There is always a minimal disruption that we cause to an electron’s velocity through our measurement of its motion . . . and so we are faced with a quantum mechanical balancing act.

—Brian Greene describing the Heisenberg Uncertainty Principle (Greene 1999).

Kim: Whoever makes these books [science curriculum unit] must not have a life. Spend their time typing.
Angelique: I think they have a lot of time on their hands.

—Two eighth graders discussing a highly rated curriculum unit used in this study, 2002.

The purpose of this chapter is to present the rationale for a six-year study of highly rated, middle school science curriculum units and to discuss features of the project that affect scale-up as we begin our third year. A goal of this university/public school district collaboration is to help close achievement gaps in an ethnically, linguistically, and socioeconomically diverse suburban school district. By providing research evidence to guide implementation and scale-up, we study how and why the distinctive features of the curriculum units affect outcomes for diverse student populations. In this chapter, we:

- Provide the rationale for the study and its scale-up design;
- Discuss our collaborative scale-up research using a framework that includes discussion of interventions, the environment, and interactions between them (Cohen, Ball, and Barnes 2001);
Rationale, Design, and Early Findings

Rationale

The purpose of our research is to study the implementation and scale-up of three highly rated, middle school science curriculum units in Montgomery County Public Schools (MCPS), a large suburban school district in Maryland. The aim is to close achievement gaps in a school district that has an increasingly diverse student population. Our work has two phases. In the implementation phase, we introduce a new science curriculum unit to all of the students in a particular grade in five middle schools. In this quasi-experiment, five schools receiving the intervention are demographically matched with five comparison schools where students receive instruction guided by a menu of approved curricular approaches. Mean student outcomes for these two curriculum conditions are compared for two iterations of the implemented units (two consecutive years). We disaggregate the outcome data in order to find out if mean scores hide data patterns indicating that some groups of students are disadvantaged by either curriculum condition. We also study an enactment of the intervention in one classroom, in depth, using video ethnography. This provides a sense of how the unit is actually functioning from the students’ viewpoints, and can help guide our interpretations of the quasi-experimental group data.

If outcome measures indicate that the new unit is more effective than the comparison condition, the research moves to the scale-up phase in which the unit is progressively introduced to fifteen additional middle schools, then to all thirty-six MCPS middle schools in the subsequent years of the study. Thus, in this sense, scale-up means the deliberate expansion to many settings of an externally developed intervention that has been previously used in a number of settings (Stringfield and Datnow 1998, as cited in Coburn 2003). Although the focus of this research is on scale-up in this basic sense, this university/school district research partnership also satisfies aspects of a more expansive and ambitious definition of scale-up—one that includes a fundamental change in the core of instruction, and the relationship among students, teachers, and knowledge (Coburn 2003; Elmore 1996).

This study can also be viewed as research on the implementation and scale-up of education policy (Spillane, Reiser, and Reimer 2002) because its impli-
cations extend beyond the boundaries of this university/school district collaboration. The curriculum units chosen for implementation have been selected by applying a set of criteria developed by the American Association for the Advancement of Science (AAAS) Project 2061. Project 2061 launched a major effort to identify curriculum materials aligned with benchmarks that meet a rigorous set of criteria consistent with current theories of learning, and that have content-specific instructional strategies that support learning (Kesidou and Roseman 2002). The Curriculum Analysis of Project 2061 consists of categories of questions for analyzing a unit of instruction. The first step involves content analysis and ensures that the curriculum unit focuses on a discrete science benchmark or set of target ideas, at an age-appropriate level. The second step is an instructional analysis that inquires if the curriculum unit: (a) starts from ideas that are familiar or interesting to children; (b) explicitly conveys a sense of purpose; (c) takes into account student ideas, specifies prerequisite knowledge and skills, and conveys suggestions for teachers to find out what their students think about the phenomena related to the benchmark; (d) is engaging; (e) provides for firsthand experiences with phenomena; and (f) has students represent their own ideas about phenomena and practice using the acquired knowledge and skills in varied contexts (Kesidou and Roseman 2002; Roseman, Kesidou, and Stern 1996). (See the appendix to this chapter for a complete set of criteria.)

All three of the middle school science curriculum units in this scale-up study have been rated according to the Project 2061 criteria. All three have stronger “instructional support profiles” than is typical, and each unit focuses on a particular curriculum standard or benchmark (forces and motion, reasons for the Earth’s seasons, conservation of matter). Research evidence (AAAS 1993) shows that the concepts targeted by each unit are challenging, not only for middle school students, but also for adults.

However, curriculum analyses conducted by teams of experts in settings divorced from classrooms, no matter how well informed and meticulous, may fall short of clarifying “what really works” for diverse learners. Culture and language cannot be separated from curricula, “content learning,” or the classroom environment (Lee and Songer 2003; Lynch 2000, 2001; Moje 2001a; Warren et al. 2001). What works in one context may not in another, at least not without modifications. For this study, the important questions are: What actually works and for whom? How well does it work and why? What modifications may be necessary to make a unit of study more effective for specific groups of diverse learners? And finally, what happens to an implemented unit when it goes to scale? Is its effectiveness enhanced or diluted by the process of scale-up? If science curriculum units with these specified characteristics can be shown to be more effective than others using disaggregated outcome data, and if video analyses help us to
understand how specific instructional characteristics contribute to student success (or failure), then there are important policy implications for this study.

Research Questions, Design, and Scale-Up Plan

Researchers at the George Washington University (GWU) and science educators in Montgomery County Public Schools (MCPS) received an Interagency Educational Research Initiative (IERI) planning grant, with a follow-on five-year research grant, *Scaling up highly rated curriculum units for diverse student populations: Using evidence to close achievement gaps* (Lynch et al. 2001, 2002). Data collection for each question is spread over year 0 (planning grant year) through year 5, and not all questions are examined every year. The research questions and the schedule for addressing them are:

1. Does the implementation of a highly rated science curriculum unit result in higher mean scores (on student outcome measures) than the mean scores of students in the comparison condition? Does disaggregating outcome data and testing for interactions between demographic groupings and curriculum condition reveal important patterns not captured in the reporting of mean scores of treatment and comparison groups (years 0, 1, 2, 3)?

2. Using ethnographic methods to analyze a complete enactment of a unit for a group of four middle school students, how does the unit function in a diverse middle school classroom (years 0, 1, 2, 3, 4)?

3. The Impact of Experience: Does the effectiveness of curriculum units increase as schools and teachers become more experienced by implementing them from year to year (years 0, 1, 2, 3)?

4. The Impact of Scale: Is there a difference in the effectiveness of highly rated curriculum units when comparing small-scale treatments (i.e., five treatment schools) with large-scale treatments (i.e., twenty or thirty-six treatment schools) (years 3, 5)?

5. Fidelity of Implementation: Does the degree to which curriculum units are taught as intended by their authors (measured by a fidelity of implementation classroom observation instrument) affect student outcomes (year 4)?

One of the interventions under investigation in this study, *Chemistry That Applies (CTA)* (Michigan Science Education Resource Project 1993), was introduced to eighth graders in five MCPS middle schools in year 0 (the planning grant year, 2001–2002). *CTA* is a six- to ten-week curriculum unit targeting a benchmark/standard on conservation of matter in chemical reactions and how the idea of atoms helps to explain this phenomenon.
Five treatment schools had been matched with five comparison schools that used a variety of approved curricular options to teach this benchmark/standard. In year 1 (2002–2003), the implementation study was replicated in the same ten middle schools. The results were encouraging and will be discussed in more detail below. In the subsequent two years, CTA will scale up to twenty schools, and then to all thirty-six MCPS middle schools.

