongoing teacher professional development built around day-to-day lessons, and regular in-class support of teachers by expert peer coaches. One cohort of students at three Philadelphia middle schools using the model was followed from the end of fourth grade through seventh grade. Their gains in science achievement and achievement levels were substantially greater than students at 3 matched control schools and the 23 district middle schools serving a similar student population. Under school-by-school comparisons, these results held for the two schools with adequate implementation. Using widely available materials and techniques, the model can be adopted and modified by school partners and districts.

Periodicals, Inc. *Sci Ed* 90:1005-1027, 2006

INTRODUCTION

U.S. urban schools serving high-poverty, high-minority populations face great challenges due to the characteristics of their neighborhoods, student backgrounds, teacher preparation,

The opinions expressed herein are those of the author and do not necessarily reflect the views of the grant sponsors.
Correspondence to: Allen Ruby; e-mail: aruby@csos.jhu.edu
Contract grant sponsor: Interagency Education Research Initiative (IERI 84.305W).
Contract grant number: R305W020003.
Contract grant sponsor: National Science Foundation.
Contract grant sponsor: Institute of Education Sciences, U.S. Department of Education. Contract grant sponsor: National Institute of Child Health and Human Development.
This paper was edited by former Editor Nancy W. Brickhouse.

and district- and school-level resources. These schools show large differences in student school experiences and student outcomes versus low-poverty schools and even high-poverty nonurban schools (Lippman, Burns, & McArthur, 1996).

These differences are reflected in middle grades' student achievement in science. Results from the 1996 and 2000 U.S. National Assessment of Educational Progress (the NAEP is the only ongoing representative survey of U.S. student achievement) show that student science achievement significantly worsens under the conditions of inner-city middle schools (O'Sullivan, Lauko, Grigg, Qian, & Zhang, 2003). In the 2000 NAEP, 39% of all eighth-grade students were rated Below Basic in science (with another 32% scoring at Proficient or better). However, this percentage reached 70% of students attending high-poverty inner-city schools as seen when scores are broken out for Black and Hispanic status and lower socioeconomic status (with only another 6-12% at Proficient or better). Middle schools cannot take all the blame for these results, as the NAEP shows only slightly better results for fourth graders, but they also cannot receive any credit for offsetting the low levels of science achievement among poor minority elementary students. The recent international 2003 TIMSS study also documents the poorer science performance of middle grade students attending schools with higher concentrations of students with lower socioeconomic status not only in the United States but also around the world (Martin, Mullis, Gonzalez, & Chrostowski, 2004).

The status of science achievement among students in these schools poses significant challenges both to individual student's academic careers and to the success of educational policies focused on improving science education. Students achieving at Below Basic are not prepared for the more demanding content and less forgiving grading policies of high school science courses, especially challenging college preparatory classes. They are unlikely to be prepared for onging national and state efforts to advance science education through the adoption of more rigorous science standards and benchmarks, curricula, and assessments. As they have not yet mastered what is now expected of them in their earlier grades, they may be less able to make adequate progress toward meeting the expectations for their future grades. As a result, successful science education reform must address those students already considered behind in their elementary and early middle grades.

In this paper I describe a whole-school intervention to improve science teaching and learning and its impact on one cohort of middle grades students over 3 years at three inner-city schools in Philadelphia, PA. The intervention focuses on improving classroom instruction through a combination of providing a complete day-to-day science curriculum with intensive and ongoing teacher support in its use. The principle behind this intervention is that a well-taught, academically rigorous, and motivating science curriculum will improve instruction for all students in a school leading to greater learning. The caveat is that under conditions faced by schools serving poor minority students, attention must be kept focused on making available the day-to-day lessons, the materials to teach them, and the ongoing teacher support for their implementation. The need for this approach in such schools is described next.

OBSTACLES TO IMPROVING SCIENCE ACHIEVEMENT

The Center for Social Organization of Schools of Johns Hopkins University began working with urban middle schools serving high-poverty, high-minority student populations in Philadelphia, PA, in the mid-1990s. We identified a set of constraints concerning curriculum and teacher preparation that needed to be addressed in order to directly improve science classroom instruction and learning.

The primary science curricula constraints were the lack of consistency within the school and the lack of the means to implement the curriculum. In many cases, the curriculum varied

between teachers in the same grade. As a result, students in the same grade covered different content and teachers were not in a position to support one another. The in-grade variation in curriculum led to problems in the between-grade curriculum. In some cases, the same content was covered in several grades. Upper grade teachers were not sure what content their new students had covered and correctly assumed that they had different preparation. As a result, upper grade teachers often did not build on what had been taught in the lower grades.

At the time, the district had a policy of site-based curriculum development guided by district-set standards (School District of Philadelphia, 1995). However, when time, resources, and leadership were not available to faculty (even for regular science faculty meetings), schools did not establish a science curriculum and as a result individual teachers set their own. Promotion of thematic curricula within small learning communities inside the schools also created in-school variation, especially as lack of resources and teacher turnover often led to underdeveloped curricula not connected to major concepts or the district's standards (Christman, 2001).

The second key curriculum constraint was the lack of the common means to implement a curriculum. The lack of day-to-day lessons, materials, equipment, and teacher understanding of the curriculum led to a reduction in science content instruction. Teachers often fell back on the materials they had regardless of their quality further contributing to the variation in what was being taught. In schools that had established a science curriculum, the lack of the means for all faculty to implement it led to a de facto variation in quality and topic coverage. Even where curriculum materials, such as textbooks, kits, or individual lessons, had been obtained for all, teachers often picked and chose among them for the parts they felt comfortable teaching.

Many teachers did not feel qualified to teach middle school science. The majority of them were certified as K-6 teachers but allowed to teach middle school under a district exemption (Useem, Barends, & Lindermayer, 1999). The lack of background in science often led to reduced coverage of content, especially physical science, that was seen as difficult. Coupled with this was a lack of in-service professional development necessary to learn the materials available. Under a National Science Foundation grant known as the Urban Systemic Initiative, the district was training science leaders for each school who were to be responsible for training the other teachers. The approach succeeded in creating a leadership corps and pockets of improved instruction but a lack of time and resources prevented the envisioned turn around training and the widespread schoolwide implementation of a reform science program (Blanc & Ballenger, 1999). The continued lack of teacher preparation contributed to variation in curriculum as teachers taught what they knew. It also contributed to a high rate of teacher turnover not only as science teachers left the school but also as they stayed but switched from teaching science to subjects they felt more prepared to teach (Ruby, 2002). Such turnover impeded the creation of a single curriculum, with all teachers prepared to teach it.

These conditions were not unique to the district and their occurrence has been documented across the country. The lack of a coherent, within-school curriculum has been found in other districts serving high-poverty minority students (Smith, Smith, & Bryk, 1998). Students in these districts are commonly taught science by some of the nation's least prepared and least experienced teachers (Lankford, Loeb, & Wyckoff, 2002; National Center for Education Statistics [NCES], 1999; Neild, Useem, Travers, & Lesnick, 2003) who as a result often provide less capable instruction (Grossman, 1990). Lack of preparation often stems from a lack of formal education, lack of teaching experience, and/or a lack of ongoing relevant professional development. Lack of formal education occurs with undercertified, uncertified, and some alternatively certified teachers as well as teachers teaching outside their field

(Darling-Hammond, 1992, 2000; Ingersoll, 1996, 1999; Neild, 2001). For example, in the 1999-2000 school year, uncertified teachers made up 25% of Philadelphia's teachers at the highest poverty middle schools (Neild, 2001). Lack of teaching experience fostered by high teacher turnover contributes to less successful instruction (Betts, Rueben, & Danenberg, 2000), the loss of teachers with higher qualifications as they are the most likely to leave (Lankford et al., 2002), and blocks attempts to improve instruction and learning (Bodilly, 1998; Fullan, 1990; Schaffer, Nesselrodt, & Stringfield, 1997; Useem, Christman, Gold, & Simon, 1997). While national surveys show average teacher turnover rates of 12%, rates for high-poverty, high-minority schools are higher (Ingersoll, 2001; NCES, 1997). In Philadelphia's high-poverty schools, 46% of middle school teachers were in their first 2 years of teaching in 1999-2000 (Neild, 2001).

