RUNNING HEAD: FIDELITY OF IMPLEMENTATION IN SCALE-uP

The Evolving Definition, Measurement, and Conceptualization of Fidelity of Implementation in Scale-up of Highly Rated Science Curriculum Units in Diverse Middle Schools

Sharon Lynch
Carol O'Donnell
The George Washington University, Washington, DC

Paper presented at the symposium on "Fidelity of Implementation"
at the Annual Meeting of the American Educational Research Association,
Montreal, Canada.
April 7, 2005

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

The Evolving Definition, Measurement, and Conceptualization of Fidelity of Implementation in Scale-up of Highly Rated Science Curriculum Units in Diverse Middle Schools

"…the curriculum that counts is the curriculum that is enacted. If we want the intended curriculum best to contribute to the enacted one, we must find ways to design the first with the second clearly in view."
(Ball & Cohen, 1996, p. 8)

Introduction

In this paper, we report on our progress in defining, conceptualizing, and measuring fidelity of implementation as we complete the fourth year of a six-year study on the scale-up of highly rated middle school science curriculum units in a large diverse public school system within the metropolitan area of Washington, DC (Lynch, Kuipers, Pyke, & Szesze, in press). To begin, we describe our initial attempts to examine fidelity of implementation within our scale-up study. Given that evaluation studies are prerequisite to scale-up studies, we then discuss the dilemma we faced when studying fidelity of implementation of curriculum materials that had not yet been "proven" effective through evidence-based research designs. In addition, the focus of our research is on improving outcomes for diverse student populations, so it is critical that the effects of the curriculum materials on subgroups of students are determined to guide scale-up.

Curriculum Materials and Diverse Student Populations

There is a compelling need to better understand how curricula, instruction, and student diversity affect student achievement in K-12 classrooms. For the U.S., major barriers to improving science education include the quality of U.S. science curriculum frameworks and standards overall; the quality of curriculum materials used by teachers and students (U.S. National Research Center for TIMSS, 1996); and, teachers' difficulties with implementing reform-based curricula (Lynch, 1997; Lynch, 2000). National systemic reform initiatives point to the need for more focused science curricula and better curriculum materials for teachers to use (Lynch, 2000). Therefore, it is reasonable to assume that improved curriculum materials aligned with science education standards that encourage students to learn with understanding could accelerate the reform process (given qualified teachers and reasonable resources).

There are several national organizations that are involved in efforts to stimulate the creation of improved curriculum materials or to develop rating systems to identify promising, extant standards-based curricula (Lynch et al., in press). The AAAS Project 2061 has launched a major effort to find curriculum materials aligned with benchmarks that meet a rigorous set of criteria consistent with current theories of learning and content specific instructional strategies that support learning (Kesidou & Roseman, 2002). The Project 2061 Curriculum Analysis consists of seven categories of criteria for analyzing curriculum materials (see AAAS, 2003). This analysis inquires if a curriculum unit: starts from ideas that are familiar or interesting to children; explicitly conveys a sense of purpose; takes into account student ideas, and conveys suggestions for teachers to find out what their students think about the phenomena related to the benchmark; provides for first-hand experiences with phenomena; and, has students represent their own ideas about phenomena and practice using the acquired knowledge and skills in varied contexts (Roseman, Kesidou, & Stern, 1996; Kesidou & Roseman, 2002). These criteria seem to be at the heart of effective instruction for diverse learners. They are so comprehensive and sound that if curriculum materials met them, it seems likely that all students would be well on the way to learning targeted benchmarks (Lynch, 2000).

However, there are three issues to consider. First, any curriculum analysis conducted by expert panels divorced from classroom settings runs the risk of falling short of clarifying "what really works", especially for diverse learners. Still the Project 2061 Curriculum Analysis seems to hold a great deal of promise as a first step in identifying effective curriculum materials for diverse populations. Curriculum materials that pass the first screen—being reasonably well rated using the Project 2061 criteria—seem likely candidates for scaling-up. Moreover, the meticulously wrought rating system in itself holds distant promise for scale-up, because if it can be determined that curriculum materials with these characteristics or a subset of these characteristics induce better instruction and learning for diverse students, then they might be used as design principles for creating new curriculum materials with similar characteristics (Lynch, 1997). Project 2061 is currently involved in collaborative efforts with curriculum developers and science education researchers to do exactly this--develop materials using the curriculum analysis criteria as design principles (Jo Ellen Roseman, personal communication).

The second issue is associated with using the Project 2061 Curriculum Analysis to identify science curriculum materials for use in schools. Project 2061 has not been able to identify hardly any science curriculum materials that received a high rating. (Math materials faired better.) But even those science materials that have been highly rated using this system are likely not to have been evaluated using evidence-based research, which requires a comparison of their efficacy with some reasonable control group. Should scale-up occur without results from sound evaluation studies?

Finally, there is a contrast between the need to provide a single set of curriculum materials for all students within a school setting and the recommendations to modify curriculum and instruction to meet the needs of diverse learners. This contrast creates a quandary for scale-up research. The constraints of schooling often require that a single set of curriculum materials suffice for thousands of students and presumably help each student obtain scientifically appropriate understandings of science concepts. But theories of conceptual change and understanding the social-historical contexts of schooling suggest that schools must be flexible and responsive to increasingly diverse student populations (Lynch et al., in press). In the next section we describe how our study of Scaling Up Highly Rated Science Curriculum Materials for Diverse Populations has navigated these issues, and their connection to *fidelity of implementation*—defined by the public health field as the determination of how well an innovation is being implemented in comparison with the original program design (Mihalic, 2002).

