March 28, 2007

A model for fidelity of implementation in a study of a science curriculum unit: Evaluation based on program theory

Sharon J. Lynch

Prepared for the annual meeting of the American Educational Research Association, April 2007, Chicago

This work was based on the work conducted by SCALE-uP: A collaboration between George Washington University and Montgomery County Public Schools (MD); Sharon Lynch, Joel Kuipers, Curtis Pyke, and Michael Szesze serve as principal investigators of SCALE-uP. Funding for SCALE-uP was provided 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 author and do not necessarily reflect the position, policy, or endorsement of the funding agencies.

Correspondence to: Sharon J. Lynch; E-mail; slynch@gwu.edu; Work Phone: 202-994-6174 and FAX: 202-994-3365

Abstract

This paper proposes a conceptual framework for fidelity of implementation and its measures in the context of a quasi-experimental effectiveness study of a middle school science curriculum unit, *Motion and Forces (M&F*). In this study of *M&F*, measures of fidelity were based upon the intervention's program theory, and included the intervention's processes and structure from the points of view of both teacher and students. By using multiple measures of fidelity of implementation for treatment and comparison classrooms and correlating fidelity measures with student outcomes, fidelity of implementation not only provides evidence for the internal validity of the study, but also provides a rich view of how and why the implementation "works", and supports the intervention's program theory.

A model for fidelity of implementation in a study of a science curriculum unit: Evaluation based on program theory

Introduction

Recent emphasis on "evidence-based research" in education (National Research Council [NRC], 2005; U.S. Department of Education [USDOE], 2003) has been accompanied by a renewed and keen interest in the construct "fidelity of implementation" (USDOE, 2003; NRC, 2004; National Science Foundation [NSF], 2000). *Fidelity of implementation* may be defined as the extent to which the delivery of an intervention adheres to the program model originally developed, and confirms that the implementation of the independent variable in outcome research occurred as planned (Mowbray, Holter, Teague & Bybee, 2003). Not only does the understanding and measurement of fidelity of implementation have increasing import for education researchers, practitioners in school systems have also found this construct useful (and vexing) as they seek to understand the effectiveness of interventions initiated in response to pressures to improve student performance in the era of No Child Left Behind (USDOE, 2001).

Reasons for current interest in fidelity of implementation criteria are straightforward: As there are increased efforts to introduce interventions to improve student outcomes and to demonstrate the scope of their effectiveness among diverse students, for a given intervention, there is a need to know:

- If the intervention was actually implemented (a fidelity check);
- The extent to which the intervention was implemented (using instruments that can provide a range on fidelity indicators);

- Whether there is a relationship between fidelity of implementation and student outcomes (an indication of the intervention's "worth", requiring fine-tuned measures of fidelity capable of discriminating among classroom enactments of interventions); and,
- Whether the intervention was much different than typical practices in "comparison" classrooms (necessary to understand the "value added" for an intervention).

Effectiveness research has sufficiently advanced that black-box outcome studies are unacceptable (Mowbray et al., 2003). Researchers should present evidence not only that the intervention "worked", but why and how it worked based upon its program theory. This requires an understanding of the role of theory in evaluation research, and such program-theory-based evaluation research not only is a means of ascertaining the efficacy or effectiveness of an educational intervention, it becomes a means of theory-testing (Clements, 2007).

The purpose of this paper is to offer a conceptual framework for understanding and measuring fidelity of implementation for middle school science curriculum units in a large six-year study of the effectiveness and scale-up of curriculum materials (Lynch, Kuipers, Pyke, & Szesze, 2002). It focuses on *Exploring Motion and Forces: Speed Acceleration, and Friction (M&F)*, a curriculum unit developed by the Harvard-Smithsonian Center for Astrophysics (2001), and explores the relationship between student outcomes and fidelity ratings. Convergent measures of fidelity in the study of *Motion and Forces (M&F)*, in turn, should:

- Support or refute the learning theory driving this study;

- substantiate the framework offered for understanding fidelity of implementation; and,
- suggest ways to improve *M&F* and/or its delivery in classrooms.

Study Context

Overview of SCALE-uP

In 2001, researchers at The George Washington University (GWU) and educators in Montgomery County Public Schools (MCPS), Maryland, initiated a six-year study of three middle school science curriculum units that have highly specified instructional characteristics in common. The research included one planning grant year (Year 0: 2001-2002) and five years of curriculum implementation studies (Years 1 – 5: 2002 – 2007). The purpose of the Scaling up Curriculum for Achievement, Learning, and Equity Project (SCALE-uP) was to conduct research on the implementation of three science curriculum units using quasi-experimental research design and scale them up if they proved to be effective. If a unit was found to be effective after two trials (outcomes for students in five treatment middle schools were compared with a matched set of five comparison schools for at least two consecutive years), then SCALE-uP explored additional research questions about their movement to scale in 35 MCPS middle schools. *Scale-up* can be defined as the transition from idiosyncratic adoption of curriculum units to broad, effective implementation across a large and diverse school system, or as the deliberate expansion to many settings of an externally developed school restructuring design that has previously been used successfully in one or a small number of school settings (Stringfield & Datnow, 1998). SCALE-uP also explored how these units functioned

culturally and linguistically in classrooms, using ethnographic inquiry into video data of complete enactments of the units.