High-quality professional development provided in a consistent and sustained fashion improves science teachers' instruction (Cohen & Hill, 2001; Desimone, Porter, Garet, Yoon, & Birman, 2002) and so offers a means to address teacher underpreparedness. However, teacher surveys show a lack of certain professional development opportunities such as mentoring and teacher networks, low intensity and unsustained formal training, and a failure to focus on the features that make professional development effective (NCES, 1999; Porter, Garet, Desimone, Yoon, & Birman, 2000). The quality of professional development has been found to vary among teachers within the same school and from year to year. Schools and districts are often unable to implement effective professional development and lack the resources to provide it to all their teachers (Porter et al., 2000).

Along with these obstacles to student achievement, ours and others' work has also identified positive resources that can support improved science instruction in these schools. Foremost is a high level of student interest in science (Wenner, 2003). Building on this interest often requires addressing the diverse cultural and ethnic backgrounds of the students. This background offers intellectual resources and examples from daily life that can be tapped to help students understand science concepts (Lee, 2003; Warren, Ballenger, Ogonowski, Rosebery, & Hadicourt-Barnes, 2000), can be built upon to develop student ownership in their learning (O'Neill & Barton, 2005), and can be stimulated by the recent reform methods of science instruction (Cuevas, Lee, Hart, & Deaktor, 2005). At the same time, the differences in backgrounds combined with the lower socioeconomic status of the students often require teachers to initially provide more explicit instruction both on the expectations for students (e.g., the level of work to be done, the rules of discourse) and on the methods of science that may not have been learned at home (Baxton, Carolone, & Carlone, 2005; Lee, 2003). Through this explicit instruction, teachers are able to help students take on a greater role in their own learning.

Another resource is those science teachers who have learned to successfully work within the constraints of these schools and to tap the resources of their students. By the nature of the profession, these teachers are often isolated in their classrooms. A number of techniques are available to give these teachers greater opportunity to pass their knowledge on to other teachers by supporting individual teachers, for example, through mentoring and co-teaching, or influencing the entire faculty and direction of the science program, e.g., through regular science faculty meetings or taking on science leadership roles (Rhoton & Bowers, 2002; Spillane, Diamond, Walker, Halverson, & Jita, 2001; Tobin, Roth, & Zimmerann, 2001).

The resources available in urban middle schools serving high-poverty, high-minority students are often overwhelmed by the lack of preparation of the majority of the student body and teaching corps. Currently, these schools are being called upon to prepare their students for challenging state and national science standards (Lippman et al., 1996; NCES, 1992), but without adequate supports, their teachers will continue to lack the resources and skills necessary to both help students attain on-grade status and succeed at more difficult

standards-based instructional programs (National Research Council, 2000). Our conclusion was that individual schools required external support (from the district or an outside partner) focused on these classroom-level obstacles if science instruction and achievement were to improve.

THE TALENT DEVELOPMENT MODEL

Over the past 10 years, the Center for Social Organization of Schools has been developing a teacher-support model for high-poverty secondary schools as an integral part of the Talent Development (TD) Middle Grades reform model (Balfanz, Ruby, & Maclver, 2002). The aim is to create a teacher-support system that addresses the reality of large numbers of poorly prepared teachers facing difficult teaching conditions, including a lack of curriculum resources and underprepared students, resulting in high teacher turnover. The goal is to ensure that a school's entire science faculty, including new, inexperienced, and alternatively certified teachers, as well as veterans, can provide standards-based academic science instruction daily while improving their own knowledge of subject matter and instructional pedagogy. The TD model is based in the on-the-ground findings that urban teachers need both a detailed curriculum (including the required materials) and the professional development in its use, and the research literature that has shown professional development is most effective when it is long term, school based, collaborative, focused on students' learning, addresses specific teaching practices, and linked to curricula (Desimone et al., 2002; Hiebert, Gallimore, & Stigler, 2002). The model provides four tiers of support: (1) an implementable curriculum, (2) ongoing intensive teacher professional development, (3) in-classroom support from peer coaches, and (4) mechanisms to foster and sustain changes in science instruction, including opportunities for science teacher cooperation and development of teacher leaders.

An implementable curriculum means a common science curriculum within grades accompanied by the day-by-day lessons needed to implement it, including materials and equipment that meet the district and state science standards. This is not a scripted approach to teaching but one that provides a foundation of academic lessons. Teachers' time is freed from having to plan every day's lessons and scrounge for equipment. Inexperienced teachers can use this time to focus on learning classroom management and pedagogy, while experienced teachers can modify the curriculum to class interests. Same grade teachers are able to learn and work collaboratively when all are using the same lessons. The TD model uses science curricula developed with the support of the National Science Foundation including the Full Option Science System (FOSS) developed by the Lawrence Hall of Science, University of California—Berkeley, and Science and Technology for Children (STC) developed by the National Science Resources Center. These science materials focus on depth of understanding of a topic rather than breadth over multiple topics. Built around hands-on activities but requiring significant amounts of student planning and analysis, these materials allow students who may be behind in not only science but also reading and math to readily take part in lessons and learn the concepts and vocabulary.

Outside the classroom professional development is provided through a multiyear sequence of monthly workshops that is grade, curriculum, and content specific for which teachers can earn graduate credit. During the first year, the workshops focus on the materials and lessons teachers will be teaching in the next month, the content knowledge behind the lessons, and the pedagogical techniques to use while teaching them. This focus on what will be taught in the next lessons distinguishes the model's professional development from the traditional generic workshops that have failed to change teacher practices and student achievement (Killon, 1999; McLaughlin & Oberman, 1996). In later years, the focus shifts to deepening teacher content knowledge, increasing student-generated inquiries, and

pedagogic challenges. At the same time, the basic professional development continues to be provided for new teachers in response to teacher mobility.

In-classroom support available on a weekly basis over the entire school year is provided by expert peer coaches familiar with the curriculum and middle grades' pedagogy. Teachers and classes vary in their abilities and needs. Peer coaches see the teachers and classes in action and help teachers adapt the professional development and materials to the specific needs of their classes using such techniques as model teaching, co-teaching, and critical observation with confidential feedback. In this way, they help the teacher customize the science reform to their class and their own needs in a way that maintains its desired impact. For example, teachers may not be comfortable with use of hands-on activities done in cooperative groups because they themselves lack the skills to manage this type of work or their students cannot successfully work together. In response, a teacher may turn to demonstrating the activities, which solves their own classroom management concerns but frustrates the purpose of using a hands-on centered curriculum. A peer coach in the classroom sees why the teacher has turned to demonstrations. They can offer the teacher specific responses to the problem, such as explicit teaching of social skills students need to work in groups, strict student roles within groups that offer the teacher more control, or better teacher preparation and time management to avoid lags during class, which will allow the use of student activities. As both teacher and students become accustomed to hands-on instruction, the peer coach can advise the teacher on reducing the structure and increasing the role of students.

In-school leadership and supporting structures are needed to maintain the sustainability of the instructional changes. Peer coaches' knowledge of the science faculty and the curriculum, and their ability to work with the school administration allows them help establish such structures as regular science faculty meetings, times for teachers to observe others' teaching, and mentoring of new teachers. These structures allow teachers to support one another, share ways to improve instruction, and help induct new teachers. Such structures built into the school schedule (versus remaining informal) are easier to maintain and receive greater support from school administration. Peer coaches are in position to identify excellent teachers who can become the trainers for new science teachers. These teachers receive additional professional development in how to train new teachers, then act as co-trainers with the coach, and finally take the lead in training workshops.