In year 2 (2003–2004), our study calls for the addition of two new highly rated curriculum units: ARIES: Exploring Motion and Forces: Speed, Acceleration, and Friction (Harvard-Smithsonian Center for Astrophysics 2001) for sixth-grade students in five middle schools and Great Explorations in Math and Science (GEMS): The Reasons for Seasons: Sun–Earth Connections (Lawrence Hall of Science 2000) for seventh graders in another five schools. To select this set of study schools, MCPS’s thirty-six middle schools were placed in five “pools,” such that each pool had a similar demographic and achievement profile. Then we randomly selected two schools from each of the pools, and randomly assigned Motion and Forces to one-half of the pair, and Seasons to the other. The comparison condition consisted of the same set of schools, counterbalancing for the curriculum unit being implemented (i.e., if a school is assigned sixth-grade Motion and Forces as a treatment, then its seventh-grade science classes become the comparison group for the Seasons unit). In this way, the study aims to generalize to all schools in the district.

To prepare teachers to implement the units, in summer of 2003 we conducted a two-day professional development workshop. The workshop was led by authors of the highly rated units or by a highly skilled professional developer sent by the publisher. We also presented information on the goals of the scale-up research, the results from year 0, and discussed the importance of good data collection efforts by teachers. The main focus of the workshop, however, was on understanding and working through the lessons in highly rated curriculum units with the professional developers who were assisted by Implementation Teams of MCPS teachers (about six teachers per unit). The Implementation Teams became the on-site professional development leaders after the publishers’ specialists returned home. The Implementation Teams help any teachers who did not attend the summer workshop learn how to use the units, and also answer implementation questions as they arise, work with researchers to provide feedback from the research, and engage participating teachers in discussions about the efficacy of the units, especially for diverse learners.

**CTA Study Results Years 0 and 1**

The student sample selected for study in years 0 and 1 was different from the population targeted for years 2 to 5. In year 0, we oversampled ten MCPS middle schools with the highest levels of ethnic, linguistic, and
socioeconomic diversity, creating five matched pairs of schools (with about 3,000 eighth-grade students). Half of each pair was randomly assigned to either treatment (CTA) or to comparison conditions. This sampling method produced two equivalent samples, rich in diversity, with enough students to provide power for significance tests on disaggregated subsamples. Students in both curriculum conditions were given pre-, post-, and delayed posttests for outcome measures, including the Conservation of Matter Assessment and five scales that measured students’ self-reports of engagement and goal orientation (Lynch et al. 2005; Pyke et al. 2003).

The results from this quasi-experimental study showed that students who used CTA had significantly higher scores on the Conservation of Matter Assessment than did those in the comparison groups. In order to ascertain if overall effects masked a different pattern of results for various groupings of students, we performed a series of ANCOVAs for each demographic grouping—Free and Reduced Meals Status (FARMS), English for Speakers of Other Languages (ESOL) status, ethnicity, and gender. We found significant main effects for the curriculum condition in all analyses, favoring CTA (with moderate effect sizes). All groupings listed above also had significant main effects for subgroups (except for gender, where the scores of girls and boys were about the same). For example, for the FARMS condition, students who had never been eligible for FARMS (Never) significantly outscored those who had once been eligible (Prior) and both of those subgroups outscored those who were currently eligible for FARMS (Now). There were no significant interactions between curriculum condition and the groupings analyzed, except for the ESOL condition. Students who were currently in ESOL (Now) had similar Conservation of Matter Assessment scores to their peers in the comparison condition. Prior ESOL students with CTA, however, outscored their Prior comparison peers, as well as Never comparison students (those who had never been eligible for ESOL). Other outcome measures for engagement and goal orientation either were not statistically significant or were statistically significant favoring CTA students in all groupings, with low effect sizes (Pyke et al. 2003).

To explore how CTA functioned in a classroom of diverse students, we videotaped a group of four students in one classroom for one complete enactment of the curriculum unit. The students’ speech and selected activities were carefully recorded, transcribed, and digitally linked to the video. This portion of our study sought to explore how a diverse group of four students worked with one another and their teacher as they experienced CTA. Each student in the video is not intended to “represent” a demographic group. CTA offered diverse students varied and numerous entries to understanding the target concept. Students seemed to form a functional group with the aim of marching through the CTA lessons. The students as-
sumed different roles as the curriculum was enacted. Overall, our data show students with measurably distinctive but consistent patterns of verbal and nonverbal interaction as they encountered a CTA curriculum unit and responded to its requirements. As the unit progressed, all four students exhibited increasing levels of involvement, science talk, and discussions consistent with science habits of mind (AAAS 1993). We documented individual variations in the amount of talking, manipulating of materials, use of scientific terms, seeking clarification, giving or taking direction, and other characteristics of student discourse. Students’ real-time interpretations of CTA’s lessons provide an insider’s perspective to criteria. Our goal is to continue to make connections between the ethnographic and quasi-experimental portions of the study, in order to explain how CTA functions for diverse learners.

Preliminary analyses of the year 1 quasi-experimental data on the Conservation of Matter Assessment indicate that students with CTA, once again, had significantly higher mean scores than their counterparts in the regular curriculum condition. This was true for the entire sample, as well as for subgroups of students taking CTA. Later in this chapter, we describe some of these preliminary results for year 1 in more detail. We also videotaped another enactment of CTA in a different middle school than in year 0.

In summary, this research on the scale-up of highly rated curriculum materials is progressing well. We have formed an effective school district/university collaborative where considerable numbers of teachers, students, and schools are involved. The study is on schedule, and no major problems have been encountered as we enter the third year of research. The school/university divide is slowly but surely being bridged as we work together to move the project forward and share our areas of relative strength. We have research results that support our hypotheses and are contributing to a reduction in achievement gaps. The ethnographic study is developing means to explore how the instructional elements of CTA contribute to improved student outcomes. The next section of this chapter attempts to identify features of our work that are making this scale-up research possible.

**FEATURES OF SCALE-UP RESEARCH: INTERVENTIONS, ENVIRONMENTS, AND INTERACTIONS**

Theories on how and why education interventions go to scale, or fail to do so, are still developing (Coburn 2003; Cohen, Raudenbush, and Ball 2003). Theoretical anchors and perspectives for thinking about the scale-up of educational interventions offered by Cohen, Ball, and Barnes (2001) provide a useful structure for discussing our work. Three crucial dimensions affecting scale-up are: the nature of the intervention’s design, the features of the
instructional environment that affect scale-up, and the interaction of designs and environments, including interveners, schools, and teachers.

The interventions in our study are three middle school science curriculum units that have much in common with one another, and have identical attributes for instructional support. Although the “interventions,” written small, are the three curriculum units, a more expansive view of the interventions includes their characteristics for instructional support that are specified in the Project 2061 Curriculum Analysis (see appendix). We believe that these curricular attributes contribute substantially to successful implementation and scale-up. In addition, a main goal of our study is to provide research evidence, both experimental and ethnographic, that helps researchers and MCPS science educators understand how the curriculum units function to improve the learning environment and outcomes for a diverse population. We want to better understand how the Project 2061 Curriculum Analysis criteria affect students in science classrooms. In addition, our research may lead to the identification of other characteristics of curriculum materials not included in the Project 2061 criteria that contribute to the success or failure of the interventions.