The three curriculum units studied in SCALE-uP are:

- 8th Grade: *Chemistry That Applies (CTA)* (Michigan Department of Education, 1993);
- 7th Grade: *GEMS: The Real Reasons for Seasons: Sun-Earth Connections* (Lawrence Hall of Science, University of California at Berkeley, 2000); and,
- 6th Grade: *ARIES: Exploring Motion and Forces: Speed, Acceleration, and Friction (M&F)* (Harvard-Smithsonian Center for Astrophysics, 2000).

MCPS is the seventeenth largest school system in the U.S. and an ideal testbed to explore scale-up because of its size, the diversity of its student population, and its well-organized administrative structure. Before scaling-up these science units to thousands of MCPS middle school students, however, it was imperative to know that the units were both effective and equitable. Although each unit had undergone field tests conducted by its developers, no experimental or quasi-experimental effectiveness studies had been done on any of the units prior to SCALE-uP. Moreover, to determine whether a unit was suitable for scale-up in this culturally, linguistically, and socio-economically diverse school system, a treatment unit should not only show significantly higher student outcomes than the comparison condition overall, it should also be effective for diverse subgroups of students when data were disaggregated by gender, ethnicity, or eligibility for Free And Reduced-price Meal System (FARMS), English for Speakers of Other Languages (ESOL), or special education services (Lynch, Kuipers, Pyke, & Szesze,

2005; Lynch, Taymans, Watson, Ochsendorf, Pyke, & Szesze, 2007). Only then would it be eligible for scale-up.

Given the scope of SCALE-uP (nearly a hundred thousand students and their teachers in 35 middle schools were involved over six years), it was important to be able to attribute student outcomes for each intervention to the implementation of the intervention unit itself, rather than to extraneous, uncontrolled variables (Clements, 2007; Gersten et al., 2005; Lynch et al., 2006). The validity for each unit's effectiveness studies was tempered by evidence that each was implemented with acceptable fidelity.

Effectiveness Studies for M&F Over Three Years

This paper focuses on *M&F's* implementation in Year 4 of SCALE-uP (2005-06). *M&F* is a six-week physical science curriculum unit designed for students in grades 5-8. Its 18 explorations are activity-based, with an emphasis on students' direct experience with phenomena to build individual conceptual models of motion and force. *M&F* curriculum materials include a *Teacher Manual* and a student *Science Journal*.

M&F presented special challenges for SCALE-uP because the results from the first two sets of trials in Years 2 and 3 (n = 2,170 and 2,252 students, respectively) were ambiguous; overall effect sizes were small, Cohen's d = .10.and .06 respectively; and, patterns of disaggregated outcome data were troubling for the equity litmus test—*M&F* showed positive outcomes for some subgroups, but the comparison condition seemed more favorable to others (Rethinam, Pyke, & Lynch, 2007). Consequently, rather than scaling-up *M&F* in Year 4 (2005-06) as planned, instead SCALE-uP replicated the effectiveness study of *M&F* in four new matched pairs of middle schools (n = 1,761 students), carefully controlling for threats to internal validity by virtually eliminating

possible pre-test effects and employing several convergent measures of fidelity of implementation.

Results for *M&F* in Year 4 were far more encouraging than for Years 2 and 3. A 1 X 2 between-groups Analysis of Variance indicated a statistically significant main effect in favor of the *M&F* treatment, with $F(1, 1760) = 24.49$, $p < .01$, Cohen's $d = .23$. Tests for interactions for various subgroups of students revealed that curriculum condition interacted with ethnicity and with FARMS status. In addition, mean outcome scores were higher for all subgroups for *M&F* compared to the comparison condition, with the exceptions of students with prior FARMS status and for Asian American students (for which there were no statistically significant differences between treatment and comparison subgroups). Unlike *M&F's* pattern of results for Years 2 and 3, in Year 4 nearly every subgroup had greater achievement with *M&F* than with materials used in the comparison condition (Watson, Pyke, & Lynch, 2007). In addition, HLM analyses of Year 4 data controlling for individual and classroom-level variables, found an effect size of .56 favoring *M&F* at the classroom level (see Rethinam, Pyke, & Lynch, 2007).

The focus of the present paper is to present and explore a conceptual framework for understanding fidelity of implementation for the *M&F* effectiveness study in Year 4. Because effectiveness studies that have well-developed conceptualizations of fidelity of implementation for science and mathematics curriculum materials are not common (Clements, 2007; NRC, 2004), a goal of this paper is to present a framework for understanding and measuring fidelity and demonstrate how the framework informs program theory undergirding the intervention.