Both our case studies and a growing body of literature indicate that this four-tiered model of teacher support, with its emphasis on curriculum and implementation-focused professional development, has a positive impact on classroom instruction by increasing the frequency with which standards-based lessons are effectively implemented in classrooms (Balfanz, Maclver, & Ryan, 2002; Cohen & Hill, 2001; Desimone et al., 2002; Kennedy, 2002). In addition, several similar university-based programs note these results with their curriculum and instruction-focused programs. For example, the Center for Learning Technologies in Urban Schools (LeTUS), a collaboration between the University of Michigan and Detroit Public Schools, has developed middle school science modules and provided extensive professional development in their use to teachers at urban schools (Blumenfeld, Fishman, Krajcik, & Marx, 2000; Schneider, Krajcik, & Blumenfeld, 2005). This type of program differs somewhat from the TD model in its development of modules, greater emphasis on the use of computer technology, initial focus on a subset of science teachers in a school, and less emphasis on peer coaches but retains the overall goal of improving science instruction through a curriculum that includes the day-to-day lessons to teach it and providing the intensive and ongoing teacher support to use it.

A next step is to determine whether this model improves science achievement. In this study, I examine the science achievement of one cohort of students passing through Grades 5-7 at three high-poverty, inner-city, middle schools in Philadelphia using this model.

Their achievement (measured using both standardized test score and achievement level) is compared to students', in the same cohort, achievement as they pass through three closely matched schools in Philadelphia not using the model as well as through the 23 middle schools in the district serving high-poverty, high-minority student populations.

BACKGROUND

The study runs between the fall of 1998 and spring of 2001. During this period, the district was in the middle of an ambitious reform program entitled Children Achieving (School District of Philadelphia, 1995). The program's centerpiece was a school accountability index composed of student gains on the Stanford Achievement Test (SAT 9), student and teacher attendance, and grade promotion. Expected levels of annual gains were established for each school and rewards and sanctions were applied accordingly. Small learning communities, which divided schools into smaller units of teachers and students, were established in middle schools with differing levels of autonomy. Subject standards and benchmarks were being established. An Urban Systemic Initiative funded by the National Science Foundation supported districtwide science reform by training science teacher leaders, offering summer institutes open to all teachers, and providing small grants to encourage schools to adopt reform-based science curricula. Both the experimental and control schools had the benefit of these policies and programs.

The Philadelphia reform did not seek to create a districtwide curriculum in science across and within schools. Schools were responsible for developing a science curriculum that met the district's science standards and were free to develop their own, choose an already available one, or work with an outside partner. Within schools, small learning communities and often teachers were free to do the same. In practice, the lack of time to develop a curriculum and train teachers in it, the lack of funds to purchase enough materials for schoolwide implementation of the curriculum and train the teachers in its use, and the turnover of science teachers made it difficult to create and maintain a standard science curriculum throughout a school. As a result, neither experimental nor control schools were implementing a single science curriculum consistently across and within their grades at the start of this study. Over the course of the study, none of the control schools adopted a new curriculum schoolwide, though individual small learning communities did.

These district factors also affected the ability of the three experimental schools to fully adopt the new science curriculum resulting in the differential levels of program implementation. School- and grade-level variation also affected implementation through such factors as teacher willingness to learn and use new curriculum materials, administrative support, the availability of resources (including funds and professional development time), and the level of teacher turnover.

The impact of any educational reform can be linked directly to its level of implementation in the classroom and the school (Berman & McLaughlin, 1977, 1979; Crandall et al., 1982; Feiner et al., 1997; Stallings & Kaskowitz, 1974; Stringfield et al., 1997). Table 1 presents the level of implementation of the program at each of the three schools by grade and year. The bolded conditions in the table were those faced by the cohort examined in this study. Students in this study at School I had 2 years of low implementation (fifth and sixth grades) and 1 year of medium implementation while students at School 2 received 1 year of medium and 2 years of high implementation and at School 3 they had 2 years of medium implementation. The nonbolded implementation conditions in Table 1 were faced by later cohorts and while not relevant to the analysis presented here they do illustrate two lessons of educational interventions. First, implementation increased year by year at the three schools

TABLE 1
Implementation of TD Science Program by School by Year

Grade	School 1	School 2	School 3
1998-1999	**Low in all grades**	**Medium in all grades**	
1999-2000	Medium in fifth grade	**High in all grades**	**Medium in sixth grade**
	Low in sixth grade		Low in seventh grade
	Low in seventh grade		
2000-2001	High in fifth grade	**High in all grades**	High in sixth grade
	Medium in sixth grade		**Medium in seventh grade**
	Medium in seventh grade		

[a]School 3 does not have a fifth grade.

as teachers received professional development and became convinced of its usefulness. Second, implementation varied not only between schools but also within school by grade.

School 1 had the weakest implementation for the cohort studied due to an initial lack of funds to purchase materials. In the first year, the program provided supporting materials and training through classroom visits and after-school trainings. FOSS and STC science materials were purchased first with grants from the Urban Systemic Initiative and later using school funds. These were incorporated into the curriculum first in the fifth grade and moving up through the sixth and seventh grades over time, often too late to have a major effect on the relevant cohort, and teacher in-class and after-school training were provided.

School 2 had the strongest implementation for the cohort studied. After developing a schoolwide science curriculum, School 2 used school funds to purchase the science materials and supplemented these with others from the Franklin Institute Science Museum. Additional materials for all grades were purchased with a grant from the Urban Systemic Initiative. With the provision of monthly professional development and in-class support, implementation was strong for the next 2 years of the study, especially for the lower grades. Seventh grade suffered from extreme science teacher turnover but the continued provision of the basic level of professional development maintained a strong implementation.

School 3 had a medium level of implementation for the studied cohort, but these students received 1 less year of exposure to the treatment as they entered the school in sixth grade. In Year 1, the science faculty developed a standard curriculum and attended after-school and in-class professional' development built around the science materials purchased by the school for the sixth grade. The next year saw the purchase of materials for the seventh grade and the continuation of the same level of professional development leading to strong implementation in the sixth grade and medium implementation in the seventh grade.

These differing and suboptimal levels of implementation are not unusual for educational interventions and can justify two views regarding the results of this study. First, the results provide a lower-end estimate of the impact of the intervention as full implementation did not yet occur. Or second, that the results give a realistic estimate of the intervention's impact, given the expected variation in implementation of any educational reform. These results then provide a realistic estimate of the impact but one that could be increased through ongoing efforts to improve implementation.

STUDY DESIGN AND DATA

This study tests the impacts of a standardized implementable curriculum, based on NSF-supported curricula materials, used within and across grades schoolwide, and supported by

ongoing professional development and in-class coaching focused on that curriculum. These conditions compose the TD science program and so are tested together.

I use a nonequivalent group design (Campbell & Stanley, 1963), a type of quasi-experimental design, to evaluate if the TD science program improves student achievement in science. The design includes three treatment schools paired with three matched control schools. Matching was done on the basis of such school characteristics as minority composition, poverty level, and average student test scores prior to our involvement. Using a longitudinal design, a cohort of students is followed from the end of fourth grade through seventh grade in 1998-2001. During this period, the three treatment schools implemented the program while the three matched control schools did not. Standardized science test scores were collected from the same students in the spring of fourth grade, before students began attending middle school, and the spring of seventh grade. Because students in the treatment schools may differ in their exposure time to the program (due to the mobility of the student body and differences in the timing of student transfer to middle school), the exact exposure time of each student in the treatment schools is measured. The program's effectiveness is judged by whether the amount of exposure time is significantly associated with gains in science achievement from fourth to seventh grade, controlling for different matched paired schools as well as a variety of individual student characteristics, including race/ethnicity, gender, behavior, and attendance. I examine both the quantitative change in student achievement over this period as well as any change in qualitative achievement levels, which more clearly shows whether students are reaching a basic level of achievement in the subject. If students in a treatment school show significantly greater gains than similar students in the paired control school, we have evidence that the TD science program improves students' science achievement.

While students were not randomly assigned to treatment or control groups due to practical constraints, the quasi-experimental design approximates the randomized experimental design. The purpose of randomization is to make the treatment and control groups comparable on both observed and unobserved characteristics. The experimental and control schools are as comparable as possible on many observed characteristics. The matching of the pairs of schools was done with expert input from the school district to increase the similarity of the students being compared and their school environments. The comparability is further improved by statistical models controlling for the matched school pairs, the cohorts of students, and the interaction among school pairs and exposure time to the treatment. In addition, the statistical models also control for a variety of individual student observed characteristics as well as unobserved individual traits such as their innate ability. These two features enhance the comparability between the treatment and control groups.