Intervention Design

In order for a school district to even begin to consider the adoption of an intervention, especially on a large scale, the intervention must possess certain design characteristics, each of which is likely to affect the course of scale-up (Cohen, Ball, and Barnes 2001). These characteristics are:

1. Instructional focus—the extent to which the design targets instruction directly, as opposed to indirect methods such as resource allocation or school organization
2. Comprehensiveness—the degree to which an intervention includes many elements of instruction and related facets of schooling such as grouping, literacy demands, use of assessment information, teacher collective work, or involvement of parents
3. Academic demand—the difficulty of academic content for both teachers and students and the complexity of associated components of instruction such as sequence, or the relationship of assessment information to instructional strategies
4. Organizational demand—the extent to which an elaboration requires extensive collaboration and cooperation
5. Elaboration—the degree to which an intervention’s target goals and practices are well articulated and support the implementers who are attempting to use them
6. Development—the extent to which the intervention provides direct support for learning and change, implied by the design’s complexity

We discuss the design features of the interventions used in our study briefly, noting that it is the interaction between environments and interventions that gets at the fundamentals that allow this research on scale-up to occur. The science curriculum units implemented in our study have currency in part because they were developed by university organizations (Harvard, University of California, Berkeley, and Michigan State) through NSF grants (CTA is the exception, funded by well-known private foundations). These units were developed in the last ten years, and can be characterized as “reform-based.” They are probably more alike than different because they share similar theoretical and epistemological perspectives—they all appear to be constructivist in philosophy, based on notions of conceptual change, and require that students interact with one another in groups to make sense of the evidence and build cases. Concomitantly, their instructional approaches are similar; they are “hands-on” units that ask students to learn by experiencing phenomena directly and gathering evidence leading to understanding.

Our evidence for these claims is based on a curriculum analysis process designed by Project 2061 (Kesidou and Roseman 2002) that has considerable concordance with the features of interventions listed above in Cohen, Ball, and Barnes’s (2001) frame. Table 4.1, the results of using the Project 2061 Curriculum Analysis, compares the interventions. They have similar instructional characteristics in several areas, but certainly not all. In addition, for comparison purposes, table 4.1 contains the ratings of a more traditional, commercial middle school physical science textbook. Its low ratings show why we refer to the intervention curriculum units in our study as “highly rated.”

The three intervention science curriculum materials all directly target benchmark/standard-based ideas that are difficult for middle school students, as well as adults, to learn (AAAS 1993). The units’ expectations are for student conceptual change. They have teachers’ guides that explain why specific instructional approaches are taken. The research base provided in each unit shows they are aimed appropriately at their target age group. These units have the potential to be “comprehensive” (the extent to which an intervention includes changes in instructional practices, grouping, assessment, and use of assessment information). But the comprehensiveness of enactments would seem dependent on teachers’ individual interpretations. Some teachers might implement the units as they are intended, and they could be judged “comprehensive.” Other teachers might march through the lessons not attending to features that require thoughtful integration of components that could change the educational core.
Table 4.1. Criterion-Level Instructional Analysis Ratings for Chemistry That Applies*,
GEMS: The Real Reasons for Seasons**, ARIES: Exploring Motion and Forces***, and a
middle school physical science textbook published by Macmillan/McGraw-Hill*

<table>
<thead>
<tr>
<th>Instructional Categories / Ratings***</th>
<th>Chemistry That Applies</th>
<th>ARIES—Exploring Motion and Forces</th>
<th>GEMS—The Real Reason for Seasons</th>
<th>Macmillan/McGraw-Hill Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Identifying a Sense of Purpose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conveying unit purpose</td>
<td>o</td>
<td>o</td>
<td>N/R</td>
<td>o</td>
</tr>
<tr>
<td>Conveying lesson/activity purpose</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Justifying lesson/activity sequence</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>II. Taking Account of Student Ideas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending to prerequisite knowledge and skills</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Alerting teacher to commonly held ideas</td>
<td>o</td>
<td>o</td>
<td>N/R</td>
<td>o</td>
</tr>
<tr>
<td>Assisting teacher in identifying own ideas</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Addressing commonly held ideas</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>III. Engaging Students with Relevant Phenomena</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Providing variety of phenomena</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Providing vivid experiences</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>IV. Developing and Using Scientific Ideas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introducing terms meaningfully</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Representing ideas effectively</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Demonstrating use of knowledge</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Providing practice</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>V. Promoting Student Thinking about Phenomena, Experiences, and Knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraging students to explain their ideas</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Guiding student interpretation and reasoning</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Encouraging students to think about what they've learned</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>VI. Assessing Progress</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aligning assessment to goals</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Testing for understanding</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Using assessment to inform instruction</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>VII. Enhancing the Science Learning Environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Providing teacher content support</td>
<td>n/a</td>
<td>o</td>
<td>o</td>
<td>n/a</td>
</tr>
<tr>
<td>Encouraging curiosity and questioning</td>
<td>n/a</td>
<td>o</td>
<td>o</td>
<td>n/a</td>
</tr>
<tr>
<td>Supporting all students</td>
<td>n/a</td>
<td>o</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

*These ratings were done by Project 2061.
**These ratings were done by GWU/MCP8 SCALE-ap exams.
***Excellent=●; Very Good=●; Satisfactory=●; Fair=●; Poor=O; N/R=No Rating
The organizational demands of the units are high because they require the extensive use of laboratory equipment, the organization of students into groups in order to do the labs, and perhaps most important, the ability of teachers to guide students in the interpretation of the physical phenomena and related data collection so that new understanding (conceptual change) can occur. This means, for many teachers, a different, demanding approach to science teaching, albeit one encouraged by science education reform and supported by elaborated teachers’ guides. Indeed, a complaint that one sometimes hears about the units is that they are too “scripted.” However, their conceptual and organizational demands may make this concern trivial for many middle school science teachers who do not ordinarily teach science in this way. Their elaboration features may help such teachers to implement the units. In summary, these interventions are strong in several factors listed above, including: instructional focus, academic demand, elaboration, and development.

Environmental Features

Three features of the environment that affect scale-up, which are identified by Cohen, Ball, and Barnes (2001) are:

1. Coordination—the nature of the instructional guidance provided by school governance structures and its consistency with intervention’s goals, resulting in more support for the implementation of the intervention

2. Incentives for improved teacher practice and student performance—the accountability system adopted by states or local school districts for demanding academic performance and improved instruction (Strong incentives to adopt an intervention likely lead to a higher quality of implementation.)

3. Resources available for the implementation of the intervention—the capacity of individuals (the knowledge, skills, and will to use the intervention), the organization’s culture and structures, and the money, time, space, and ability to mobilize resources to support the implementation

MCPS is a large Maryland school district (136,000 students) located in the Washington, D.C., metropolitan area. It is perceived as one of the best large school districts in the area, if not the United States, due to the affluence and educational level of its community, its organization and facilities, and its performance rankings. For instance, MCPS is consistently among the top five school districts in the state of Maryland on state assessments. At the same time, MCPS is part of the Washington, D.C., metropolitan area and is
rapidly becoming more socioeconomically, ethnically, and linguistically diverse (Perlstein 2003). In MCPS, the largest proportion of new immigrants is Spanish speaking and blue collar (primarily from Central America). There are also large numbers of Koreans, Southeast Asians, and East Asians entering the schools. In the last five years, the proportion of white students fell below 50 percent, and there is no ethnic majority in many schools.

The MCPS superintendent, Dr. Jerry Weast, from his first days on the job set a clear and public priority of improving achievement for a diverse student population. If MCPS is to maintain its status as a top-performing school district, disaggregated achievement data indicate that it must do more to identify and implement effective curricular and instructional practices in science and mathematics for diverse learners. In a recent state assessment, there were slight decreases in achievement in eighth grade mathematics and science. For instance, in 1999, 63.3 percent of students met the eighth grade state science standard, while the 2000 data showed that only 62.4 percent did (Bulletin 2000).