A Conceptual Framework for Fidelity of Implementation

Program Theory and Fidelity of Implementation

In order to understand the fidelity of implementation model for this study, it is important to understand fundamental premises adopted by the SCALE-uP researchers in 2000. SCALE-uP was initially conceived as the study of curriculum units "highly rated" according to AAAS Project 2061's Curriculum Analysis (Kesidou & Roseman, 2002). In the mid-1990's, Project 2061 had developed a system for analyzing science and mathematics curriculum materials that were focused on national science content standards/benchmarks. Project 2061's Curriculum Analysis relied upon the research base of the learning sciences, including general findings on knowledge organization in expertise, the role of prior knowledge in learning (conceptual and cultural), and conditions that facilitate the transfer of knowledge (Kesidou & Roseman, 2002). Table 1, derived from Kesidou and Roseman's article (2002), shows the relationship of the first five Categories of the Curriculum Analysis to Project 2061's review of the research on student learning, c. 1996. The *program theory*, or scientifically sound theory of action (cf Chatterji, 2004; Clements, 2007; Mowbray et al., 2003; NRC, 2004), to be tested by SCALE-uP, therefore, was embodied in Project 2061's Curriculum Analysis (AAAS, 2003; Kesidou & Roseman, 2002).

[Insert Table 1 about here, please.]

Project 2061's Curriculum Analysis consists of criteria consonant with theories of learning on the learning cycle (Categories I, III and IV), conceptual change (Categories II and V), social constructivism (Category V), and cognitive apprenticeship (Categories IV and V). There is overlap among Categories as there is among the learning theories cited in Table 1 (J. Roseman, personal communication, December 11, 2006). In addition,

Category II, asks whether curriculum materials' developers have attended to the specific research base on student conceptions that are the object of their object, implicitly including a "learning model or trajectory" for concepts and skills in a particular science domain that are research-based (Clements, 2007).

SCALE-uP's initial program theory was based upon Project 2061's Curriculum Analysis (Lynch et al., 2005; Lynch, 2006a). SCALE-uP intended to explore how three different sets of "highly rated" curriculum materials functioned as tools for teaching and learning if implemented with reasonable fidelity in a large school system. SCALE-uP adopted this program theory, not as unswerving commitment to a sacred text, but rather to explore Project 2061's broad set of criteria for evaluating written curriculum materials in the context of a circumscribed set of field studies. Was each of the three units chosen for study effective? Equitable? None had been evaluated in this way. SCALE-uP was agnostic about whether curriculum materials having high ratings on the Curriculum Analysis would be effective for all subgroups of students (Lynch, 2006a); some students might be disadvantaged by such materials because of cultural differences in how science is best learned (cf. Lee, ___; Lynch, 2000). However, there is little empirical research on how different subgroups of students respond to science curriculum materials (Lee,). An initial goal of SCALE-uP, therefore, was to better understand whether three curriculum units vetted and described via Project 2061's Curriculum Analysis could help all subgroups of students to learn important science concepts. SCALE-uP would provide three cases of curriculum materials that supported or problematized the program theory embodied by the Project 2061 Curriculum Analysis.

In summary, although Project 2061's Curriculum Analysis had its origins in learning theory and research, a Curriculum Analysis conducted by science education experts to evaluate *written* curriculum materials is not the same as field-based trials of *implemented* curriculum materials. Consequently, SCALE-uP's program theory begins with that of Project 2061, and seeks to provide evidence to support, extend, refute or complicate it, for each middle school science curriculum unit explored.

However, as the research progressed, it was increasingly important to determine that the units were implemented with fidelity. If they were, then such evidence would link instructional strategies built into the units with classroom enactments. Roseman points out that:

> A legitimate criticism of the 2061 criteria might be that they are more about effective teaching than about effective curriculum materials. We know very little about how teachers use curriculum materials. Including in the teacher guide all the supports recommended in the 2061 criteria will have little impact on student learning if the teachers don't read or understand or make use of the guidance provided (J. Roseman, personal communication, December 11, 2006).

Using the Project 2061's Curriculum Analysis meant that each middle school science curriculum unit chosen for SCALE-uP had to focus on its own challenging science standard/benchmark/set of target ideas. Moreover, the print curriculum materials for each unit had to show evidence of containing 16 instructional strategies in Categories I-V of the Curriculum Analysis (see the first column of Figure 1).

[Insert Figure 1 about here please with CTA M&F and Seasons Categories 1-5.]

However, when the three units were subjected to the Project 2061 Curriculum Analysis at the start of SCALE-uP, evaluators determined that *CTA* had incorporated most of the instructional strategies identified in the Project 2061 Curriculum Analysis into its print materials, but *Seasons* and *M&F* were judged as meeting about half of the instructional criteria (see the results for each unit in Figure 1). Although all three units had better ratings than a typical textbook (shown in the last column of Figure 1), SCALE-uP's initial assumptions that all three units would be highly rated (that is, rate *Excellent* or *Satisfactory* on most of the Project 2061 criteria) were not fulfilled. Moreover, the three units had different profiles for criteria met. .