Description of SCALE-uP

This research program, called Scaling up Curriculum for Achievement, Learning, and Equity Project (SCALE-uP), is a collaborative effort between The George Washington University (GWU) and Montgomery County Public Schools (MCPS)—a large suburban school district outside of Washington, DC that is highly diverse (with no ethnic majority as of 2001). SCALE-uP is in its fourth year of a six-year program. The study is designed to understand how highly rated science curriculum materials, aligned with reform goals improve educational outcomes for diverse student populations on a large scale (Lynch et al., in press). This work aims *to improve education achievement by providing scientifically-based knowledge and skills that lead to sustainable learning changes across diverse student populations* through interdisciplinary research that informs practice and can be implemented in real, complex, and varied educational environments (National Science Foundation, 2000).

SCALE-uP is designed to implement and scale-up three different middle school science curriculum units over a six-year period. The three units are: an eighth grade curriculum unit, *Chemistry That Applies* (CTA) (State of Michigan, 1993); a seventh grade unit, *Great Explorations in Math and Science (GEMS): The Real Reasons for Seasons* (GEMS, Lawrence Hall of Science, 2000); and, a sixth grade unit, *ARIES: Exploring Motion and Forces* (Harvard-Smithsonian Center for Astrophysics, 2001). We have conceived of the research as having two major components: implementations studies for each unit; and, scale-up studies for each unit. The goal of the implementation studies is to evaluate the efficacy of a unit in the context of MCPS using a quasi-experimental design (see Lynch, Pyke, Kuipers & Szesze, in press, for a complete description). If a unit proves to be effective, then it can be scaled up from five middle schools, to 20 middle schools, then eventually to all 37 MCPS middle schools. The implementation study research question is:

> 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?

Affirmative answers to this question guide the decision to scale up a unit.

The schedule for the implementation and scale-up studies is:

Grant Year	Description of Scale-up Efforts
Year 0 2001-02	5 matched pairs of Treatment and Comparison schools with Chemistry That Applies (CTA) ~ 3000 8th graders.
Year 1 2002-03	Replicate CTA with same schools ~ 3000 8th graders. CTA in 20 schools altogether ~ 6000 8th graders.
Year 2 2003-04	The Real Reasons for Seasons (Seasons) in a matched sample of 5 schools ~ 3000 7th graders. Motion and Forces (M&F) in a matched sample of 5 schools ~ 3000 6th graders.
Year 3 2004-05	CTA in 37 schools (~ 12,000 8th graders). Replicate study for Seasons and M&F (~ 3000 7th and ~ 3000 6th graders).

In 2005-06, we are slated to study a research question targeted at scale-up:

> *Fidelity of Implementation*: Does the degree to which a curriculum unit is taught with fidelity affect student outcomes?

We anticipated using Years 1-3 to develop an instrument to measure fidelity of implementation. However, as will be apparent in the section that follows, matters pertaining to defining, conceptualizing, and measuring fidelity of implementation were unavoidable as early as Year 0. Although we began SCALE-uP with a limited and somewhat naïve notion of fidelity of implementation, a review of the literature conducted by O'Donnell (O'Donnell, 2004), eventually resulted in the following framework for discussing fidelity of implementation (adapted by O'Donnell from Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003). The framework has five components:

1. Adherence to the unit – Whether the unit is delivered as designed or written. The implementation of particular activities, investigations, and methods is consistent with the way the unit is written and with professional development.
2. Exposure – Whether the number of lessons implemented, the length of time spent on the unit, and the type of concepts and skills received by the students is consistent with the intent of the unit developer.
3. Program differentiation – Whether critical features that distinguish the unit from the standard or traditional curriculum are present or absent from unit implementation.
4. Quality of delivery – How a teacher implements a unit. The theoretical or pedagogical ideals evident during implementation are consistent with the way the unit is written. The skill, enthusiasm, preparedness, and attitude in using the techniques or methods prescribed by the unit are evident in the implementation.
5. Participant responsiveness – The extent of student engagement. Their involvement in the activities and content of the program is consistent with the intent of the unit developer.

Evolution of Fidelity of Implementation in Our Study of Implementation and Scale-up: Years 0 - 3

Throughout our scale-up study, our definitions, conceptualizations, and measures of fidelity of implementation evolved. In the next section, we trace the year-by-year evolution, for each curriculum unit studied.

Year 0 (2001-02): Planning Grant

Evolving Notion of Fidelity of Implementation

In 2001-02, which we call Year 0, we received a relatively small planning grant from the Interagency Educational Research Initiative (IERI) to explore the possibilities of scaling-up CTA. CTA had a high rating from Project 2061, had been used with literally thousands of students, and had its provenance with a Michigan State University research group that had done some quasi-experimental evaluation on a companion unit (*Matter and Molecules*) that showed it to be effective. Therefore, we had many reasons to believe that CTA would be more successful than the standard range of options to which it was to be compared in MCPS. Through the planning grant,

CTA was implemented in five highly diverse middle schools (~1500 students) using pre-, post-, and delayed post-tests. CTA was found to be more effective than the standard range of options in five matched comparison schools. Disaggregated data showed that, with almost every subgroup tested, students who used CTA performed significantly better than their comparison peers, with effect sizes in the moderate range (see Figure 1). These encouraging results gave us reason to apply for and receive a 5-year scale-up grant from IERI to study the scale-up of CTA and two other middle school curriculum units in MCPS.