The MCPS mathematics program has received a great deal more public scrutiny than science, but the lessons learned are not likely to have escaped MCPS science educators. Three separate, extensive evaluation studies of the mathematics program showed great variability in student achievement between and within schools, coupled with large and pervasive gaps in student achievement by ethnicity/race (Weast 2000). Moreover, schools with similar demographics showed differing patterns of achievement and a large variance in achievement gaps. A major factor in poor student performance—possibly greater than poverty or language development—is the structure and implementation of the mathematics curriculum (Weast 2000). No systematic relationship was found between various teacher attributes (level of education, certification, years of experience, etc.) and student performance, leading to the conclusion that what teachers do in the classroom is a more important factor than measurable attributes of training (Larsen 2000). However, Larsen's study was not able to pin down factors that made some teachers more effective than others. A second report indicated that students had diverse learning needs, which many teachers felt unprepared to address (Gross 2000). Finally, a major curriculum audit in K–12 mathematics (Phi Delta Kappa International 2000) demonstrated that the situation in MCPS mirrors the problems uncovered by the Third International Mathematics and Science Study (TIMSS) and other national studies—an unfocused mathematics curriculum, a plethora of textbooks, idiosyncratic curriculum implementation by teachers, and lack of alignment of curriculum and instruction with mandated assessments. The report concluded that priority attention be given to changes in expectations for minority students and the instructional behaviors of their teachers, and that this be done systematically through rigorous evaluation of curriculum and instruction (emphasis ours).
On the whole, then, the MCPS administrators’ and teachers’ concerns help create an environment to improve curriculum and instruction for diverse learners. This research on scale-up of highly rated curriculum materials is timely, particularly if it can help to close achievement gaps.

By most standards, MCPS middle school science is well resourced in terms of lab classrooms, equipment, and supplies. Science teachers are accustomed to receiving resources to support projects coordinated between science leadership from the central office and their building resource teachers (RTs), who are similar to department chairs. However, even an affluent school district has an ebb and flow of money for resources, and budgets are tightly managed. The leadership, consequently, makes a point of finding grants to supplement budgeted funds. The structural support provided throughout the system is strong, and flows directly from the science leadership to coordinate a reform agenda for K–12 science. For instance, MCPS has an individual whose full-time job is to make up “science kits” for teachers. This saves the school district money (commercial kits are more expensive) and encourages teachers to conduct hands-on lessons.

Perhaps the most important environmental feature affecting our scale-up research is the recent history of middle school science educators in systemic reform efforts. In 1997, MCPS received an NSF-funded five-year, $1,400,000 Local Systemic Change grant, called Science Connections. Its goal was to provide comprehensive training to middle school science teachers in science content, pedagogy, performance-based teaching and assessment, technology applications, and new science units adopted from NSF and other national curriculum projects (emphasis ours), with a focus on inquiry-oriented instruction (Hagan 2000). Although Science Connections devoted much effort to the implementation of new reform-based curriculum materials, virtually none have scaled up to reach students in thirty-six middle schools. The reason for this seems to have been the lack of a causal evaluation design that could discern the effectiveness of each unit on student achievement. (Evaluation efforts were more global and sought to capture the effects of the many concurrent Science Connections activities on student outcomes.) Consequently, there was no basis on which to make curriculum decisions, unit by unit. The units that teachers saw as most successful were retained on the list of approved middle school science curriculum materials, and were implemented idiosyncratically across the district, while less popular units faded from use (B. Hansen, personal communication, September 19, 2003).

However, Science Connections resulted in other measurably positive effects. It created a trainer-of-trainers professional development model that identified a master science teacher (MST) for each MCPS middle school. The role of the MSTs was to disseminate improved teaching practices within their buildings, and over the life of the grant. There was a planned rotation of MSTs so that most teachers were involved. Data from a recent
study indicated that the more frequent the participation of MSTs in *Science Connections* workshops and meetings, the greater the improvement in science achievement at the school level. There was also some evidence showing the narrowing of achievement gaps in schools of MSTs with high participation rates (MCPS Science Office 2003). In addition, the MSTs became sophisticated in their understanding of reform goals and alignment issues, and generally responded well to requests for evaluation data from the MCPS Office of Shared Accountability.

It seems fair to say that closing achievement gaps is a major concern of MCPS educators, including classroom science teachers. In order to close gaps, there must be measures of student outcomes, but the state of Maryland dropped its state science assessment program in 2001, and new middle school science assessments are not slated for implementation until 2006. This situation has positive and negative effects on efforts to change science curriculum and instruction in MCPS. On one hand, the accountability system has temporarily been disconnected and incentives weakened, although MCPS does participate in NAEP and other science assessments less influential than the state science assessments. On the other, the lack of pressure for accountability gives science educators, from the top levels to the classroom teachers, the freedom to make sensible changes in curriculum and instruction, with a reprieve from immediate repercussions from within the school system and from without. Because the nature of the proposed state-level middle school science assessment system is not yet well defined, MCPS is temporarily free to develop science curriculum and instruction in ways that it sees as best for the district and in line with national reform efforts and internal guidelines, as it keeps a weather eye on assessment developments at the state level.

In summary, according to Cohen, Ball, and Barnes’s (2001) framing of environmental features important for scale-up, MCPS in general, and its science program in particular, have a healthy environmental profile for coordination and resources. The weakening of incentive structure, especially the state’s hiatus from science assessment, could either work for or against efforts to scale up interventions, depending on the alignment of our study interventions with the assessments under development.

**Interactions: Designs and Environments, Interveners, Implementers, and Schools**

Cohen, Ball, and Barnes (2001) point out that scale-up is not simply affected by the design of an intervention, or only by the features of the environment in which it is implemented. Rather, interactions between design and environment determine if an intervention is successfully implemented and goes to scale. Such interactions are the reflection of two sets of per-
spectives, those of the interveners and those of implementers. Our study's progress is located at the nexus of the problem space of these interactions and includes:

1. Interveners' perspectives—creating (or in our case, choosing) designs that: enable major instructional improvement; are affordable to create and/or implement; can be sustained or improved through time as they are implemented; are capable of recruiting motivated implementers; and do not so greatly increase costs as to make them unsustainable (Cohen, Ball, and Barnes 2001). We think that another feature is also crucial to this problem space: The interveners must have an intervention, the effects of which can be measured in ways that are convincing to implementers and other stakeholders. This makes the intervention defensible.

2. Implementers' perspectives—finding ways to improve instruction and secure resources, while disturbing current practices as little as possible; and maintaining familiar roles, social organization, and associated ideas about learners, schools, and instruction, while retaining a responsive stance toward agencies in the environment.

Selection of Intervention Curriculum Units

For this study, the implementers are MCPS science teachers. The interveners can be seen as the university researchers and MCPS science administrators who did not create the interventions, but chose them from extant curriculum materials. A driving interest of the researchers was to learn how a middle school curriculum unit highly rated according to Project 2061 criteria would fare in science classrooms characterized by high levels of student diversity. Thus, the first curriculum unit chosen for study, CTA, was selected because it was one of the very few middle school science curriculum units evaluated by Project 2061 that had a satisfactory profile. Moreover, it was the only one rated by Project 2061 that aligned with the state of Maryland and MCPS's middle school science curriculum frameworks. Some MCPS middle school science educators were familiar with the Project 2061 initiatives, including the Curriculum Analysis that was introduced to teachers in Science Connections professional development workshops conducted by Project 2061 staff. However, the Curriculum Analysis was not seen as sufficiently practical or useful to MCPS educators to have warranted its incorporation into the curriculum materials selection process until the advent of this study.

Although our study introduced CTA to the five treatment schools for the first time in 2001, this was not CTA's debut in MCPS. A handful of teachers had picked up and implemented CTA during their participation in Science
Connections, which “pilot tested” a variety of curriculum materials but did not evaluate their effects on student outcomes. Other teachers found chemistry units from other publishers more to their liking, and implemented those units. This entailed buying the books and commercial equipment kits. Thus, within MCPS, there is a little pro-CTA bias, and some anti-CTA bias, especially among teachers who had invested in other curriculum materials for chemistry.