This raises important issues about program theory for each unit and for measures of fidelity of implementation. Figure 1 shows that, with the exception of *CTA*, the program theories of *Seasons* and *M&F* were not as congruent with Project 2061's Curriculum Analysis as SCALE-uP had initially expected. Direct discussions (oral, by email, and by formal written responses to structured queries) with the developers of *Seasons* and *M&F* confirmed this. Each developer reported areas of agreement and disagreement with Project 2061's (and SCALE-uP's) program theory, reflected by Curriculum Analysis ratings in Figure 1 (see Ochsendorf, Lynch, & Pyke, 2006; and, O'Donnell, Watson, Pyke, & Lynch, 2006 for more information on each analysis). Consequently, a framework for fidelity of implementation for *M&F* in Year 4 had to capture not only the classroom enactments of instructional strategies consistent with Project 2061's Curriculum Analysis, it also had provide a close reading of the fidelity to the unit as it was conceived by the developer.

Developing a Conceptual Framework for Fidelity of Implementation

Process and Structure Criteria for Fidelity of Implementation

Fidelity of implementation is a construct that is routinely employed and measured in intervention studies in the fields of mental health (Bond, Evans, Salyers, Williams, & Kim, 2000) and in behavior change studies conducted in health education (Resnick et al., 2005). But fidelity of implementation measures are less developed and seem undertheorized in effectiveness studies of curriculum and instruction where the results are based on student learning of discipline content (Clements, 2007; Gersten et al, 2005; Lynch & O'Donnell, 2005). For example, in an evaluation of the quality of K-12 mathematics curriculum material evaluations, the NRC reported that 33 of the 63 "*at least minimally methodologically adequate"* comparative studies reviewed by the NRC did not report any measure of fidelity; and, of the 30 studies that did measure fidelity, only 1 reported and adjusted for it when interpreting its outcome measures (NRC, 2004, p. 115, emphasis author's). O'Donnell's review of the education literature on fidelity of implementation (2007) found that only 5 of 25 primary studies of program effectiveness met the methodological criteria for measuring the relationship between fidelity of implementation to K-12 core curriculum materials and participant outcomes.

To develop a conceptual framework for fidelity in the present science curriculum unit effectiveness study, the SCALE-uP fidelity research group found extremely useful two reviews of the literature from the mental health fields (Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003). In addition, an article by Mowbray et al. (2003) on the development, measurement, and validation of fidelity criteria made it clear that there are no existing simple formulae or set of guidelines to apply when studying fidelity in curriculum materials. Rather, it is up to the researcher conducting the

evaluation and the researcher/developer who designed the intervention to create fidelity criteria and measures. The developer's role, ideally, is to clarify the intervention's program theory or the theoretical design underlying the intervention, and to determine its critical components, identifying the key aspects of process and structure crucial to the internal validity of the study. The evaluator's role is to determine how to capture these elements by developing criteria for fidelity and designing corresponding measures.

Mowbray et al. (2003) suggest that such criteria may fall into two general categories, *process* and *structure.* Process fidelity is the way that the program is delivered. In the mental health literature, process criteria can include program style, client-staff interactions, client-client interactions, or emotional climate. Rating performances for specified process criteria tends to be subjective and is often conducted via observations and interviews. Process fidelity is roughly equivalent to Gersten et al's fidelity *quality* (how well the intervention was implemented). Capturing process fidelity is labor intensive and expensive; reliability can be an issue (Mowbray et al., 2003). If an intervention's program theory is not clearly articulated by its developers, then process measures are especially challenging for evaluators.

Program structure fidelity may more directly address whether the intervention was implemented at all, and is seems close to Gersten et al's *surface* fidelity (the expected intervention was in fact implemented). Process fidelity includes a frame for delivery of service; in the mental health literature this may include measures of staffing levels and their characteristics, procedure checks, and frequency of contact with program participants. Structure fidelity can be captured via checklists and other straightforward

frequency devices. Structure performance measures for fidelity may be less costly and more reliable than process measures.

It is likely that both process and structure criteria are required to convincingly capture fidelity of implementation. Neither alone would provide sufficient information to be assured of study validity. Model drift (deviating from what was intended) may occur more often on the process side of fidelity because processes are more subject to the implementer's discretion (Teague, Drake, & Ackerson, 1995). However, process fidelity measures provide differentiation that is more likely to discriminate between fidelity levels among implementers.

Teacher and Student Roles in Studying Fidelity of Implementation

A commonly held view in the education community on curriculum implementation is that the delivery of curriculum materials is linear and mediated by the teacher. In this view, outcomes are dependent on the fidelity with which the teacher delivers the curriculum to the students, depicted in the top part of Figure 2. The teacher has the central role, while students are the recipients of the teacher's instruction, no matter the instructional materials. (There are alternative views that hold that outcomes are determined by the teacher's ability to *modify* the curriculum materials for his or her students and this view places the teacher in an even more central position; a system to measure teacher modifications and fidelity is far beyond the scope of the present study.)

[Insert Figure 2 about here please.]