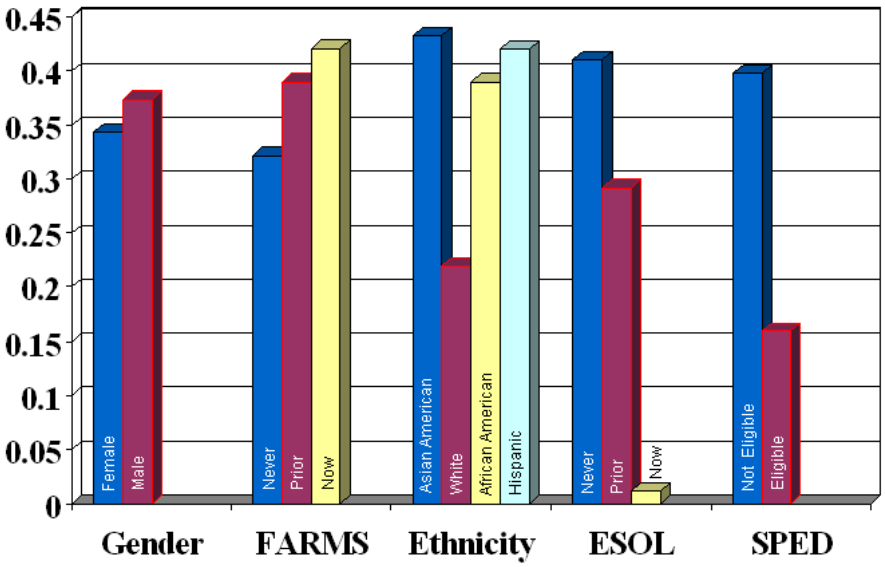

Figure 1: Effect size CTA Year 0 (2001 - 2002).

As we implemented CTA in Year 0, we were aware that fidelity of implementation was an issue, but we had limited funding to measure it. We were not very sophisticated in our conception of FOI, but our goal was to at least pay one visit to each CTA classroom to make sure the unit was being implemented. Thus, our crude measure for fidelity of implementation was a binary measure of *adherence*—that is, either CTA was being implemented or it was not. (All but one teacher did implement the unit, but at least two failed to complete it.) Classroom visits by GWU researchers showed a great deal of variation in implementation by teachers: Implementation ranged from the non-implementer, to struggling implementers who had little experience with chemistry laboratory work at the middle school level, to competent implementers, to over-the-top implementers who were not only implementing the unit but making profligate additions to the unit that seemed unlikely to preserve its integrity of purpose and design.

At this point in our study, modifications to the unit were expected. The equity literature suggested that diverse learners would require modifications to science curriculum and instruction, and we had intentionally selected for our sample the most diverse schools in MCPS. We wanted to capture teachers' modifications and set up a web-based interactive communication system that teachers could log on to and chat about modifications that they were making daily to CTA lessons. However, only two teachers out of 15 used the system, and they did not discuss modifications. Moreover, at professional development meetings, implementing teachers proved to be not so interested in modifications either. Rather, they wanted to talk more about the difficulties in implementing the 18 chemistry labs required by CTA. Simply getting the labs to "work" and understanding the chemistry behind them were the most important issues raised. In addition, teachers reported high levels of student engagement, so the need to find problems in the unit, which required modifications, faded.

Summary of Fidelity of Implementation Year 0

Year 0 taught us the first important lessons about fidelity of implementation for a challenging unit like CTA in this implementation stage:

- **Definition of Fidelity:** Modifications (improvements) were not on teachers' minds at this stage, nor were researchers thinking about ramifications for fidelity of implementation that would soon arise. Rather, the main concerns for teachers included gaining comfort with the lab requirements and equipment outlined by CTA's Teachers' Guide and keeping the lessons "moving." The main concerns for researchers were to gather and analyze the data on the Year 0 implementation study.
- **Measurement:** At this point, it was important to simply know whether teachers were actually implementing the unit and completing it or not. Visiting the teachers in the classroom proved to be a crude way of measuring *adherence* (one measure of fidelity).
- **Hedges' Conceptualization of Scale-up:** Because this was an implementation study, conceptualization for scale-up was not applicable. However, the assumption operative was Hedges (2004) "tailored manufacturing model" (see Lynch & O'Donnell, 2005).

Year 1 (2002-03): The First Year of SCALE-uP

Evolving Notion of Fidelity of Implementation

During the 2002-03 school year, we began our five-year SCALE-uP research project. The goal for Year 1 was to replicate Year 0 of our study with CTA in the same five pairs of schools. The teachers were more familiar with the unit and they knew that SCALE-uP was going to be a part of their science program for the next five years. There were funds built into the grant for observations in science classrooms by the observers from the MCPS Office of Shared Accountability (OSA). Having a MCPS observer in the classroom was a practice that middle school science teachers were accustomed to. This meant MCPS observers, rather than GWU researchers, were going to be conducting observations (although GWU researchers continued to visit classrooms). The observations would provide a crude measure of *adherence* to the unit.

At this point we made an important decision about the replication of CTA. We decided that it was far better to ask teachers to implement the unit with fidelity, than ask teachers to make modifications to the unit. This meant teaching all 18 lessons in order, without adding accessory materials. This would allow a better evaluation of the efficacy of CTA, and would take pressure off the teachers to invent modifications to a unit that was quite challenging to implement. Teachers seemed to understand the rationale for this important adjustment to the research rationale, and raised no major objections.

Given this request that CTA be taught with "fidelity," one goal for Year 1 was to develop a valid and reliable observation instrument that could be used by MCPS observers to capture *quality of delivery*. Because the unit was highly rated by Project 2061, the rating scheme for the 22 criteria in the Project 2061 curriculum analysis was used as the basis for the design of the Classroom Observation Instrument to capture *quality of delivery*. We had many discussions about the fact that the Project 2061 criteria were designed for written curriculum materials, rather for classroom observations. However, we reasoned that if the written curriculum materials possessed instructional attributes that affected the way teachers taught and the way students learned, then these instructional moves might be observed in the classroom if it were taught with fidelity. Susanne Merchlinsky of MCPS Office of Shared Accountability worked with the authors of this paper to create the Classroom Observation Instrument, and this joint enterprise is an important aspect of our collaborative research.

In addition to measuring the teachers' *quality of delivery*, we sought to measure the students' response to the curriculum unit. The quality of curriculum implementation seems to rest not only with the teacher, but also with the students (*participant responsiveness*) as the unit was enacted. Consequently, we built into the Classroom Observation Instrument prompts designed to capture student moves, as well as teacher moves.