In addition, I also compare the treatment results to those from the 23 other district middle schools serving high-poverty, high-minority populations. This comparison does not include matching the schools by specific characteristics. However, it does give us a sense of how the treatment results compare to the general state of student achievement in the district at schools facing similar conditions.

Measurement

A standardized science test is used to measure student science achievement. As the United States lacks both a national science curriculum and national tests, many school districts use one of the commercially developed achievement tests such as the Stanford Achievement Test, the Iowa Test of Basic Skills, or the Comprehensive Test of Basic Skills to measure student achievement. These tests are based on surveys of curricula around the country using reviews of content standards, curriculum, and textbooks and are widely piloted and then

field tested to create national norms. During the period of this study, Philadelphia gave the Stanford Achievement Test version 9 (SAT 9) to fourth and seventh grades. The SAT 9 science test has 40 multiple-choice items each covering a content area (Life, Physical, or Earth and Space Science) and a process type (Using Evidence and Models, Recognizing Constancy and Patterns of Change, and Comparing Form and Function) (Harcourt Brace, 1996).

Three characteristics of the SAT 9 made it appropriate for use in this study. First, because science curriculum was a school-based decision, the test used should be a general test of student science knowledge (both process and content). Any test closely tied to the curriculum used by the experimental schools would give those students an unfair advantage over the control students who may have received a very different curriculum. As achievement gains are measured over 3 years of study, students have the time to make broad gains in science knowledge that can be picked up on a general test like the SAT 9. Second, the SAT 9 was integrated into the district's testing schedule and was taken seriously by the schools. At the time, Philadelphia students were heavily tested using both the SAT 9 (in English, math, and science) and a state test (in English and math). Adding an additional test near the end of the year would raise the problem of test fatigue among both teachers and students, leading to indifference and underperformance.

Third, SAT 9 scores from different grade levels can be compared using a continuous score scale (Harcourt Brace, 1997). The publisher scored the tests and converted them to scaled scores based on a national equating program. The difference in the scaled score between the fourth-grade test and the seventh-grade test can then be used to assess the effect of the program. Normal curved equivalents (NCEs), derived from the scaled scores by converting percentile rankings to normalized z-scores, are also available. Measuring the change in the score for the same students eliminates the effects of any unobserved, time-invariant, student characteristics that may be related to student achievement.

Using student responses to specific sets of questions on the SAT 9, students can be placed into one of four criteria-based achievement levels (Harcourt Brace, 1997). For Philadelphia, these levels were named: Below Basic, Basic, Proficient, and Advanced. While these categories have the same labels as those used in the NAEP, they are not equivalent. District-standardized tests have been found in many cases to show higher student achievement (fewer students in the Below Basic category) than the NAEP (Standard & Poor's, 2005). One of Philadelphia's goals during this period was to reduce the number of students achieving Below Basic. To examine whether the TD science program supported the district's goal, I use a second, qualitative, dependent variable concerning the change in achievement level between fourth and seventh grade looking specifically at the movement between two revised categories, Below Basic versus Basic and Above, created by collapsing the original four categories.

The treatment is the TD science program and the key explanatory variable is the number of years students are exposed to the program. I use years of exposure rather than a 0-1 dummy variable because I expect greater exposure to lead to a greater impact on achievement and students differ in their exposure for two reasons. First, in two of the pairs of experimental and control schools, students entered middle school in Grade 5, but in the third pair, students entered in Grade 6. Second, students transfer in and out of schools leading to different levels of exposure for students in the same experimental school. The value of this variable is each student's years of exposure to the program (or years at the school during this period). Fifth and sixth grades are given the value of 1 year of exposure apiece while the seventh grade year is coded at 0.75 years because students took the SAT 9 in early spring rather than at the end of the academic year. Students in the dataset have 0, 0.75, 1.75, or 2.75 years of exposure to the science program with students at the control schools having the 0 values.

TABLE 2
School Pair Matches: Data from 1996 to 1997

	School Pair 1		School Pair 2		School Pair 3	
Characteristic	School 1	Control 1	School 2	Control 2	School 3	Control 3
Race						
Black (%)	26	19	75	65	98	100
White (%)	13	20	1	11	1	0
Other (%)	61	61	24	24	1	0
SAT 9 average reading comprehension scaled scores	648	638	647	641	658	638
SAT 9 average math scaled scores	653	642	644	644	658	639
SAT 9 average science scaled scores	639	629	636	638	645	633
Attendance rate	86	77	83	83	86	81
Percent low income	86	95	87	92	71	81

In order to control for differences between the experimental and control schools that may be related to differences in student achievement, a careful matching of control with experimental school was done when we first began working with each school. In cooperation with the Philadelphia school district, control schools were chosen based on similarities to their matched experimental schools in student poverty levels (using free and reduced price school lunch participation), student attendance, racial and ethnic make-up of the student body, and previous years' performance on the standardized achievement tests, primarily using a combination of math and reading scores on the SAT 9 (Table 2). Two of the three experimental schools have a smaller percentage of White students than their controls. In most cases, the experimental schools' test score averages are slightly higher than their controls. Using growth in test scores rather than absolute scores helps adjust for these differences.

I control for student observable characteristics that may be related to student achievement in order to reduce these differences between our experimental and control students. Specifically, the analyses control for race/ethnicity, gender, English as a Second Language (ESL), special education status, attendance, and behavior. Race/ethnicity and gender are time-invariant variables and the majority of their impact will be eliminated by our modeling technique. However, they are included to capture any unobserved, time-varying effects they may have although these should be modest. Gender, ESL, and special education are 0-1 dummy variables. Race/ethnicity is composed of four categories, Asian, Black, Hispanic, and White, each represented by 0-1 dummy variables. Attendance is the ratio of days attended over total days of school averaged over the years at middle school. Behavior is based upon behavior ratings (1, 2, and 3 representing Excellent, Satisfactory, and Unsatisfactory, respectively) given by each student's teachers every quarter. These values were averaged for each teacher, then averaged across teachers, and then averaged across the years of attendance at middle school. The number and type of science courses are standard at these grade levels, so no controls for these are necessary.

Descriptive Analysis

Table 3 provides the basic descriptive statistics, means, and standard deviations, regarding student exposure to the science program and student achievement for the whole sample

TABLE 3
Descriptive Data—Means and Standard Deviations by School Pair and Cohort

School Pair	Variable	Treatment Students at 3 Experimental Schools	Control Students at 3 Matched Schools	Control Students at 23 District Schools
All	Exposure (years)	2.0 (0.7)	0	0
	Fourth grade scale score	600 (29)	610 (28)	605 (29)
	Net growth in scale score from fourth to seventh grade	34[3,23] (28)	25 (28)	30 (27)
	n	630	463	3851
Pair 1	Exposure (years)	2.4 (0.7)	0	
	Fourth grade scale score	610 (30)	611 (26)	
	Net growth in scale score from fourth to seventh grade	27 (28)	26 (27)	
	n	201	152	
Pair 2	Exposure (years)	2.2 (0.8)	0	
	Fourth grade scale score	597[3] (25)	608 (27)	
	Net growth in scale score from fourth to seventh grade	44[3] (26)	29 (24)	
	n	140	141	
Pair 3	Exposure (years)	1.6 (0.3)	0	
	Fourth grade scale score	605 (28)	610 (30)	
	Net growth in scale score from fourth to seventh grade	35[3] (28)	21 (30)	
	n	260	170	

[3 or 23] Significantly different from students at the 3 or 23 control schools at the 0.01 level.

and also breaks it down by school pair (each matched experimental and control school). Statistically significant differences are noted using the superscripts "3" to show a difference of the treated students from the students at the 3 matched control schools and "23" to show a difference with the students at the 23 district middle schools with large percentages of poor and minority students. Our matched sample of 1093 students contains 630 students with some exposure to the TD science program and 463 control students at the 3 matched control schools (3851 students at the 23 district schools). The distribution of the students across the school pairs is 353 students for School Pair 1, 281 for School Pair 2, and 430 for School Pair 3.