However, research on how students learn from media or from health education interventions commonly place students in a more active and central role in the instructional process (Elias, Zins, Graczyk and Weissburg, 2003). For instance, an

evaluation of a health education intervention on smoking cessation asked participants to report on how motivating the class sessions were, or the degree to which they report incorporating the changes suggested in their own lives (Williams, Minicucci, Kouides et al., 2002 cited in Resnick et al., 2005). In another example, research on student classroom use of and learning from Internet resources places students at center stage when gathering information from web searches; the teacher is seen in an supervisory role to help students to conduct searches or evaluate the quality of the information obtained, rather than as the agents of information transmission (Kuiper, Volman, & Terwel, 2005). These alternative views of student agency and involvement with classroom-based curriculum materials suggest that not only are teachers implicated in curriculum implementation, so too are the student participants. The curriculum is enacted by both, together. In this view, depicted in the bottom half of Figure 2, students are responsible actors in an activity system that uses curriculum materials as a tool for learning, with teachers crucial to facilitation (Lynch, 2006b).

Fidelity of Implementation for M&F

The conceptual model developed for fidelity of implementation in SCALE-uP reflects both process and structure criteria, informed by both teacher and student participation. Each aspect of this model is discussed below.

Teacher Process Fidelity and Project2061-ness. For SCALE-uP's studies of middle school science curriculum materials, process fidelity initially seemed at the heart of the matter. If curriculum materials contain certain instructional strategies consistent with learning theory endemic to Project 2061's rating system, then such treatment materials ought to produce higher student outcomes for the target concept than those used

in comparison classrooms where the same content is taught without the benefit of such curriculum materials. Thus, a teacher's ability to find and implement instructional strategies built into the unit's student guide and further described in the teacher manual would be critical to the unit's effectiveness. We refer to this process fidelity component as "Project2061-ness," meaning the number of and extent to which the instructional strategies found in the Project 2061 Curriculum Analysis are in evidence (corresponding to 16 criteria in Categories I-V of the Analysis). Thus, if a curriculum unit met many of these criteria, it had a lot of "Project2061-ness." If a teacher was observed using many of these instructional strategies when implementing the unit, then that teacher was teaching with "high Project2061-ness." Teacher process fidelity (or teacher Project2061-ness) was captured by classroom observation instruments based on SCALE-uP's interpretation of Project 2061's criteria for evaluating curriculum materials. (See Figure 3.)

[Insert Figure 3 about here, please.]

Student Process Fidelity and Project2061-ness. Classroom observations of teachers implementing curriculum materials allowed a clear view of the teacher's use of the instructional strategies designed into them. However, because many student activities were conducted in lab groups, whole class observations were unsatisfactory for capturing student process fidelity. SCALE-uP's video data on student group discourse during lessons made a convincing case that students were interacting with the curriculum materials in important ways (Kuipers, Viechnicki, Massoud, & Wright, in press). Video for groups of four students working on lab activities showed students referring to the written curriculum materials, reading them aloud, responding to written questioning prompts orally and in writing, and actively interacting with the text. This occurred

whether the teacher was present in the group or not. Students were asked by *M&F* written materials to make predictions, gather and analyze lab data, and to reason from the evidence gathered about the physical phenomena central to each unit. Students could choose to participate, or not. Although the teacher is important to implementing the curriculum unit, once it is underway and the pattern of the lessons is understood by student groups, then the teacher's overt centrality seems to fade, ineluctably. Activities were not mediated "through" the teacher, but appeared to be a set of interactions between students in the lab group, curriculum materials, and physical phenomena at hand, with the teacher drifting in and out. (However, the teacher was crucial to administrative tasks and keeping order, often explaining procedures and scaffolding student understanding of results.) SCALE-uP developed a measure for student process fidelity of implementation based upon student self-reports on a questionnaire inquiring about their perceptions of their experiences with the instructional strategies in the unit. This approach is consistent with what Dane and Schneider (1998) called *participant responsiveness*, defined as the level of participation and enthusiasm of the participants.

Fidelity as Teacher Adherence to Structure. As measures of fidelity to Project2061-ness were developed, two problems became apparent. The first was that the curriculum developers had not always adopted all of the instructional strategies contained in the Project 2061 Curriculum Analysis and had somewhat different notions of program theory; therefore, a class observation instrument based upon such process criteria alone was likely to be inadequate. A second problem was that the instructional strategies in the Project 2061 Curriculum Analysis (and designed into the curriculum units to some extent) were also often also observed in comparison classrooms in this quasi-

experimental study. Comparison teachers had attended professional development workshops emphasizing similar instructional strategies and had incorporated them into their daily practice. Capturing Project2061-ness, therefore, was a helpful but insufficient measure of fidelity of implementation for *M&F,* and fidelity criteria more closely and uniquely linked to the *M&F* curriculum unit were required. We needed a more exact and direct method to ascertain whether the treatment unit was taught as intended by the developer—adherence. Such direct measures of adherence to *M&F* would be based on its written materials for teachers and students, and thus can be seen as fidelity to the unit's structure. Necessarily, adherence fidelity measures would be limited to treatment classrooms only. (It was impossible to develop parallel measures for the comparison condition because teachers chose from a range of curricular options.) An instrument was created to capture what *M&F* teachers did or did not include in each observed lesson, based on what the developer explicitly intended for a lesson. Each lesson was "mapped" according to the *M&F* student *Science Journal* and *Teacher Manual*, and a classroom observer used a checklist to indicate the points on the map followed, or not. Fidelity as teacher adherence could be measured by observing how closely teachers followed each component of a *M&F* lesson.