During Year 1, MCPS observers conducted two rounds of observations, visiting a total of 16 classrooms implementing the CTA curriculum unit. All of the teachers visited were implementing CTA and adhering to the unit. *Quality of delivery* seemed to vary, according to the analysis of the data collected with the Classroom Observation Instrument. Although we were not satisfied with the reliability of the instrument, it seemed we were making progress in understanding and measuring fidelity of implementation for CTA.

The results of the implementation study for CTA in Year 1 (with new emphasis on fidelity) mirrored those of Year 0. CTA once again proved to be more effective than the comparison condition overall (see Figure 2); CTA had greater effects on students than did the comparison curriculum for students in all subgroups. Consequently, CTA was ready for scale-up to 15 additional classrooms the following year.

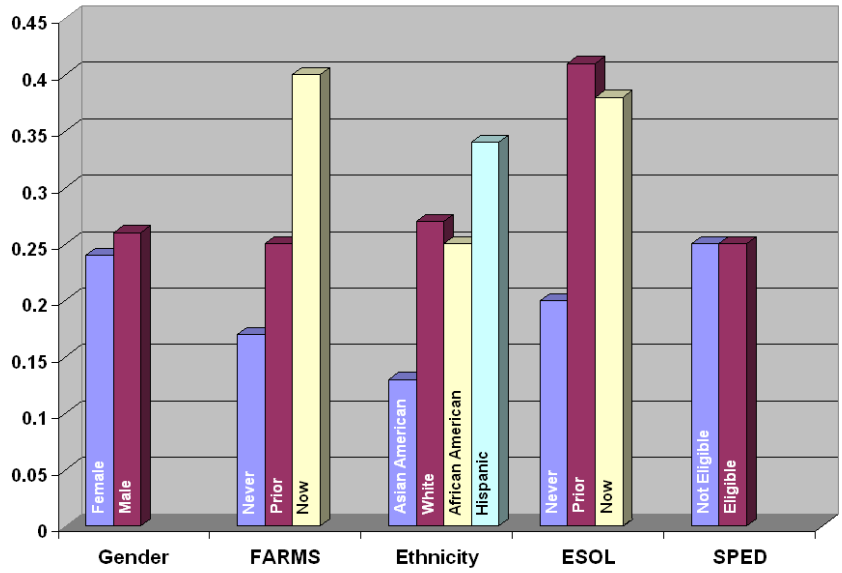

Figure 2: Effect size of CTA Year 1 (2002-2003).

Summary of Fidelity of Implementation for Year 1

Year 1 taught the second set of important lessons about fidelity of implementation:

- **Definition of Fidelity**: Fidelity was defined as *adherence* to the unit, measured crudely by classroom observations; *quality of delivery* (as measured by implementation of the 2061 instructional strategies); and *participant responsiveness* (students' participation in the activities of the unit).
- **Measurement:** Visits using a Classroom Observation Instrument focusing on teachers' implementation and participants' responsiveness (see Figure 3 for a portion of that instrument, which was later revised due to problems with reliability and validity).
- **Hedges' Conceptualization of Scale-up:** This is not exactly applicable because this was an implementation study. However, on the basis of Hedges' (2004) conception of scale-up and fidelity of implementation (see Lynch & O'Donnell, 2005), Year 1, the study fell under "mass production" model in Year 1, rather than the "tailored manufacturing model" followed in Year 0, the year in which modifications were encouraged and allowed.

Year 2 (2003-04): Scaling-up CTA and Implementation Studies of Seasons *and* Motion and Forces

Evolving Notion of Fidelity of Implementation

As we went into Year 2 (2003-2004), much to our surprise, fidelity of implementation issues began to take on a larger and larger role, much of this initiated by the teachers.

Scale up of CTA. The data for two consecutive years of CTA implementation studies showed that the unit was more effective than the standard range of curricular options. Most encouraging, the largest effects sizes were for demographic subgroups of students that were often placed most at risk in science classrooms. Consequently, we were comfortable with the decision to scale-up CTA with fidelity rather than to encourage modifications to the unit. This meant we were using the mass production metaphor for scale-up. While this engineering metaphor might not seem as attractive as others, it nonetheless fits the data we were obtaining and the context of the school system in which we were operating. MCPS has a well-organized, effective K-12 science department and network of science teachers, had been the recipient of many grants to improve curriculum and instruction, and has an effective

professional development and management program in place. Communication occurs via the Resource Teachers (department chairs) and other professional networks of middle school science teachers. Thus, this mass production model seemed optimal, given this context and the results of the implementation studies.

We continued to monitor *adherence* to CTA (in the gross sense—that is, were teachers implementing CTA?) by sending MCPS observers into classrooms. Moreover, we made certain to include some of the self-contained special education classrooms in the observations. These observations convinced us that although some of the special education teachers were "using" CTA, the modifications that they necessarily had to make for cognitively challenged students resulted in low fidelity of implementation. Thus, it was hard to say these students were "receiving" CTA as it was intended. We eventually decided to exclude these classrooms from the study because we could not be sure that all special education teachers were attending the CTA-related professional development or implementing the unit. This occurrence aligns with Hedges' observation that the path of scale-up in the mass production model might be changed through measures of fidelity, such as observations (Hedges, 2004).

On the basis of an analysis of the data from the Classroom Observation Instrument collected in Year 1, we were concerned about its discriminatory power. To test the hunch that the instrument was not discriminating between Treatment and Comparison instruction, we conducted pre-CTA observations on some of the teachers slated to implement CTA (see Figure 3 for a section of our revised classroom observation protocol with data from our pre- and during-unit observations). Our reasoning was that if the measures from the pre-CTA observations were about the same as the measures taken during CTA for the same teacher, then we were not capturing the *quality of delivery*. Thus, our instrument would need more work to improve validity and observer reliability.