The average level of exposure to the TD science program for students in experimental schools is 2 years for the whole sample. This level is lower (1.6 years) for School 3 because it does not contain a fifth grade and highest at 2.4 years for School 1.

Initial student achievement, measured by fourth-grade test scores, is significantly lower for treated students (6 scaled points) than for the matched school students (but not for students at the 23 district middle schools). When the sample is broken down by pairs, this significant difference remains only for School Pair 2.

Growth in student achievement is significantly greater for the students taking part in the TD science program (9 scaled points versus the matched schools). This significant difference continues to occur for School Pair 2 and School Pair 3.

ANALYSES AND FINDINGS

Two multivariate analyses are carried out on the impact of the TD science program on student achievement. The first examines the impact of student exposure to the program on the change in their science achievement test scores between fourth and seventh grades. The second examines the impact on their change in science achievement level between the two grades. The statistical models used for each and their results are discussed below.

Analysis 1: Change in Student Test Scores

Our first analysis examines the impact of the TD science program on the change in student test scores on the SAT 9 science achievement test between fourth and seventh grades. I examine the relationship between the level of student exposure to the program and the change in test scores for one cohort of students who matriculated through three pairs of matched middle schools while controlling for student characteristics. I expect to find a positive relationship between exposure and gain in test score.

Statistical Models. I specify two regression models corresponding to Eqs. (1), (1 a), and (2). Model 1 specifies the main effect of treatment controlling for other student control variables and can be used with students from both the 3 matched control schools and the 23 district middle schools with large percentages of poor and minority students. An extension, Model la, includes a control for School Pair and so can only be used with students from the 3 matched control schools. Model 2 includes the interaction between treatment and School Pair in order to examine any differences in the treatment's effect among schools.

$$y_i = \beta_0 + \beta_1 E_i + \beta_4 X_i + \varepsilon_i \quad (1)$$

$$y_i = \beta_0 + \beta_1 E_i + \beta_2 S_i + \beta_4 X_i + \varepsilon_i \quad (1\,a)$$

$$Y_i = \beta_0 + \beta_1 E_i + \beta_2 S_i + \beta_3 E_i S_i + \beta_4 X_i + \varepsilon_i \quad (2)$$

where the subscript i is for students. Y_i is the gain or loss in test score (scaled score or NCE) between fourth and seventh grades. E_i is a continuous variable representing years of exposure to the TD science program. S_i is a vector of two dummy variables representing the three school pairs (with Pair 1 as the reference). X_i is a vector of student-level control variables including race/ethnicity, attendance, behavior, female, special education, and ESL.

Substantively, the models provide several views of the relationship between exposure to the TD science program and growth in student science test scores. Model 1 will show the relationship for all students either between treated students and untreated students in the 23 district schools or the 3 matched control schools. Model 1 a will show the relationship for all students controlling for School Pair. Model 2 will show the relationship by School Pair giving the most complete picture of the relationship.

Results. Table 4 shows the relationship of exposure to the TD science program to student gains in science achievement for the three models using scale scores and Table 5 uses NCEs. The positive significant coefficients for Model 1 show that students at the treatment schools benefited from the program gaining about 3.5 scaled points or 2 NCEs more for each year of exposure in comparison to students at the matched control schools (or 2 scaled points and 1.2 NCEs versus students at the 23 district middle schools).

TABLE 4
Coefficients for 1 Year Effect of TD Science Program on Change in SAT 9 Science Test Scale Scores: Fourth—Seventh Grade

Exposure	Model 1: 23 District Schools	Model 1: 3 Matched Schools	Model 1 a: 3 Matched Schools	Model 2: 3 Matched Schools
For students in all schools	2.08[a] (0.58)	3.47[a] (0.79)	3.89[a] (0.80)	
For students in School Pair 1				—0.60 (1.20)
For students in School Pair 2				7.21 [a] (1.35)
For students in School Pair 3				7.27a (1.61)
n	4474	1062	1062	1062

Notes: Model 1 controls for: race/ethnicity, attendance, behavior, female, special education, and English as a Second Language.
Model la adds a control for school pair.
Model 2 adds an interaction term between exposure and school pair.
Of the estimates for the coefficients of the control variables from Model 2, only the coefficient for behavior was significant.
[a]Significant at $p < 0.01$ level.

TABLE 5
Coefficients for 1 Year Effect of TD Science Program on Change in SAT 9 Science Test NCEs: Fourth—Seventh Grade

Exposure	Model 1: 23 District Schools	Model 1: 3 Matched Schools	Model 1 a: 3 Matched Schools	Model 2: 3 Matched Schools
For students in all schools	1.20[a] (0.31)	1.93[a] (0.45)	2.17[a] (0.45)	
For students in School Pair 1				—0.48 (0.68)
For students in School Pair 2				4.09[a] (0.76)
For students in School Pair 3				4.21[a] (0.91)
n	4474	1062	1062	1062

Notes: Model 1 controls for: school pair, cohort, attendance, behavior, female, special education, English as a Second Language, and race/ethnicity.
Model la adds a control for school pair.
Model 2 adds an interaction term between exposure and school pair.
[a]Significant at $p<0.01$ level.

However, the simplified Model 1 does not consider whether the results vary by school. When School Pair is included as a control variable in Model la, the results remain consistent as treated students score almost 4 scaled points and 2.2 NCEs more than the control students in the matched control schools.

The estimates from Model 2, which relaxes the assumption that the students in all schools share the same coefficient, address how the results differ by school. Students receiving the treatment in Schools 2 and 3 have significant positive coefficients (gains of over 7 scaled points or 4 NCEs per year of exposure) while students in School 1 show no significant differences from their counterparts in the matched control school.

TABLE 6
Effect Size for Coefficients from Tables 5 and 6

Model	Exposure	Scale Scores	NCEs
Model 1 (23 district schools)	For all students	0.06[a]	0.06[a]
Model 1 (3 matched schools)	For all students	0.14[a]	0.14[a]
Model I a Model 2	For all students	0.16[a]	0.15[a]
	For students in School Pair 1 —	0.02	—0.03
	For students in School Pair 2	0.20[a]	0.20[a]
	For students in School Pair 3	0.18[a]	0.19[a]

[a]Significant at $p<0.01$ level.

Table 6 shows the effect sizes of the exposure to the TD science program for those coefficients found significant in Models 1, la, and 2. These are calculated by multiplying the coefficient for exposure by the standard deviation of the treatment divided by the standard deviation of the test score. A 1 standard deviation change in exposure to the TD science program at Schools 2 and 3 leads to about a 0.20 standard deviation gain in test score as compared to a student in the matched control schools. Traditionally, effect sizes of this magnitude have been considered small (Cohen, 1988). However, recent research specifically concerned with field studies of educational interventions has found that effect sizes of this type represent serious practical impacts (Kane, 2004; Keller, 1995). Kane (2004) notes that the national samples used to norm the math SAT 9 found an effect size of 0.50 when moving from fourth to fifth grade, i.e., a whole school year is associated with one-half of a student-level standard deviation gain in test score. Assuming a similar finding for the science SAT 9, the effect of the TD science program would be equivalent to the effect of about 40% of a school year.

Our first set of models offer a significantly and substantially positive picture of the impact of the TD science program on student achievement at two of the three treatment schools. At the other school (School 1), students did no better than those at their matched control school but the low level of implementation at School 1 for the first 2 years may have contributed to this result.

Analysis 2: Change in Science Proficiency

The analysis in the previous section focuses on the gain in science SAT 9 scores as a result of exposure to the TD science program. The SAT 9 score is typically used as a relative rating of students against one another along a continuous scale. Its developers also created four criteria-based achievement levels to identify the science proficiency of each student. These categories include: Below Basic, Basic, Proficient, and Advanced. During the period of this study, the Philadelphia school district sought to move students categorized as Below Basic to Basic or higher.