Fidelity as Student Adherence to Structure. A parallel criterion for student adherence to the intervention unit was suggested by Songer and Gotwals' measure of fidelity of implementation to the *BioKids* science curriculum units (2005). Songer and Gotwals had used the number of worksheets that students had completed as an indicator of levels of student structure fidelity, and found that this simple measure predicted

student outcomes well in regression analyses. Thus, a direct measure of student involvement with each lesson could indicate student adherence to structure.

M&F included a student *Science Journal* that had question prompts for students for each lesson. We used students' responses to the *Science Journals*, grouped by classroom, to determine student adherence to structure. This was done by counting the number of questions each student answered over the entire unit, and averaging them by classroom.

Summary

Figure 3 depicts four categories for fidelity of implementation and the measures used to explore them. Figure 3 is a conceptual rather than structural model, and its arrows represent hypothesized relationships. Although it may be more straightforward to consider fidelity to structure (did the intervention occur?) prior to considerations to fidelity to process (how the intervention occurred), SCALE-uP initial goal was to explore curriculum materials that possessed the instructional strategies contained in Project 2061 Curriculum Analysis. Consequently, we will focus on process first, providing a view not only a view of what was occurring in M&F classrooms, but also of comparison classrooms. This paper explores some relationships in Figure 3, linking fidelity measure to outcomes mathematically, and represented by solid arrows. Other relationships no doubt exist, depicted by dotted arrows in Figure 3. Intuitively, there are likely relationships between structure and process fidelity, in general, as well as relationships between student and teacher fidelity in classrooms where there are teacher and students work together to support enactments of *M&F*. Although these relationships depicted by

dotted arrows in Figure 3 are beyond the scope of the present paper, they are important to the conceptual model.

Given the framework in Figure 3 for understanding and measuring fidelity of implementation, we use the Year 4 *M&F* implementation study to:

- Describe the fidelity with which the unit was being implemented according to teacher and student structure and process criteria;
- Compare process measures of fidelity of implementation in treatment classes with what was occurring in comparison classrooms to grasp similarities and differences; and,
- Explore the relationship between fidelity of implementation and student outcomes, to infer aspects of implementation critical to student learning.

References

American Association for the Advancement of Science. (2003). *Project 2061 middle grades science textbooks: A Benchmarks-based evaluation*. Retrieved June 1, 2004, from http://www.project2061.org/tools/textbook/mgsci/index.htm.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Researc*h, 2(2), 75–87.

Chatterji, M. (2004). Evidence on "what works": An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher, 33*(9), 3-13.

Clements, D. H. (2007). Curriculum research: Toward a framework for 'research-based curricula'. *Journal for Research in Mathematics Education, 38*, 35-70.

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.

Elias, M.J., Zins, J.E., Graczyk, P.A., Weissburg, R.P. (2003). Implementation, Sustainability, and Scaling Up of Social-Emotional and Academic Innovations in Public Schools. *School Psychology Review*, 32, 303-319.

Harvard-Smithsonian Center for Astrophysics. (2001). *ARIES: Exploring motion and forces: Speed, acceleration, and friction.* Watertown, MA: Charlesbridge Publishing.

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(6),* p. 522-549

Kuiper, E., Volman, M., & Terwel, J. (2005). The web as an information resource in K-12 education: Strategies for supporting students in searching and processing information. *Review of Educational Research, 75*(3), 285-317.

Kuipers, J.C., Viechnicki, G.B., Massoud, L., & Wright, L.J. (in press). Science, culture, and equity in curriculum: An ethnographic approach to the study of a highly-rated curriculum unit. In K. Richardson and K. Gomez (Eds.), Talking science, writing science: The work of language in multicultural classrooms. Mahwah, NJ: Lawrence Erlbaum Associates.

Lawrence Hall of Science. (2000). *The real reasons for seasons—sun-earth connections.* Berkeley: The Regents of the University of California.

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

Lee, O., & Luykx, A. (in press). *Science education and student diversity: Synthesis and research agenda.* New York: Cambridge University Press.

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

Lynch, S. (2006a, April). *An overview of SCALE-uP and results for Chemistry That Applies. Are "highly-rated" middle school science curriculum materials effective and for whom?: Results from a set of implementation studies.* Paper presented at

the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Lynch, S. (2006b). *ISO metaphor and theory for scale-up research: Eagles in the Anacostia and activity systems.* Manuscript submitted for publication.