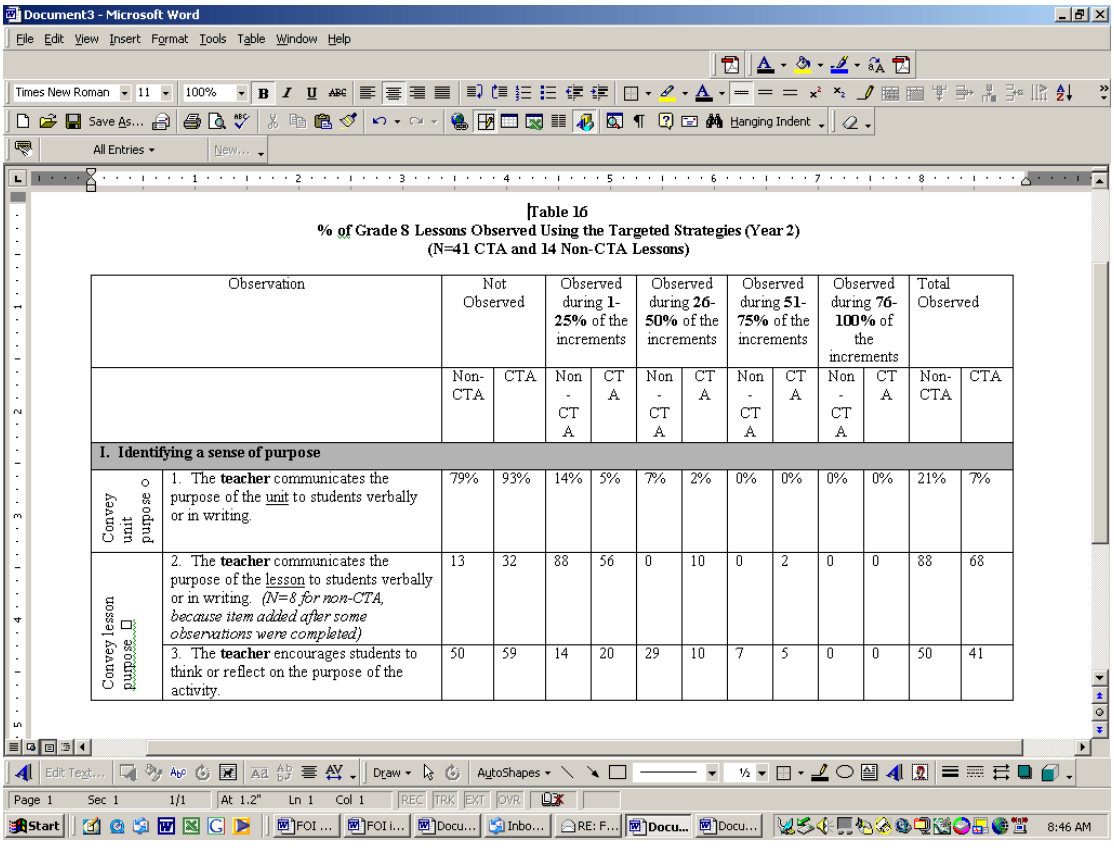

Table 16
% of Grade 8 Lessons Observed Using the Targeted Strategies (Year 2)
(N=41 CTA and 14 Non-CTA Lessons)

Observation		Not Observed		Observed during 1-25% of the increments		Observed during 26-50% of the increments		Observed during 51-75% of the increments		Observed during 76-100% of the increments		Total Observed	
		Non-CTA	CTA	Non-CTA	CTA	Non-CTA	CTA	Non-CTA	CTA	Non-CTA	CTA	Non-CTA	CTA
I. Identifying a sense of purpose													
Convey unit purpose o	1. The **teacher** communicates the purpose of the unit to students verbally or in writing.	79%	93%	14%	5%	7%	2%	0%	0%	0%	0%	21%	7%
Convey lesson purpose □	2. The **teacher** communicates the purpose of the lesson to students verbally or in writing. *(N=8 for non-CTA, because item added after some observations were completed)*	13	32	88	56	0	10	0	2	0	0	88	68
	3. The **teacher** encourages students to think or reflect on the purpose of the activity.	50	59	14	20	29	10	7	5	0	0	50	41

Figure 3: Revised quality of delivery instrument showing data from pre- and during-unit observations

Implementation of Seasons *and* Motion and Forces. Five additional matched pairs of middle schools were selected for implementation studies of *Motion and Forces* (a 6th grade unit) or *Seasons* (a 7th grade unit). In this case, we did not oversample from the most diverse schools, but rather chose a stratified random sample such that the population of middle schools in the entire county was represented. The five pairs of schools selected might also be

considered as test-beds, in which to better test the efficacy of CTA under identifiably varied circumstances. MCPS observers visited 28 *Seasons* classrooms and 34 *Motion and Forces* classrooms during Year 2. As a result of this data collection, we were confident that *Motion and Forces* and *Seasons* were being implemented in MCPS classrooms (our crude measure of *adherence*). However, the observers agreed that the Classroom Observation Instrument was not capturing *participant responsiveness.* It was impossible to observe enough student conversations to make anything other than extremely high inference judgments about how students were experiencing the units. We agreed that for Year 3, we would drop items that focused on the student participation and create a separate instrument for measuring *participant responsiveness*.

Q*uality of delivery* items also needed refinement; it proved difficult for observers to measure the frequency of the 22 indicators reliably. We knew that we would have to revise the Classroom Observation Instrument again in Year 3.

Teachers Request Fidelity Guidelines

Teachers of the two new units participated in a two-day professional development workshop conducted by the developers of the units and follow-up meetings during the school year. They began to ask about the classroom observations, and requested a definition of "fidelity of implementation." Teachers, in earnest, wanted to know the extent to which they could modify the unit. For instance, if they had a favorite video that they usually showed on astronomy, was it permissible to show this as the *Seasons* unit was progressing? What about supplemental textbook readings? Could they use laboratory exercises not included in the unit? Could they develop tests? Assign homework? Give lesson starter "warm-ups"?