The selection of the other two curriculum units in this study, GEMS Reasons for the Seasons and ARIES Motion and Forces, provided a different set of challenges. The MCPS K–12 science director had a standing concern that astronomy was neglected at the middle school level and asked us to find two acceptable astronomy units for implementation, one for sixth and the other for seventh grade. After much searching, we concluded that we could find only one commercially available unit (Reasons for the Seasons) that fit the standards of learning frameworks for Maryland and the MCPS, and also had a prayer of being highly rated according to the Project 2061 Curriculum Analysis, and was affordable if it went to scale in thirty-six MCPS middle schools. Some units approved by MCPS and already in use in some schools were rejected for our study, given their costs if scaled up. Such units often require the purchase of student textbooks, student workbooks, teachers’ guides, and materials kits for a unit of study that would last only several weeks. We estimated the cost to the school district for one such unit to be about a quarter of a million dollars over the life of the study, and afterward, materials would still need to be replenished. Such units were too expensive to be considered sustainable at scale, and were eliminated from consideration.

We set out in search of another unit of instruction that might fit the lock and key mechanism that had (perhaps inadvertently) evolved when the state of Maryland and MCPS created science standards of learning and well-defined curriculum frameworks. It was incredibly difficult to find units for particular standards at a particular grade level, given assumptions of prerequisite knowledge and sequencing that further complicated the search. For instance, if a curriculum unit was created by a publisher for one large state, it would be unlikely to fit curriculum frameworks developed by different states—a lock-and-key situation that has developed in spite of national science standards/benchmarks. After cursory examination and rejection of many, many units on the basis of our three criteria (alignment, the potential for a high rating, and cost), we finally located the sixth grade-level ARIES physical science unit, Motion and Forces, which was aligned with MCPS’s curriculum framework and seemed similar to CTA in its instructional approach.

Project 2061 had not evaluated these two new units using their Curriculum Analysis (the units were created after Project 2061 focused on middle
school science). Consequently, in order to find out how these units compared to CTA based upon the same set of well-defined criteria, we organized two teams of science educators, and adhering as closely as possible to Project 2061 procedures, analyzed the two units (Faubert et al. 2003; Ochsendorf et al. 2003). The ratings resulting from our teams’ analyses can be seen in table 4.1. Although the Curriculum Analyses for the two new units did not result in as strong instructional profiles as CTA’s, table 4.1 shows that they have more in common with one another than with more typical middle school physical science textbooks.

So, the first important interaction between intervention and environment involved choosing from candidate interventions—middle school science curriculum units—most of which simply did not fit MCPS and Maryland standards because of misalignment of topic focus and grade level, and their prohibitive costs at full scale. In addition, for the handful of units that were affordable and had topic and grade level matches, a perfunctory examination of their instructional approaches showed that they would not meet Project 2061 Curriculum Analysis criteria. Although publishers of science curriculum materials might disagree with the Project 2061 criteria, an inescapable fact is that the standards movement, as interpreted by states and school districts, has created a situation where it is difficult to create standards-based materials suitable for many states. Consequently, the immediate future of general, traditional textbooks seems secure. This seems the opposite of the intent of systemic reform, and has not gone unnoted (Cohen, Raudenbush, and Ball 2003).

Incentives for Implementation and Scale-Up

MCPS is an enormous school district with a well-coordinated central administration, headed by a superintendent of schools who has a reputation as a strong and demanding leader. The core of support for science curriculum and instruction is a K–12 science director (co-PI for this study), a middle school science coordinator, and a project director (hired for this study). Historically, these individuals (and their predecessors) are known for their activist approach to improvements in middle school science curriculum and instruction, and are able to assert some influence on state education policy, in part because of the stature of MCPS. In addition to the centralized science leadership, each school has designated resource teachers (RTs), similar to department chairs, responsible for the oversight of science curriculum at their schools. They receive incentives of released time and extra pay for their efforts. The central science administration meets frequently to share its initiatives with RTs and other teachers to move a reform-based agenda. All necessarily must keep an eye on the state accountability system for middle school science, on hold since 2001. The current study relies increasingly on the RTs
for the coordination of implementations and data collection at the level of individual middle schools, as the units are scaled up.

In addition to RT network, the Science Connections grant created a rotating group of master science teachers (MSTs) who assisted in instructional efforts, large and small, at the school level. The MST network dissolved in 2001 at the close of the Science Connections grant, but working relationships among these teachers remain. This strong network of science educators is one of the most important influences on the progress of the current study. Without this experienced network (an excellent example of distributed knowledge), it seems likely that our current efforts to implement relatively demanding science curriculum units and systematically collect assessment data would be much more difficult. For these professionally active teachers, a powerful incentive is the opportunity to work with one another and provide professional development to other science teachers in the system.

Aware of this, we built Implementation Teams into the current project, patterned somewhat on the old MST network. Implementation Team teachers (like MSTs) are paid about $60 per hour to organize much of the professional development of science teachers who will teach the new highly rated units. We also had learned from experience that some, but not most, science teachers were interested in the research aspects of the project. So we created a small Research Team of four science teachers who work with scientists and researchers on demanding grant-related activities such as the curriculum analyses and development of assessments.

Together, the Implementation and Research Teams act as liaisons between intereners and implementers, and add credibility and buy-in for the project. In order to offer further opportunities to develop these science teachers, we provide some funding for travel to the National Science Teachers Association annual conference, so that they can formally discuss this scale-up research from their points of view. Some joint writing projects are also under way.

According to the MCPS project director (B. Hansen, personal communication, September 19, 2003), a different sort of incentive is also operating for science teachers, with implications for the sustainability of the interventions over time. All three of the intervention science units require student workbooks and lots of materials for hands-on investigations that would take considerable time for teachers to gather and organize. Consequently, the MCPS project director purchases the written materials centrally, and has materials kits made up for each unit. This is a considerable incentive for implementation of the units, especially attractive for less experienced teachers, teachers new to the system, or teachers whose licensure is in elementary, special, or ESOL education rather than science education. All such teachers could face considerable challenges in putting together the materials required for a hands-on unit of instruction, and implementing it.
The Nature and Role of Evidence

The results of the year 0 planning grant (2001–2002) which focused on CTA seemed critical to the progress of this research program for collaborators at MCPS, GWU, and funding agencies. For this pilot study of CTA, we oversampled to select five pairs of schools with the highest levels of socio-economic, ethnic, and linguistic diversity in MCPS. MCPS educators and the public in general were very aware of the achievement gaps in MCPS (including state test scores, SAT scores, scores on MCPS assessments, and participation in honors courses). There has also been a steady barrage of coverage on this issue in the Washington Post, local publications, and MCPS internal reports. A great deal of the reporting focused on gaps between African American and Hispanic students versus white and Asian American students. Ethnic backgrounds are confounded with socioeconomic status and literacy concerns.

Consequently, when we constructed the population for the pilot study, we knew that if we could show that CTA helped close the gaps for Hispanics or African Americans, students eligible for FARMS, or students who were learning English, then the chances for moving this research agenda forward in MCPS and securing funding for the work were greatly enhanced. We were able to report that CTA produced statistically significant, higher scores for treatment students than comparison students on outcome measures, especially the Conservation of Matter Assessment that targeted the benchmark. We used a variety of means to display these results, but an example of the methods of display that seemed to make the most impact on implementers can be seen below in figure 4.1.

We were careful to couch figure 4.1 in the context of overall assessment results and to use accurate statistical language. However, the words “statistically significant” and the descriptive data illustrated in figure 4.1 seemed to have the most impact on educators, while discussion of interactions and effect sizes seemed to have less impact. CTA did not close gaps—slopes for different groups appeared to be nearly parallel in many year 0 analyses while the comparison curriculum condition seemed to widen gaps. In some instances posttest scores were not much higher than pretest scores for the subgroups of students of most concern (see Lynch et al. 2005). In contrast, the only subgroup in year 0 with flat scores who had taken CTA were the Now ESOL students.