Lynch, S., Kuipers, J.C., Pyke. C., & Szesze, M. (2002). NSF/IERI proposal, *Scaling up highly rated science curricula in diverse student populations: Using evidence to close achievement gaps.* Washington, DC: The George Washington University.

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

Lynch, S. & O'Donnell, C. (2005, April). The evolving definition, measurement, and conceptualization of fidelity of implementation in scale-up of highly rated science curriculum units in diverse middle schools. In S. Lynch (Chair), *The role of fidelity of implementation in quasi-experimental and experimental research designs: Applications in four studies of innovative science curriculum materials and diverse student populations.* Symposium conducted at the Annual Meeting of the American Educational Researchers Association, Montreal, Canada.

Lynch, S., O'Donnell, C., Hatchuel, E., Rethinam, V., Merchlinsky, S., & Watson, W. (2006, April). *What's up with the Comparison group?: How large quasi-experimental study of highly rated science curriculum units came to grips with unexpected results.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lynch, S., Taymans, J. Watson, W., Ochsendorf, R., Pyke, C. & Szesze, M. (2007). Effectiveness of a highly-rated science curriculum unit for students with disabilities in general education classrooms. Exceptional Children, 73(2), 202-223.McDonald, S.K., Keesler, V.A., Kaufman, N. J., & Schneider, B. (2006) Scaling-up exemplary interventions. *Educational Researcher, 35*(3), 15-24.

Michigan Science Education Resources Project. (1993). *GEMS: Chemistry That Applies.* The State of Michigan.

Mowbray, C., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluatio*n, *24*(3),315-340.

National Research Council (NRC). (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials. Mathematical Sciences Education Board, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NRC. (2005). *Advancing scientific research in education.* Committee on Research in Education. L. Towne, L. L. Wise, and T. M. Winters, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

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

Ochsendorf, R., Lynch, S., & Pyke, C. (2006). *Rating a science curriculum unit: Perspectives on the program theory.* Manuscript submitted for publication.

O'Donnell, C. (2007). *Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research.* Manuscript submitted for publication.

O'Donnell, C. L., Watson, W. A., Pyke, C., & Lynch, S. (2006). *Understanding a quasi-experimental study of a seasons unit: Examining the intended, implemented, and attained curriculum.* Manuscript submitted for publication.

Resnick, B., Bellg, A. J., Borrelli, B., DeFrancesco, C., Breger, R., Hecht, J., Sharp, D. L., Levesque, C., Orwig, D., Ernst, D., Ogedegbe, G., & Czajkowski, S. (2005). Examples of implementation and evaluation of treatment fidelity in the BCC Studies: Where we are and where we need to go. *Annals of Behavioral Medicine, 29,* 46-54.

Rethinam, V., Pyke, C., & Lynch, S. (in press). Using Multilevel Analyses to Study Individual and Classroom Factors in Science Curriculum Effectiveness. *Evaluation and Research in Education.*

Songer, N. B., & Gotwals, A. W. (2005, April). Fidelity of implementation in three sequential curricular units. In S. Lynch (Chair), *"Fidelity of implementation" in implementation and scale-up research designs: Applications from four studies of innovative science curriculum materials and diverse populations.* Symposium conducted at the meeting of the Annual Meeting of the American Educational Research Association. Montreal, Canada.

Stringfield, S., & Datnow, A. (1998). Scaling up school restructuring designs in urban schools. *Education and Urban Society, 30*(3), 269-276.

Teague, G. B., Drake, R. E., & Ackerson, T. H. (1995). Evaluating use of continuous treatment teams for persons with mental illness and substance abuse. *Psychiatric Services, 46*, 689-695.

Watson, W., Pyke, C., Lynch, S., Ochsendorf, R. (2007, April). Understanding the effectiveness of curriculum materials through replication. Paper to be presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

U. S. Department of Education (USDOE). (2003, December). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide*. Retrieved August 29, 2005 from http://www.ed.gov/print/rschstat/research/pubs/rigorousevid/guide.html.

USDOE, Office of Elementary and Secondary Education. (2001). *No Child Left Behind: A Desktop Reference,* Washington, D.C. Retrieved January 3, 2007 from http://www.ed.gov/admins/lead/account/nclbreference/page.html.