These questions resulted in a serious dialogue between researchers, developers, and MCPS science educators and teachers who had taken leadership roles in the project about what constituted fidelity of implementation. The GWU Project Director/Senior Research Associate, Carol O'Donnell, had done considerable research on fidelity of implementation, and our grasp of the concept was improving. SCALE-uP researchers and teachers together were able to reach a decision about fidelity of implementation—that is, if this was to be a valid study of the efficacy of the two new units, then the units needed to be implemented with a good degree of fidelity. Teachers' main concerns were over *adherence* criteria, rather than *quality of delivery*. They wanted to know what they could or could not include, rather than focusing on how best to deliver the unit according to the instructional prompts provided in the Teachers' Guides.

Consequently, a set of guidelines were drawn up and were given to implementing teachers (O'Donnell, Lynch, & Hansen-Grafton, 2004). The guidelines seemed well accepted because the teachers were beginning to be drawn into the logic of the implementation studies. While a few teachers objected to the loss of autonomy, the guidelines did recommend that if a teacher thought students would suffer from the fidelity guidelines, that the final decisions about modifications should be based upon student needs, not research design integrity. The guidelines are printed below in bold-face:

Fidelity is:

- **Adhering to unit and lesson purpose, goals, and objectives**
- **Adhering to unit pedagogical approaches (incorporating the learning cycle or 5 E's if present in the unit; introducing the lesson with a lesson-related warm-up followed by hands-on or minds-on inquiry; utilizing lesson closure; addressing misconceptions; etc.)**
- **Following lesson sequence (i.e. teaching all of the unit lessons and teaching them in order)**
- **Using the recommended equipment or materials (i.e. don't substitute equipment unless it is broken or has proven to not work)**
- **Making an adaptation to the lesson that does NOT change the nature of the lesson's intent (e.g. listing the lesson questions on a worksheet and adding lines for students to write on; using a thicker cardboard for a disk launcher; having students discuss questions in a group before reporting them out to the class; addressing prerequisite skills that your students may not have prior to a lesson, such as taking time to have students practice using a stopwatch).**

Fidelity is not:

- **Reducing or modifying unit goals and objectives**
- **Reconfiguring the lesson so that your standard instructional repertoire gradually replaces parts of the new unit as originally designed**
- **Reducing the amount of behavioral change expected from participants (e.g. not having students write or address their individual and current ideas about a topic prior to starting the hands-on portion of the lesson)**
- **Varying grouping strategies outlined in the unit - that is, conducting the lesson as a whole class and demonstrating the investigation instead of allowing each group to perform the investigation firsthand**
- **Changing the unit's organizational patterns (i.e. rearranging order of lessons); varying the lesson schedule (i.e. stopping the unit before it is over); reducing the number of lessons; eliminating or skipping lessons**
- **Substituting other curriculum materials or lessons for those described by the unit; adding reading selections or video tapes not agreed upon by the study**

This set of guidelines is especially noteworthy because the definitions and ranges were demanded by the teacher participants in the study. Both teachers and researchers discussed what might or might not be acceptable in the range. Most notable is that teachers seldom (if never) inquired if they were still teaching with fidelity if they neglected to teach with aspects of the Project 2061 criteria in place, i.e., quizzing students for their misconceptions, or curtailing students' opportunity to reason from evidence, and so on. In short, our first definition of fidelity that emanated from the Project 2061 criteria was usurped by teachers' practical or instrumental definitions of fidelity. Despite professional development that often emphasized the Project 2061 criteria and their importance to the unit, the criteria seemed far less salient than instrumental aspects of the curriculum delivery. Additionally, we believed that these guidelines, developed *a priori*, would guide the development of *structural* fidelity of implementation criteria related to *adherence, exposure,* and *program differentiation* during Year 4.

Summary of Fidelity of Implementation for Year 2

Year 2 helped us to continue to refine our definition, measurement, and conceptualization of fidelity of implementation as follows:

- **Definition of Fidelity**: Fidelity continued to be defined as *adherence* to the unit, measured crudely by classroom observations; *quality of delivery* (indicated by implementation of the 2061 instructional strategies inherent in the curriculum materials), and *participant responsiveness* (students' participation in the activities of the unit). But we now realized that the latter two criteria needed to be separated in our definition on the Classroom Observation Instrument.
- **Measurement:** Classroom visits and the Classroom Observation Instruments focused both on *quality of delivery* and *participant responsiveness.* However, due to validity concerns, we realized that these two ideas would have to be measured separately. In addition, we began to document, through interviews, the length of time (*exposure*) teachers were spending on teaching each unit.
- **Hedges' conception of scale-up and fidelity of implementation:** CTA was scaling up to 20 schools, via something close to "mass production model" because the unit had proved to be effective for two prior consecutive years in MCPS. (MCPS might also be considered an individual test-bed for CTA, because we do not think that the MCPS context is generalizable to all school districts across the U.S.) For *Motion and Forces* and *Seasons*, although these were first year implementation studies, it was decided that teachers should adhere to the unit as closely as possible, and modifications were discouraged.

Year 3 (2004-05): Scaling up CTA and Replication of Implementation Studies of Seasons *and* Motion and Forces

Evolving Notion of Fidelity of Implementation

Due to difficulties with inter-rater reliability during Years 1 and 2, the *quality of delivery* protocol for the Classroom Observation Instrument was simplified dramatically during Year 3 (see Figure 4 for a portion of the revised protocol). Only indicators that reflected instructional strategies evidenced by the teacher's actions were now included and student-related indicators were deleted (since a new, separate instrument for *participant responsiveness*

was developed). Frequency counts for each indicator were replaced by a simpler rating system—Not Evident, Evident, or Evident with Emphasis. A rating of "Evident" means that the indicator, as demonstrated by the teacher, is observed in the classroom. Examples from actual classroom observations from Years 1-2 are included in the protocol to serve as a way to operationalize "Evident." "Evident with Emphasis" is checked when there is an indication of either increased frequency or quality in the indicator. If an indicator cannot be rated or has not been observed, the observer assigns a value of "Not Evident". Unlike Years 1 and 2, in which only a classroom period was observed, the observations in Year 3 include a full lesson, even if the lesson runs to one, two, three or more class periods.