Not only did the year 0 planning grant provide encouraging research results, it seemed to prove to researchers and science educators in the school district that we could work well with one another, sharing a common set of goals that could be forwarded by collaboration. At the same time that we were providing specific evidence about the efficacy of CTA in teaching the conservation of matter benchmark and presenting the results in various
forums (science teachers’ meetings, administrators’ meetings, MCPS and GWU websites, and professional conferences), we made a strategic decision not to link the research data with specific schools or teachers. Although one could argue with this strategy from a school improvement standpoint, we think that if the project became a race between individual teachers and individual schools, we would lose both teacher support and focus on the research problem: What is it about these highly rated curriculum units that helps diverse students to achieve improved outcomes?

Although teachers and administrators in schools seem inundated with accountability systems, measuring the impact of these curriculum materials appears to be a powerful aspect of scale-up. In contrast, *Science Connections* had implemented similar units in MCPS middle schools for five years, but they did not “stick.” As the interventions in this study go to scale, MCPS educators seem to be influenced by the act of measuring itself. We began this chapter with a quote on the Heisenberg Uncertainty Principle, analogous to our collaborative research on scale-up. The act of measuring, and how it is done, seems likely to affect research outcomes.

The assessments developed for this study are narrowly focused on the ideas necessary to understand conservation of matter, *Motion and Forces*, or *Reasons for the Seasons*. We can provide feedback on student misconceptions for these topics. Some MCPS science teachers have shown interest in the design of this study, have critiqued it, and have offered alternative approaches and explanations for the data patterns. But even if a teacher is completely

---

**Figure 4.1. Results for Conservation of Matter Assessment and Ethnicity in Year 0, CTA versus Comparison (Com) Groups**
convinced that he or she can do a better job of teaching conservation of matter than teachers using CTA, for example, it is hard to argue with the disaggregated results for the large numbers of students across schools.

As we develop video data for use in professional development settings for middle school science teachers, we hope that selected clips can encourage rich discussions on how the instructional attributes of the interventions contribute to student understanding of the target benchmarks/standards. As the study progresses, video may also help us identify where each curriculum unit falls short, and how it might be modified for some groups of students.

**Fidelity of Implementation**

One of our research objectives was to measure the impact of fidelity of implementation on student outcomes for two of the interventions (Motion and Forces and Seasons) in year 4 (2005–2006) when both units reach twenty middle schools. We hypothesize that the greater the fidelity of implementation, the better student outcomes for all demographic groupings of students. This hypothesis runs counter to a position that can be found in the literature on implementation or scale-up research that assumes that during scale-up, implementers must modify an intervention for their particular environment, especially for diverse learners (Coburn 2003; Cohen, Raudenbush, and Ball 2003; Lee and Songer 2003). While this may perhaps be true, it is a question worth exploring empirically.

Our research leads us to the rival hypothesis (above) due to our recent experiences with the implementation of CTA: In year 0, we conducted a pilot study of CTA in five treatment schools rich in diversity. We did not have an instrument to measure fidelity of implementation in year 0, but did manage to visit over half of the CTA teachers. There seemed to be a wide range of fidelity, and we know that a few teachers did not complete the unit (see Cohen, Raudenbush, and Ball 2003 on “use” of interventions). We also assumed that implementing teachers would need to make many adjustments to CTA, given the high proportions of their students who were learning English, had disabilities, or who were reading below grade level. In order to capture these modifications and as a way of monitoring fidelity of implementation of CTA, we created an electronic website that asked teachers to comment on the modifications that they were making to the unit, lesson by lesson (eighteen lessons in all).

Unfortunately, teachers simply did not use the website for a variety of reasons, not the least of which were problems with Internet access. But at our follow-up meetings, it became clear that teachers had their hands full simply implementing CTA for the first time. Most of their energy was spent on management. They had to learn how to organize lab equipment and student groups, adjust lessons that yielded misleading data, and for a few, learn how
to use safety equipment and properly dispose of chemicals. The demands of 
CTA were such that creating curricular modifications for groups of students 
with different needs simply was not a priority.

Given these circumstances—encouraging teachers to make modifications 
for diverse learners but finding little evidence that they did so due to 
CTA’s implementation demands—we were surprised that year 0 data 
showed that CTA produced higher outcome scores for all subgroupings of 
students (except Now ESOL students), compared to the other curriculum 
condition. This was especially unexpected given that MCPS had launched 
a major effort to train all teachers on something called differentiation, or 
modifying curriculum and instruction to meet the needs of diverse learners. Presumably, teachers in the comparison conditions had more time to 
differentiate instruction because they were not implementing a new cur-
riculum unit.

Because year 0 results provide no instance of CTA’s producing poorer out-
comes than the comparison condition, we urged teachers to focus on implement- 
ing the unit with fidelity in the year 1 replication, rather than creating 
modifications for diverse learners. What “implementing with fidelity” 
meant became a frequent topic at our meetings with the science educators 
involved. Our operational definition came to mean using all the unit’s les-
sions in sequence without interjecting accessory material, and reading and 
following the CTA teachers’ guide.

After two years of experience with CTA, a meeting with implementing 
teachers resulted in consensus on only three modifications for all students 
using CTA: (a) changing the laboratory procedure of one lesson that did not 
yield convincing data; (b) having the teacher demonstrate one of the chem-
ical reactions that involved a butane lighter rather than having the students 
do it themselves (exploding lighters!); and (c) disembedding the student 
questions from the CTA workbook and placing them on separate work-
sheets so that students could respond more easily. The only modification 
linked to student diversity was the suggestion that CTA be translated into 
Spanish.

We currently are developing and testing a fidelity-of-implementation in-
strument based upon Project 2061 Curriculum Analysis criteria. This in-
strument will eventually be used with all three intervention units in this 
study and is designed for classroom observations by trained raters. In year 
1, raters made two visits to classrooms of each teacher implementing CTA. 
Because the instrument is still under development, we report here only 
that all teachers visited were using CTA, and that a range of fidelity was ob-
erved.

In order to implement CTA with fidelity, we assumed that teachers would 
rely heavily on the teachers’ guide because the units’ rationales and struc-
tures are complicated. The guide coaches teachers elaborately about such instructional issues as: student misconceptions, how to guide student discussions, not providing answers to questions prematurely, and so forth. Midway through the CTA implementation in year 1, we learned at a teachers’ meeting that most of them had “lost” their teachers’ guides and were not using them at all. To these teachers, fidelity of implementation did not seem to mean close reading of a teachers’ guide in order to direct student learning, as explicitly intended by CTA. We leave it to the writers of curriculum materials to ponder the implications of this observation, but suspect that teachers’ notions of fidelity of implementation are acquired primarily through using the student guides or texts, past experience, and professional development workshops about a new unit, rather than by adhering to a teachers’ guide.

We offer two additional comments on fidelity of implementation. The first is that the act of measuring fidelity of implementation is likely to increase it (Heisenberg again). It becomes a measure of teacher accountability. Our research on scale-up schedules the study of fidelity of implementation in year 4 when substantial samples of teachers will have their classrooms observed several times as they implement new units. We think that we can get valid data to answer our question about the relationship between fidelity and student outcomes. But the results may not necessarily be generalizable to the classrooms of teachers who were not visited. At this time, we do not have a solution to this methodological problem.