Our second analysis considers the impact of the TD science program on the development of students' proficiency in science. Following Philadelphia's goal, I collapse the achievement levels into Below Basic and Basic or Above. For simplicity I will refer to these categories as "Below" and "Above." I examine the change in the achievement level for the same students from Grade 4 to Grade 7. Cross-classifying the two achievement levels in the two grades yields the following four transitions: (1) remaining Below, (2) improving from Below to Above, (3) dropping from Above to Below, and (4) remaining Above. Table 7 shows these four transitions and the proportion of the total population in each cell. The situation in the

TABLE 7
Proportion of Student Population in Achievement Levels

	Grade 7		
Grade 4	Below (%)	Above (%)	Total Grade 4 (%)
Below	24	15	39
Above	22	39	61
Total Grade 7	46	54	

six schools is quite sobering: 46% of students are Below Basic in Grade 7 with 24% of them remaining there from Grade 4 and 22% dropping from the higher level in Grade 4. Table 7 shows the increase in the percentage of students achieving at Below Basic as they move from fourth grade to eighth grade (39 to 46%), consistent with the trend found in the NAEP.

I expect that the TD science program can enhance students' proficiency and prevent students from dropping in their achievement level. In particular, I expect that the probability of transition 2 versus transition 1 is positively associated with the years of exposure to the program. That is, for those students Below Basic level in Grade 4, more will move up to Above and fewer will stay Below in Grade 7 with exposure to the program. Similarly, I expect that the probability of transition 3 versus transition 4 is negatively associated with the years of exposure to the program. That is, for those students Above in Grade 4, fewer will drop down to Below Basic and more will remain Above in Grade 7 with exposure to the program.

Statistical Models. I use a multinomial logit model because of an interest in knowing how exposure to the TD science program affects the probability of each of the four transitions while considering all the transitions at once. The model specifies a set of dependent variables, each of which is the odds of two of the transitions. In this case, I am interested in two comparisons that lead to a focus on two dependent variables:

1. The odds of moving from Below to Above versus staying Below.
2. The odds of moving from Above to Below versus staying Above.

For mathematical simplicity, the dependent variable is logged in multinomial logit models, allowing the explanatory variables to be used in the same manner (a linear function) as in Eq. (1). Because the dependent variable is logged, the coefficients on the explanatory variables are called log odds.

The equation below specifies the odds of moving from Below to Above versus staying Below as a function of exposure to the TD science program *(E)* while controlling for school and student characteristics (X). *P* stands for probability and i stands for student *i*.

$$\log\left(\frac{P_i(\text{Move\_Below\_to\_Above})}{P_i(\text{Remain\_Below})}\right) = \beta_0 + \beta_1 E_i + \beta_2 X_i \qquad (3)$$

Similarly, Eq. (3') specifies the odds of moving from Above to Below versus staying Above:

$$\log\left(\frac{P_i(\text{Move\_Above\_to\_Below})}{P_i(\text{Remain\_Above})}\right) = \beta_0' + \beta_1' E_i + \beta_2' X_i \qquad (3')$$

Equation (3) is for students who were initially Below and Eq. (3') is for students who were initially Above. These two equations are estimated simultaneously. Similar to Eq. (1), they examine the overall treatment effect.

Another way of thinking of the multinomial logit model is the probability of any outcome. For example, I am interested in the change in probability of moving from Below to Above and the change in probability of moving Above to Below brought about by 1 year of exposure to the TD science program. The odds ratio coefficients need to be transformed in a nonlinear way into these changes in probability.

This achievement-level analysis parallels the previous analysis on changes in test scores: Model 3 is parallel to Model 1 (the overall treatment effects among students in 23 district schools), Model 3a is parallel to Model la (treatment versus matched school students), and Model 4 is parallel to Model 2 (allowing the treatment effect to differ by school pairs).

Results. The estimates regarding the marginal effect of exposure for Eqs. (3), (3a), and (4) are shown in Table 8 for moving from Below to Above versus remaining Below and Table 9 for moving from Above to Below versus remaining Above. These tables show the log odds coefficients and the changes in probability for a student making a specific transition between achievement levels due to an increase of 1 year in exposure to the TD science program.

Table 8 shows that exposure to the TD science program increases the chances of rising from Below to Above in all three models with an increase in such probability per year of exposure ranging from 0.03 to 0.09. The findings for Model 4 are significant at Schools 2

TABLE 8
Effect of 1 Year of TD Science Program on the Transition from Below to Above vs. Remaining Below (Transition 2 vs. Transition 1)

Model	Exposure	Coefficient (Log Odds)	Change in Probability of Outcome 2
Model 3 (*n*=4475)	For all students	0.21[a]	0.028
Model 3a (*n*= 1062)	For all students	0.21[b]	0.041
Model 4 (*n*=1062)	For students in School Pair 2	0.46[a]	0.088
	For students in School Pair 3	0.53[b]	0.072

[a]Significant at $p < 0.01$ level.
[b]Significant at $p < 0.05$ level.

TABLE 9
Effect of 1 Year of TD Science Program on the Transition from Above to Below vs. Remaining Above (Transition 3 vs. Transition 4)

Model	Exposure	Coefficient (Log Odds)	Change in Probability of Outcome 3
Model 3 (*n*=4475)	For all students	–0.12[b]	–0.025
Model 3a (*n*= 1062)	For all students	–0.25[a]	–0.056
Model 4 (*n*= 1062)	For students in School Pair 2	–0.44[b]	–0.104
	For students in School Pair 3	–0.39[a]	–0.077

[a]Significant at $p < 0.01$ level.
[b]Significant at $p < 0.05$ level.

and 3, similar to Model 2. At School 2, the coefficient is 0.46 and significant at the 0.01 level. An increase of 1 year of exposure to the TD science program increases the probability of moving from Below to Above versus remaining Below by 0.09. This is a large effect since it increases the proportion of outcome 2 by over 60% (the proportion of the population for this outcome is 0.143). I find a similar effect for School 3. The coefficient is 0.53, significant at the 0.01 level. The probability for outcome 2 increases by 0.07.

Table 9 shows that exposure to the TD science program decreases the chances of falling from Above to Below in all three models with the decreases in probability ranging from 0.03 to 0.1. For Model 4, the TD science program reduces the probability of a dropping from Above to Below versus remaining Above at both Schools 2 and 3. The coefficient for School 2 is —0.44, significant at the 0.05 level, which translates into a decline in probability by 0.10 for every 1-year increase in exposure to the TD science program. This effect is also sizable: it decreases the probability of dropping achievement levels by over 50% (the proportion of the population for this outcome is 0.194). The coefficient for School 3 is —0.39, significant at the 0.01 level, which means a reduction in the probability of 0.08 for every 1-year increase in exposure to the TD science program.

In summary, this analysis provides evidence that the TD science program can both enhance low-proficient students' achievement levels while helping maintain higher proficient students at their achievement level. As in the case of our first analysis, the results apply to schools with better implementation (Schools 2 and 3).

DISCUSSION

Science achievement at U.S. urban schools serving high-poverty, high-minority student populations is extremely low and threatens both student success in high school science and efforts to reform science education in urban districts. The results of this study offer some optimism that school—university partnerships can improve science instruction through a focus on day-to-day curriculum materials and professional development on how to teach with them. The results add to the literature regarding the success of similar partnerships in improving teacher instruction. Additionally, the results provide evidence that the changes in instruction improve student achievement as well. This consists of both gains in standardized test scores as well as movement up (or not down) through achievement levels. The latter has become increasingly important as individual schools and districts are required to not only raise test scores but also move students up to higher levels in order to avoid state sanctions. Third, the results are for whole schools. In this way, they avoid the concern that only the most motivated or qualified teachers will take part and only their students will benefit while also addressing school and district needs to make schoolwide gains.