Classroom Observation Instrument: Quality of Delivery				
Rating categories and indicators		**Not Evident**	**Evident**	**Evident with Emphasis**
IV. Developing and Using Scientific Ideas				
a. **The teacher uses technical terms specific to the target ideas only in the context of the activities** (*e.g. "Yesterday you did....scientists call what you observed...." Note: Giving vocabulary lists separately from activities, copying words and definitions off the board, etc. are NOT considered in the context of the activities.)*				
b**. The teacher encourages students to define and/or use technical terms (either written or orally), based on the context of the activity** *(Define: e.g. "The activities we just did showed.....How might you define....based on your observations in the activity?") (Use: e.g. "How might you use the term....to describe how the object....?")*				
d. **The teacher gives students opportunities to examine, critique, interpret, create, and/or use accurate representations of the content or material being taught to the students (diagrams, drawings, graphs, images, etc.)** *(Note: Can be done as homework)*				
e. **The teacher models, and/or demonstrates the skills or the use of knowledge (or asks a student to model). This can include the teacher verbally describing procedural instructions prior to the lesson**. (*The teacher is verbalizing procedural instructions NOT showing the thinking or reasoning behind them, e.g. "First you will have to tilt your styrofoam ball/Earth so that the red dot is facing the North, then everyone begin to slowly walk around the light/Sun in a circle.")*				

Figure 4: Portion of revised Classroom Observation Instrument Year 3: 2004 – 2005

Scaling up CTA. During Year 3, we scaled up CTA to all 37 MCPS middle schools, encouraging fidelity of implementation with no modifications. We could no longer arrange to observe CTA classrooms because there were too many. However, all the 8th grade teachers were asked to administer and grade a pre-and posttest for CTA, using the Conservation Of Matter Assessment developed via the SCALE-uP grant. The teachers were to report their results to the MCPS project director. It was evident that CTA had become institutionalized.

Interestingly, a small group of teachers thought that they might improve CTA with a specific modification, and have been in touch with the developer. They have asked to conduct a well-designed quasi-experimental pilot study of the modifications in Year 4 (next year at this writing). This represents a substantial change in science teachers' views of implementing curriculum materials. Four years ago, prior to SCALE-uP, "pilot study" meant having few teachers try out a new curriculum unit and voicing their opinions. Due to lack of consensus, not a single unit had ever scaled-up using that approach. Now, many teachers seem to understand the desirability of a good research design.

Replication of Implementation Studies of Seasons *and* Motion and Forces

At the start of Year 3, our understanding of how to measure fidelity of implementation was influenced by the careful reading and discussion of an article by Mowbray, Holter, Teague, and Bybee (2003). Mowbray and colleagues discuss the various ways the concept of fidelity of implementation can be used, as interventions are created and then tested. The article placed fidelity of implementation in a wider perspective than the education literature often does. Many of the examples came from discussion of fidelity of implementation in mental health interventions.

We reexamined the five general components of fidelity of implementation formulated by Dane and Schneider (1998) and summarized by Dusenbury and colleagues (2003), and began to develop fidelity criteria aligned with each component. We developed definitions for each of the criteria adapted to suit the needs of our research, created instruments corresponding to the criteria to measure each component quantitatively, and implemented a plan for checking the validity and reliability of each instrument.

In addition, we began to think about how we might distinguish the critical features in the Treatment unit from the critical features in the Comparison units and measure their presence or absence during implementation of the target ideas. To do this, we developed an instrument to measure the "flow of a lesson" or the amount of time in *one* lesson (over several class periods if needed) that is taken up by different types of activities, either teacher-centered or learner-centered. Our thinking was if the unit's lessons form scripts, then these scripts contain scenes should differ from the scripts and scenes that ordinarily occur in classrooms. This organization of time and activities could lead students to different activity systems for the highly rated units, activity systems that lead the students to do more of the kinds of thinking that results in a better understanding of the target concepts. Such an analysis would help us to better understand *program differentiation* (see Mobray et al., 2003 and Lynch & O'Donnell, 2005).

Summary of Fidelity of Implementation for Year 3

Year 3 allowed us to refine our definitions, measurements, and conceptualizations of fidelity of implementation that emerged during Years 0 – 2 as follows:

For CTA.

- **Definition of Fidelity for CTA**: *Adherence* (crude).
- **Measurement for CTA:** Teachers required to pre- and post-assess and score students with conservation of matter assessment and send results to MCPS project director.
- **Hedges' conception of scale-up and fidelity of implementation:** CTA was scaling up to 37 schools with the "mass production model" adopted.

For Seasons *and* Motion and Forces

- **Definition of Fidelity**: Definition now included *quality of delivery*, *participant responsiveness*, *exposure*, *program differentiation*, and crude measure of *adherence,* but one that was now based on the fidelity of implementation guidelines requested by teachers.
- **Measurement of FOI for *Motion and Forces* and *Seasons*:**
  --*Quality of delivery* Classroom Observation Instrument revised a third time with attention to its ability to discriminate (see Figure 4).
  --*Participant responsiveness* questionnaire that directly asks students how they are experiencing the unit's instructional features.
  --Lesson flow observation tool to describe a temporal aspect of *program differentiation.*
  --Individual teacher interviews to measure *exposure*.
- **Hedges' conception of SCALE-uP and fidelity of implementation:** *Seasons* and *Motion and Forces* are still in the implementation study phase but teachers are encouraged to implement with fidelity. This is important in order to understand whether or not these two units ought to be scaled up—less is known about these units than we knew about CTA, and we want to ensure we capture their efficacy before scale-up.