The second comment is on the tension between scale-up research and the modification of units for particular environments in order to better reach more students, especially diverse learners. How can researchers measure the efficacy of an intervention if it is being constantly modified, resulting in permutations that increase “infidelity”? Is it possible to implement a unit and change it at the same time, and evaluate its efficacy? In response to this conundrum, we ask teachers to implement with fidelity to the best of their abilities (barring a situation where students’ interests are clearly being hurt) during the study. As the study draws to a close, we may work with implementers to systematically capture consensus modifications. At that point, it might be appropriate to initiate a new study of scale-up that intentionally compares student outcomes when modifications are encouraged versus when they are not. At this time, we find it hard to reconcile conflicting commonplace notions about conducting scale-up research: measuring the effects of an intervention unit implemented with fidelity versus capturing what teachers do to modify a curriculum unit to meet their (perceptions of) diverse students’ needs. This is perhaps where ethnographic research can play a more significant revelatory role than experimental research (Cohen, Raudenbush, and Ball 2003).
Experience

One of our research questions is to measure the impact of experience on student outcomes. This is a direct, empirical measure of the interaction between implementers, the environment, and the intervention. By "experience" we do not mean the experience of an individual teacher with an intervention, because teacher mobility is high in MCPS middle schools. We operationally define "experience" as a school’s science department’s experience with a curriculum unit over time. The organizational structure of MCPS middle school science allows this definition because the resource teachers provide a strong support network for teachers new to their schools, and there is also ongoing professional development for teachers provided directly through this project.

Our first test of the experience factor compares the CTA outcome data for year 0 with year 1 (two further measures of experience are scheduled for year 3 and involve two other curriculum units). We are in the preliminary stages of analyses for year 1 data, and here will use only descriptive statistics that are straightforward. Table 4.2 shows that the pretest scores of treatment and comparison groups were similar to one another in years 0 and 1, but in year 1 the pretest means were about seven points higher than in year 0. This may be attributed to the schools’ past experience with the study and the fact that all teachers know that conservation of matter will be assessed specifically during the course of the year. Although teachers were asked to teach the chemistry unit in a specified marking period, it is likely that conservation-of-matter ideas come up at other times during eighth grade science. Teachers might emphasize these ideas more, knowing that they are the focus of this study. Another explanation for the increase in pretest scores is that in year 0 the chemistry unit was taught during the second quarter, but in year 1 it was taught in the fourth quarter. Higher pretest scores might simply reflect student maturation.

<table>
<thead>
<tr>
<th>Curriculum Condition</th>
<th>CTA</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>mean</td>
</tr>
<tr>
<td>Year 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>1118</td>
<td>21.68</td>
</tr>
<tr>
<td>Posttest</td>
<td>1052</td>
<td>42.68</td>
</tr>
<tr>
<td>Year 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>1227</td>
<td>28.15</td>
</tr>
<tr>
<td>Posttest</td>
<td>1248</td>
<td>48.71</td>
</tr>
</tbody>
</table>
The sample sizes are larger in year 1 than year 0, even though the same schools were involved. Because teachers knew that a long-term collaborative research study was under way with assessments’ administration overseen by MCPS evaluators rather than university researchers, perhaps they were more careful in gathering complete data.

For posttest data, table 4.2 shows that CTA students gained about twenty points in both years 0 and 1, and comparison students gained about eleven points in year 0 and fourteen points in year 1. We hypothesized that CTA schools would show greater student gains with their increased experience, but they held steady. For comparison schools, the likely explanation is that as teachers became more aware of the study, they spent more time and/or effort in teaching conservation of matter in year 1 than in year 0.

However, as we examine the disaggregated data for years 0 and 1, an interesting picture of the effects of schools’ experience is revealed. We will use one example, the disaggregated ESOL descriptive data for year 1 (see figure 4.2).

The first two rows of columns (reading from left to right rather than vertically) show pretest scores in years 0 and 1, for ESOL Never, Prior, and Now students. The pretest scores for the subgroupings of students are about equal in years 0 and 1. The third row of columns shows posttest scores for comparison condition, and the fourth row for CTA students. In year 0, Now ESOL students in CTA performed similarly to their peers in the comparison condition and seemed to learn little. In contrast, in year 1, Now ESOL students who had CTA made substantial gains, while their counterparts did...
not. Figure 4.2 suggests that there is an effect of school experience, and the value added by the CTA curriculum, especially for Now ESOL students. A cursory examination of the descriptive data for year 1 for other groupings (FARMS and ethnicity) shows similar patterns. The most likely explanation is schools’ experience with the unit—teachers who used CTA for the second consecutive year and who were encouraged to implement with fidelity—seemed to reach the group of students of most concern. These are encouraging results because they indicate that the highly rated curriculum unit can close achievement gaps over time, given more school experience.

Changing the Core of Education Practice

In his 1996 article, Richard Elmore suggests that a central problem of education reform is that although there is no shortage of ideas about how to improve education through research-based interventions, there is evidence of a deep systemic incapacity of U.S. schools and teachers to adopt and extend new ideas about teaching and learning. Elmore allows that some teachers, schools, and school districts are able to successfully change the “core”—fundamental relationships among teachers, students, and knowledge. But what remains elusive in most school settings is the kind of change that includes teachers’ understandings of the nature of knowledge and students’ role in learning. If changes in the core were occurring, the result would likely be different ways of organizing how students work with one another and the teacher, and the character of class work and learning. Moreover, Elmore proposes that the closer an intervention comes to changing the educational core, the less likely it is to be scaled up. Conversely, the farther away the intervention is from the core, the more likely its adoption (Elmore 1996).

In order to discuss this position, we need to clarify two ideas when discussing the interaction of interventions with environment for our collaborative research on scale-up:

1. What exactly is being scaled up and by whom?
2. What would a change in the educational core actually look like, and how could we capture evidence for it?

Although this study is structured as the gradual scale-up of three middle school science curriculum units and measuring their outcomes (especially for diverse students), these units were chosen as exemplars of science curriculum materials that have important instructional attributes. First and foremost, they are standard/benchmark based—designed to teach a specific and challenging interconnected set of ideas; and second, the curriculum materials are structured in ways that allow and encourage a change in the
education core. The latter claim is based on the specified characteristics in the Project 2061 Curriculum Analysis (see appendix) which, if enacted through a curriculum unit, has the potential to change the education core. So for this study, assuming that we can provide evidence (albeit indirect) that the Project 2061 Curriculum Analysis criteria can improve student outcomes, then we can say that our goal is not only to scale up the three curriculum units, but also to scale up the principles that underlie them.

On the other hand, if one asked our collaborators in MCPS what scale-up is, the answer might be somewhat different. Although we have a common initial goal to study the scale-up of three curriculum units, MCPS intercessors are committed to the reform principles that guide change from the state of Maryland, and more importantly from within MCPS. These reform principles include a commitment to constructivist approaches to science teaching and inquiry, increasing alignment of curriculum to MCPS and state frameworks, better use of student assessment information, and increasing achievement for diverse learners. Thus, the goals of the university researchers’ study of scale-up, MCPS’s study of scale-up, and the Project 2061 Curriculum Analysis criteria would seem to have much in common, but emphases vary. The success of the collaboration depends on understanding and respecting different emphases, and appreciating the extent to which they overlap or diverge.

The study has potential for substantial and widespread change in the core of education, given these harmonious dynamics. But how would one know if such changes were occurring? We can document changes in students’ understanding of the target benchmarks/standards through assessments designed to capture complex conceptual change (see Lynch et al. 2003) or changes in students’ reported engagement and goal orientation (Pyke et al. 2003). However, video ethnography may better capture examples of changes in the education core (Cohen, Raudenbush, and Ball 2003). Moreover, unlike many implementation/scale-up studies that focus on teachers and changes in their instruction, this study intentionally aims video analyses at students’ responses to the enacted curriculum, using Project 2061’s Curriculum Analysis criteria as a reference point.