The day-to-day curriculum was based on FOSS and STC science modules. I cannot tease out the impact of these science materials versus the professional development provided as their use was bound up with one another. However, as the standardized science tests measured both science content and procedural knowledge and much of the professional development focused on using these modules, I can assume that treatment students more successfully learned this knowledge when using these science materials with teachers trained in them. Much of the research on these modules has focused on elementary students and often taken the form of district reports. Our results add to the formal literature on their benefits for middle grades education when coupled with extensive professional development. One practical aspect of these modules is that information on them is widely available and they are easily obtainable due to their commercial availability. Districts and university partners may not have the resources to develop equally complete materials. Additionally, the years needed to develop such materials may not be practicable, given that schools are to be evaluated

on their students' gains in science in the upcoming years. University-developed materials are often less available and less known as universities normally have weaker marketing structures versus commercial firms. However, some university–school partnerships have succeeded with university-developed materials. To ensure wider opportunities for their use by other partnerships, universities could consider broadening their own marketing structures, making their materials available to commercial firms, or supporting the newly proposed policy of collecting and creating K-12 curricula materials in a national voluntary curriculum database that would be distributed for free (National Academy of Sciences, 2006).

The caveat to using this study's results in support of these science modules is the need for intensive and ongoing professional development for all science teachers at the school in their use, requiring resources greater than the materials themselves. Professionally, development was provided through workshops and in-class coaching. Four lessons regarding workshops were learned. First, they should focus on the curriculum, which requires that separate workshops be held for all the teachers in each grade at a school. Second, workshops are structured around the teachers doing every lesson in a module, just as their students will, so that teachers are fully prepared to teach it. Workshops that only give an overview or cover a few lessons can lead to faulty implementation as teachers in these schools have little time to learn the lessons on their own. This can lead to lessons being skipped or incompletely taught. In addition, content background and pedagogical techniques (including materials management) are given for each lesson to help ensure teachers feel prepared to teach and answer student questions. Third, workshops are held throughout the school year. In this way, they cover what is to be taught next so that lessons are fresh in teachers' minds. Also, this provides opportunities for teacher feedback on how best to organize the workshops and a chance for teachers to share ideas with their colleagues. Fourth, after the first year, different levels of workshops have to be held. For new teachers, the basic workshops need to be provided so that the curriculum continues to be taught by all teachers. For experienced teachers, workshops providing new information linked to teaching the curriculum help to keep them involved. For exemplary teachers, workshops on how to teach the curriculum to new teachers provide the school with a mechanism to maintain the program.

While the workshops provide the general knowledge needed to implement the curriculum, the peer coaches were needed to address the classroom-level factors of both teachers and students that determine whether any curriculum is well implemented. There is a wide literature on the diversity of poor urban students as to their backgrounds, their strengths and their weakness and how best to orient instruction to take advantage of student resources and overcome the obstacles facing them. Often it is up to teachers to integrate this information into their instruction and lesson planning. At the same time, while many of available science modules are generally developed to support these students' learning (e.g., provide scaffolding to move students toward more independent work), they cannot address every type of classroom. The peer coach acts as the intermediary between the two. With a depth of experience and the opportunity to visit the class on a regular basis, they can help the teacher identify the needs of the class and the forms of instruction that can best meet them while also identifying strengths students have that can be built on. As the year goes on and students understand and often enjoy how the science modules work, the coach can advise the teacher on how instruction can be shifted to give students a greater role. At the same time, the coach can help teachers overcome the little obstacles (e.g., missing materials, a poorly designed procedure) that often prevent the use of hands-on activities. Coaches are also instrumental in helping schools incorporate structures that allow teachers to provide professional development to one another, such as regular science meetings, into the regular school schedule and identifying exemplary teachers who with support can take on greater roles in improving instruction throughout the school.

One other factor found key to the success of the coaches is that they are viewed by teachers as trusted supporters and not evaluators. Teachers are often concerned that outsiders coming into their classrooms either take up their time with district-mandated information that they do not find useful or are there to report on them to administration. Coaches seek to avoid this perception by providing advice directly relevant to instruction in that classroom, seek to influence instruction by example and not by pressure to follow mandates, and follow a strict policy of confidentiality. In some cases, it may take almost the full school year for a teacher to trust a coach and use him or her as an adviser. While seemingly slowing implementation, this approach may actually increase teacher support of the program.

The study illustrates the need for long-term support and evaluation. This project required 3 years to achieve acceptable levels of schoolwide implementation. A 1-year analysis of student outcomes would probably not have found positive effects on student achievement. Instead, by following the science achievement gains in a school's student body over several years, the impact of the program was identified. Increased use of longitudinal analyses of changes in student achievement may be a better way of measuring the effects of attempts to improve science instruction.

The application of the TD science program is not limited to our center or to a university–school partnership. Built on a simple teacher-support model and available materials, it can easily be adopted by other school partners or by a district with the latter having the further benefit of gaining economies of scale and alignment of district policies (e.g., material purchases, use of district professional development days, and induction of new teachers). In some cases, districts have already adopted some of these approaches but applied them in an inconsistent manner. During the study period, the Philadelphia school district provided:

(1) NSF-supported science materials but not for all classes and failed to resupply them,
(2) outside the classroom professional development but primarily to school science leaders not entire school science faculties, and (3) limited in-classroom support but not by persons expert in the curricula, science content, and middle school instruction. The key is the provision of all the components for all teachers teaching science so that all students are affected throughout their middle school attendance.

Political change in the Philadelphia school district prevented the inclusion of additional years of middle school or additional cohorts to this study. In the middle of the last year of the study, the state installed a School Reform Commission, with a majority appointed by the governor, to oversee the district. Major changes including privatization and school restructuring began in the following school year. In the summer of 2002, a new CEO was appointed for the district and he has made major changes in curriculum and instruction, many similar to those used in the TD science program. A grade-by-grade science curriculum, the materials to be used to implement it, and a pacing guide were instituted for fifth and sixth grades in the 2003-2004 school year and for seventh and eighth grades in the next year. Professional development was focused more on the use of these materials and some in-class coaching was provided. Currently, the Center for Social Organization of Schools is analyzing the impact of these reforms to determine if changes like those supported at the school level by the TD science program will have similar positive impacts on science achievement when implemented at the district level.

REFERENCES

Balfanz, R., Macbier, D., & Ryan, D. (2002). Enabling "algebra for all" with a facilitated instructional program. In V. Anfara, & S. Stacki (Eds.), Middle school curriculum, instruction and assessment (pp. 181-209). Greenwich, CT: Information Age Publishing.

Balfanz, R., Ruby, A., & MacIver, D. (2002). Essential components and next steps for comprehensive whole-school reform in high-poverty middle schools. In S. Stringfield & D. Lands (Eds.), Educating at-risk students: One hundred-first yearbook of the National Society for the Study of Education, Part II (pp. 128-147). Chicago, IL: NSSE.

Saxton, C., Carlone, H., & Carlone, D. (2005). Boundary spanners as bridges of student and school discourses in an urban science and mathematics high school. School Science and Mathematics, 105(5), 302-312.

Berman, P., & McLaughlin, M. (1977). Federal programs supporting educational change: Vol. 8, Implementing and sustaining innovations. Santa Monica, CA: RAND.

Berman, P., & McLaughlin, M. (1979). Federal programs supporting educational change: Vol. 7, Factors affecting implementation and continuation. Santa Monica, CA: RAND.

Betts, J. R., Rueben, K. S., & Danenberg, A. (2000). Equal resources, equal outcomes? The distribution of school resources and student achievement in California. San Francisco, CA: Public Policy Institute of California. Blanc, S., & Ballenger, R. (1999). Philadelphia Urban Systemic Initiative interim report: 1998-1999. Philadelphia, PA: Research for Action.

Blumenfeld, P., Fishman, B., Krajcik, J., & Marx, R. (2000). Creating usable innovations in systemic reform: Scaling up technology-embedded project-based science in urban schools. Educational Psychologist, 35(3), 149-164.

Bodilly, S. (1998). Lessons from New American Schools' scale-up phase. Santa Monica, CA: RAND.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.