Closing Remarks

In summary, we have outlined the evolution of our definitions and measures of fidelity of implementation as we complete the fourth year of a six-year scale-up study, using five fidelity of implementation components adapted from the public health and mental health fields—*adherence, exposure, quality of delivery*, *participant responsiveness,* and *program differentiation*. These components serve as the foundation of developing fidelity criteria and are bound together by three categories adapted from Mobray et al. (2003): structure, process, and self-perceived effects by students/participants. Building on this work, this paper proposes a role for fidelity of implementation in research design for scale-up; provides information about how each of the five components can be used to define and measure fidelity of implementation; discusses in detail the evolution of fidelity of implementation instruments designed to capture each component, in particular *quality of delivery;* and reports on the conceptualization of and challenges to measuring fidelity as it emerged from the SCALE-uP study. Each of these issues must be considered as fidelity of implementation is examined at varied sites for diverse learners.

In addition, using two scale-up models derived from examples outside the field of education—tailored manufacturing and mass production—this paper describes the changing role of fidelity of implementation in scale-up research and suggests its importance in research using comparative designs. It is our view that curriculum developers have a responsibility for defining parameters of fidelity as they create and field test new curriculum materials. Moreover, science education researchers studying curricular interventions must consider how fidelity of implementation figures into research designs and document the evolution of contextual variables that affect such definitions and measurement. In our case, moving from tailored manufacturing to mass production seemed necessary during our implementation studies, because we were implementing units that had not yet been "proven" effective through scientifically-based research.

While it may appear that asking teachers to implement a new unit with fidelity restricts teachers' autonomy, teachers still had a great deal of latitude in how they delivered the unit, because the fidelity guidelines focused on *adherence* and *exposure*, rather than *quality of delivery*. Teachers may have been less likely to buy into this interpretation of fidelity of implementation if the units were longer than several weeks. It also seems likely that teachers may not have been interested in teaching with fidelity had they not understood the design of this large research study. In addition, as science educators, in particular, they are more likely to appreciate the need to control variables when testing for effectiveness; and, as MCPS teachers, they are also experienced with other grants and recognize that unless there is evidence of effectiveness, change is unlikely to occur.

Finally, although we have not "proven" that diverse students would not perform even better with modifications to a given curriculum unit, the implementation study of CTA in Year 1 shows that it was more effective than the comparison condition and establishes a sort of baseline of good curricular and instructional practice from which modifications can be made. Thus far, modifications suggested by MCPS teachers have focused on changes in structural elements of the unit rather than process changes likely to positively affect subgroups of diverse learners. As Lynch has argued before (Lynch, 2000), before it is assumed that diverse students need something "additional", we should be sure that they have access to high quality curriculum and instruction as a matter of course. In this case, fidelity of implementation guidelines help establish these baselines and provide a way to measure the effectiveness of the units before they are scaled-up.

References

AAAS. (2003). *Project 2061 middle grades science textbooks evaluation. Criteria for evaluating the quality of instructional support.* Retrieved July 25, 2003, from http://www.project2061.org/research/textbook/mgsci/criteria.htm.

Ball, D. L., & Cohen, D. K. (December, 1996). Reform by the book: What is: Or might be: The role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6 – 8 +14.

Dane, A.V. & Schneider, B.H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23-45.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research Theory and Practice, 18*(2), 237-256.

Great Explorations in Math and Science Program (GEMS). (2000). *The Real Reasons for Seasons: Sun-Earth Connections.* Berkeley, CA: Lawrence Hall of Science, University of California at Berkeley.

Harvard-Smithsonian Center for Astrophysics. (2001). *Exploring Motion and Forces: Speed, Acceleration and Friction.* Watertown, MA: Charlesbridge Publishing.

Hedges, L. (2004, April). *Designing studies for evidence-based scale up in education.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, April, 2004.

Kesidou, S. & Roseman, J.E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. Journal of Research in Science Teaching, 39, 522-549.

Lynch, S. (1997). Novice teachers' encounters with national science education reform: Entanglements or intelligent interconnections? *Journal for Research in Science Teaching, 34 (1),* 3-17.

Lynch, S. (2000). *Equity and science education reform.* Mahwah, NJ: Erlbaum.

Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (In press). Examining the effects of a highly rated science curriculum unit on diverse student populations: Results from a planning grant. *Journal of Research in Science Teaching.*

Lynch, S., & O'Donnell, C. (2005). *"Fidelity of implementation" in implementation and scale-up research designs: Applications from four studies of innovative curriculum materials and diverse populations.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Mihalic, S. (2002, April). *The importance of implementation fidelity.* Boulder, Colorado: Center for the Study and Prevention of Violence.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24,* 315-340.

National Science Foundation, (2000). Interagency education research initiative (IERI), Program solicitation. Arlington, VA: Author.

O'Donnell, C. (2004, June). *Fidelity of implementation: Background, definitions, and components for measuring.* Internal document: The George Washington University.

O'Donnell, C., Lynch, S., Hansen-Grafton, B. (2004). *Fidelity of implementation guidelines.* Internal document: The George Washington University.

Rogers, E. M. (2003). *Diffusion of Innovations* (5th ed.). New York: Free Press.

Roseman, J.E., Kesidou, S., & Stern, L. (1996, November). Identifying curriculum materials for science literacy: A Project 2061 evaluation tool. Paper presented for the National Research Council's Colloquium "Using the National Science Education Standards to Guide the Evaluation, Selection, and Adaptation of Instructional Materials," Washington, DC.

State of Michigan. (1993). *Chemistry That Applies*. Lansing, MI: Michigan Department of Education.

U.S. National Research Center for TIMSS, (1996). A splintered vision: An investigation of U.S. science and mathematics education. Dordrect, The Netherlands: Kluwer.