One way of indicating changes to the instructional core is recording incidences of student “trouble” and subsequent means of seeking “remedies.” Trouble is indicated by student queries (“What is this stuff?” “What question are we on?”) or expressions (“Ohmygod!” “Ewwww.”) during an enactment of a unit. Remedies are how the student seeks to fix the trouble, and are categorized by the means for resolution: asking the teacher; discussing with other students; referring back to the curriculum guide; or figuring out an answer for oneself (the latter is often done aloud, and thus can be captured by the video) (Kuipers, Lynch, and Pyke 2003). If the patterns of trouble and remedies change when comparing an intervention unit to what is typical
classroom practice, or over time among students and teachers as an intervention unit is enacted, then this can serve as evidence of change to the education core. Trouble and remedies are only two examples of how we are coding and exploring the video corpus.

Can a curriculum unit create circumstances that encourage the reorganization of the relationship among teacher, students, and knowledge? Consider this exchange, captured on video, among a group of three students at a lab table, prompted by the teacher’s stopping at the lab table and asking about their findings. In the past, the students have used balloons to collect gases given off in chemical reactions. However, in this reaction (steel wool rusts in a flask covered by a balloon, converting the oxygen in the air to ferrous oxide or rust), the balloon is sucked into the flask and students try to account for this surprising result (in lesson 9 of CTA):

Mike: Balloon’s ‘flating inside.
Kim: It’s being sucked in.
Mike: ‘Cause it wants to use up all the oxygen.
Angelique: I think that the, um, reaction took up all the air inside the flask. When the oxygen in the air ran out, it started sucking down the balloon.
Mike: Yes. Sucked in.
Kim: Vacuumed. Uh, deflated.

Satisfied that the students are on track, the teacher moves off to another group. This short dialogue represents, we think, a change in the relationship of students, teacher, and knowledge—the core. The students learn from the curriculum materials, the physical phenomena, and from discussion in the group. The teacher, in this case, does prompt them to explain, then simply offers assurance that they are on the right track.

The discourse analysis of year 0 resulted in 1,800 pages of transcription, capturing video clips such as the one above in which students attempt to explain the phenomenon that they are observing. It must be said that while much of the dialogue is procedural (“What’s the key question?” “I want to light the candle.”) or completely unrelated to the lesson (“Doesn’t my finger look really infected?” “Babies smell, they have that certain smell.”), students primarily rely on the curriculum unit’s text, the phenomena being observed, and one another in the lab portions of the lessons, rather than on the teacher. The teacher’s role is crucial, but changed. She must see to classroom organization, keep the groups on task, and turn the direction of small- and large-group discussions into more productive directions when needed. But she serves as the conductor of a rather large orchestra playing a complex and sometimes improvised piece of music (the enacted curricu-
lum) rather than a soloist who occasionally allows another instrument to repeat her theme.

**SUMMARY**

This chapter discusses the nature of interventions, the environment, and interactions in scale-up research. We view these findings as tentative and would not want to generalize to all kinds of scale-up research. Our research is based upon a university/school district collaboration that has a set of mutual and authentic goals and a hunch that highly rated curriculum materials can make a substantial contribution to student outcomes. The study’s research evidence and how it is reported to teachers and other stakeholders are important features of this work. The types of research evidence that the study provides—disaggregated group data that target outcomes on specific and challenging science standards/benchmarks and video data that show what is occurring in intervention classrooms—are crucial. Providing conceptually specific yet anonymous data to teachers and schools contrasts to current accountability systems that are not well aligned with curriculum and instruction, yet make evaluations public by identifying schools and sometimes teachers. Our approach seems likely to promote a better environment for research on improved outcomes for diverse learners, and healthier interactions among participants.

This accounting of our study leads us to find ever more attractive the notion of distributed learning (Rogoff 2001) recently reviewed by Spillane, Reiser, and Reimer (2002) in their discussion of the implementation of policy research. In our study, learning is intentionally distributed among researchers and science educators in schools, as well as among students and teachers, using curriculum materials as a communicative score (in the musical sense). We would not presume to say that these relatively short (but thoughtfully constructed) curriculum units change the core of education. But by focusing our research primarily on students as actors rather than on teachers or other features of the environment, we explore, both methodologically and with research evidence, how curriculum materials may encourage fundamental shifts in the way that students and teachers work together to construct knowledge.

**APPENDIX**

**AAAS Project 2061 Curriculum Analysis**

The following criteria for evaluating the quality of instructional support can be accessed at www.project2061.org/newsinfo/research/default.htm.
Category I. Providing a Sense of Purpose

A. Conveying unit purpose. Does the material convey an overall sense of purpose and direction that is understandable and motivating to students?
B. Conveying lesson purpose. Does the material convey the purpose of each lesson and its relationship to others?
C. Justifying activity sequence. Does the material involve students in a logical or strategic sequence of activities (versus just a collection of activities)?

Category II. Taking Account of Student Ideas

A. Attending to prerequisite knowledge and skills. Does the material specify prerequisite knowledge/skills that are necessary to the learning of the benchmark(s)?
B. Alerting teacher to commonly held student ideas. Does the material alert teachers to commonly held student ideas (both troublesome and helpful) such as those described in Benchmarks Chapter 15: The Research Base?
C. Assisting teacher in identifying own students’ ideas. Does the material include suggestions for teachers to find out what their students think about familiar phenomena related to a benchmark before the scientific ideas are introduced?
D. Addressing commonly held ideas. Does the material attempt to address commonly held student ideas?

Category III. Engaging Students with Relevant Phenomena

A. Providing variety of phenomena. Does the material provide multiple and varied phenomena to support the benchmark idea(s)?
B. Providing vivid experiences. Does the material include activities that provide firsthand experiences with phenomena when practical or provide students with a vicarious sense of the phenomena when not practical?

Category IV. Developing and Using Scientific Ideas

A. Introducing terms meaningfully. Does the material introduce technical terms only in conjunction with experience with the idea or process and only as needed to facilitate thinking and promote effective communication?
B. Representing ideas effectively. Does the material include accurate and comprehensible representations of scientific ideas?
C. Demonstrating use of knowledge. Does the material demonstrate/model or include suggestions for teachers on how to demonstrate/model skills or the use of knowledge?
D. Providing practice. Does the material provide tasks/questions for students to practice skills or use knowledge in a variety of situations?

Category V. Promoting Student Thinking about Phenomena, Experiences, and Knowledge

A. Encouraging students to explain their ideas. Does the material routinely include suggestions for having each student express, clarify, justify, and represent his/her ideas? Are suggestions made for when and how students will get feedback from peers and the teacher?
B. Guiding student interpretation and reasoning. Does the material include tasks and/or question sequences to guide student interpretation and reasoning about experiences with phenomena and readings?
C. Encouraging students to think about what they’ve learned. Does the material suggest ways to have students check their own progress?

Category VI. Assessing Progress

A. Aligning assessment to goals. Assuming a content match between the curriculum material and this benchmark, are assessment items included that match the same benchmark?
B. Testing for understanding. Does the material include assessment tasks that require application of ideas and avoid allowing students a trivial way out, like using a formula or repeating a memorized term without understanding?
C. Using assessment to inform instruction. Are some assessments embedded in the curriculum along the way, with advice to teachers as to how they might use the results to choose or modify activities?

Category VII. Enhancing the Science Learning Environment

A. Providing teacher content support. Would the material help teachers improve their understanding of science, mathematics, and technology necessary for teaching the material?
B. Encouraging curiosity and questioning. Does the material help teachers to create a classroom environment that welcomes student curiosity, rewards creativity, encourages a spirit of healthy questioning, and avoids dogmatism?
C. Supporting all students. Does the material help teachers to create a classroom community that encourages high expectations for all students, that
enables all students to experience success, and that provides all students a feeling of belonging in the science classroom?

NOTE

1. This work is supported by the National Science Foundation, the U.S. Department of Education, and the National Institutes of Health (REC-0228447). Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the position or policy of endorsement of the funding agencies. Please address correspondence to slynch@gwu.edu.