Christman, J. (2001). Powerful ideas, modest gains: Five years of systemic reform in Philadelphia middle schools. Philadelphia, PA: Research for Action.

Cohen, D., & Hill, H. (2001). Learning policy: When state education reform works. New Haven, CT: Yale University Press.

Cohen, J. (1988). Statistical power analysis for the behavioral science. Hillsdale, NJ: Lawrence Earlbaum. Crandall, D., Loucks-Horsley, S., Baucher, J., Schmidt, W., Eisman, J., Cox, P., Miles, M., Huberman, A., Taylor, B., Goldberg, J., Shive, G., Thompson, C., & Taylor, J. (1982). People, policies and practices: Examining the chain of school improvement (Vols. 1-10). Andover, MA: Network.

Cuevas, P., Lee, 0., Hart, J., & Deaktor. R. (2005). Improving science inquiry with elementary students of diverse backgrounds. Journal of Research in Science Teaching, 42(3), 337-357.

Darling-Hammond, L. (1992). Teaching and knowledge: Policy issues posed by alternative certification for teachers. Peabody Journal of Education, 67(3), 123-154.

Darling-Hammond, L. (2000). Solving the dilemmas of teacher supply, demand, and quality. New York: National Commission on Teaching and America's Future.

Desimone, L., Porter, A., Garet, M., Yoon, K., & Birman, B. (2002). Effects of professional development on teacher's instruction: Results from a three-year longitudinal study. Educational Evaluation and Policy Analysis, 24(2), 81-112.

Felner, R., Jackson, A., Kasak, D., Mulhall, P., Brand, S., & Flowers, N. (1997). The impact of school reform for the middle year. Phi Delta Kappan, 79,528-32,541-550.

Fullan, M. (1990). Staff development, innovation, and institutional development. In B. Joyce (Ed.), Change school culture through staff development (pp. 3-25). Alexandria, VA: Association for Supervision and Curriculum Development.

Grossman, P. L. (1990). The making of a teacher: Teacher knowledge and teacher education. New York: Teachers College Press.

Harcourt Brace. (1996). Stanford Achievement Test series, 9th edition: Compendium of instructional objectives. San Antonio, TX: Author.

Harcourt Brace. (1997). Stanford Achievement Test series, 9th edition: Spring norms book. San Antonio, TX: Author.

Hiebert, J., Gallimore, R., & Stigler, J. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? Educational Researcher, 31(5), 3-15.

Ingersoll, R. (1996). Out of field teaching and educational equity. NCES 96-040. Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Ingersoll, R. (1999). The problem of underqualified teachers in American secondary schools. Educational Researcher, 28(2), 26-37.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. American Educational Research Journal, 38(3), 499-534.

Kane, T. (2004). The impact of after-school programs: Interpreting the results of 4 recent evaluations. New York: W.T: Grant Foundation.

Keller, D. (1995). An assessment of national academic achievement growth. Unpublished dissertation. Newark, DE: University of Delaware

Kennedy, M. (2002). Content matters most. American Educator, 26(2), 24-25.

Killon, J. (1999). What works in the middle: Results-based staff development. Oxford, OH: National Staff Development Council.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. Educational Evaluation and Policy Analysis, 24(1), 37-62.

Lee, 0. (2003). Equity for linguistically and culturally diverse students in science education: A research agenda. Teachers College Record, 105(3), 465-489.

Lippman, L., Burns. S., & McArthur, E. (1996). Urban schools: The challenge of location and poverty. NCES 96-184. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Martin, M., Mullis, I., Gonzalez, E., & Chrostowski, E. (2004). TIMSS 2003 international science report: Findings from the IEA's Trends in International Mathematics and Science Study at the fourth and eight grades. Boston, MA: Boston College.

McLaughlin, M. W., & Oberman, I. (Eds.). (1996). Teacher learning: New policies, new practices. New York: Teachers College Press.

National Academy of Sciences, Committee on Prospering in the Global Economy of the 21st Century. (2006). Rising above the gathering storm: Energizing and employing America for a brighter economic future. Washington, DC: National Academy Press.

National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education. (1992). A profile of American 8th grade mathematics and science instruction. NCES 92-486. Washington, DC: U.S. Department of Education.

National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education. (1997). Characteristics of stayers, movers and leavers: Results from the Teacher Follow-up Survey: 1994-95. NCES 97-450. Washington, DC: U.S. Department of Education.

National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education. (1999). Teacher quality: A report on the preparation and qualifications of public school teachers. NCES 1999-080. Washington, DC: U.S. Department of Education.

National Research Council. (2000). Inquiry and the national science education standards: A guide for teaching and learning. Washington, DC: National Academy Press.

Neild, R. (2001). Distribution of certified teachers in the School District of Philadelphia. Philadelphia, PA: Philadelphia Education Fund.

Neild, R., Useem, B., Travers, E., & Lesnick, J. (2003). Once & for all: Placing a highly qualified teacher in every Philadelphia classroom. Philadelphia, PA: Research for Action.

O'Neill, T., & Barton, A. (2005). Uncovering student ownership in science learning: The making of a student created mini-documentary. School Science and Mathematics, 105(6), 292-298.

O'Sullivan, C., Lauko, M., Grigg, W., Qian, J., & Zhang, J. (2003, January). The nation's report card: Science 2003. NCES 2003-453. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Porter, A., Garet, M., Desimone, L., Yoon, K., & Birman, B. (2000, October). Does professional development change teaching practice? Results from a three-year study. Document #2000-04. Washington, DC: U.S. Department of Education, Planning and Evaluation Service.

Rhoton, J., & Bowers, P. (2002). Science teacher retention: Mentoring and renewal. Arlington, VA: National Science Teachers Association.

Ruby, A. (2002). Internal teacher turnover in middle school reform. Journal of Education for Students Placed At Risk, 7(4), 379-406.

Schaffer, E., Nesselrodt, P., & Stringfield, S. (1997). Impediments to reform: An analysis of destabilizing issues in ten promising programs. Arlington, VA: Educational Research Service.

Schneider, R., Krajcik, J., & Blumenfeld, P. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. Journal of Research in Science Teaching, 42(3), 283-312.

School District of Philadelphia, Office of the Superintendent. (1995). Children achieving: Action design. Philadelphia, PA: Author.

Smith, J., Smith, B., & Bryk, A. (1998). Setting the pace: Opportunities to learn in Chicago's elementary schools. Chicago, IL: Consortium on Chicago School Research.

Spillane, J., Diamond, J., Walker, L., Halverson, R., & Jita, L. (2001). Urban school leadership for elementary science instruction: Identifying and activating resources in an undervalued school subject. Journal of Research in Science Teaching, 28(8), 918-940.

Stallings, J., & Kaskowitz, D. (1974). Follow through classroom observation evaluation 1972-1973. SRI Project URU-7370. Menlo Park, CA: Stanford Research Institute.

Standard & Poor's. (2005, Fall). The National Assessment of Educational Progress and state assessments: What do differing student proficiency rates tell us? (Available at www.schoolmatters.com)

Stringfield, S., Millsap, M., Herman, R., Yoder, N., Brigham, N., Nesselrodt, P., Schaffer, E., Karweit, N., Levin, M., & Stevens, R. (1997). Urban and suburban/rural special strategies for educating disadvantaged children: Final report. Washington, DC: U.S. Department of Education.

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

Useem, E., Barends, R., & Lindermayer, K. (1999). The preparation of middle grades teachers in an era of high stakes and high standards: Philadelphia's predicament. Philadelphia, PA: Philadelphia Education Fund.

Useem, E., Christman, J., Gold, E., & Simon, E. (1997). Reforming alone: Barriers to organizational learning in urban school change initiatives. Journal of Students Placed At Risk, 2(1), 55-78.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A., & Hadicourt-Barnes, J. (2000). Rethinking diversity in learning science: The logic of everyday sense-making, Journal of Research in Science Teaching, 38(5), 529-552.

Wenner, G. (2003). Comparing poor, minority elementary students' interest and background in science with that of their white, affluent peers. Urban Education, 38(2), 153-172.