Table 1
Learning theory and representative studies establishing the research base for Project 2061 Curriculum Analysis (Source: Kesidou & Roseman, 2002)

Instructional Analysis Categories and Criteria	Learning Theory	Representative Studies Establishing Research Evidence
I. Identifying a Sense of Purpose Conveying unit purpose Conveying lesson/activity purpose Justifying lesson/activity sequence	Knowledge organization The learning cycle	Clear understanding of purpose, goals, and content of activity has positive effects on student learning (Boulanger, 1981; Wise & Okey, 1983) Student interest in or recognition of value of activity needs to be motivated to derive intended learning benefits from engaging in an activity (Malone & Lepper, 1987; Blumenfeld et al., 1991)
II. Taking Account of Student Ideas Attending to prerequisite knowledge/skills Alerting teacher to commonly held student ideas Assisting teacher in identifying students' ideas Addressing commonly held ideas	Role of prior knowledge in learning (conceptual and cultural) Conceptual change	Fostering student understanding requires taking time to attend to ideas for subsequent learning (Eaton, Anderson, & Smith, 1984; Minstrell, 1984; Roth, 1991) Materials alerting teachers to probable student misconceptions and suggesting strategies for identifying and addressing them can lead to improved student understanding (Bishop & Anderson, 1990; Brown & Clement, 1992; Eaton et al., 1984; Lee, Eichinger, Anderson, Berkheimer, & Blakeslee, 1993)
III. Engaging Students with Relevant Phenomena Providing a variety of phenomena Providing vivid experiences	The learning cycle	Students need opportunities to relate scientific concepts to a range of appropriate phenomena through hands-on activities, demonstrations, audiovisual aids, and discussions of familiar phenomena (Anderson & Smith, 1987) Students learn readily about things tangible and accessible to their senses and benefit from firsthand experiences with phenomena (Boulanger, 1981; Wise & Okey, 1983; Kyle, Bonnstetter, Gadsden, & Shymansky, 1988)
IV. Developing and Using Scientific Ideas Introducing terms meaningfully Representing ideas effectively Demonstrating use of knowledge Providing practice	The learning cycle Cognitive apprenticeship	Experiences with phenomena are not sufficient to understand science principles and concepts; multiple representations and explicit instruction are needed to make ideas intelligible (Driver, 1983; Smith & Anderson, 1984; Champagne, Gunstone, & Klopfer, 1985; Strike & Posner, 1985; Feltovich, Spiro, Coulson, & Anderson, 1989) Students need help understanding how ideas can be used to describe and explain phenomena, solve practical problems, or consider alternative positions on issues (Anderson & Roth, 1989) and opportunities to apply ideas in a variety of contexts with extensive practice (Hayes, 1985; Ericsson, Krampe, & Tesche-Romer, 1993)
V. Promoting Student Thinking about Phenomena, Experiences, and Knowledge Encouraging students to explain ideas Guiding student interpretation and reasoning Encouraging students to think about what they've learned	Knowledge transfer Conceptual change Social constructivism Cognitive apprenticeship Knowledge distribution	Students need guidance to make sense of experiences and ideas (Driver, 1983) Carefully chosen and sequenced questions and tasks are necessary to scaffold students' attempts to construct intended meaning of experiences or presentations of ideas (Anderson & Smith, 1987; Anderson & Roth, 1989; Arons, 1990) When students make their thinking about experiences and ideas overt to themselves, the teacher, or other students, thinking can be examined, questioned, and shaped (Needham, 1987; Linn & Burbules, 1993; Glaser, 1994)

Figure Captions

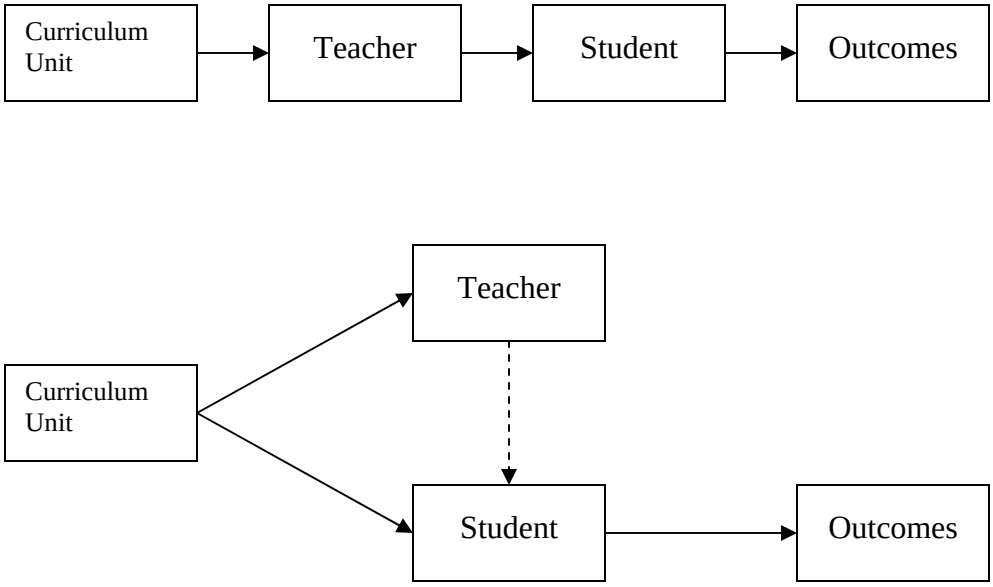
Figure 1. AAAS Project 2061 Curriculum Analysis criterion-level ratings for Categories I - V for *Chemistry That Applies*, *Exploring Motion and Forces*, *The Real Reasons for Seasons*, and a middle school physical science textbook published by MacMillan/McGraw Hill.

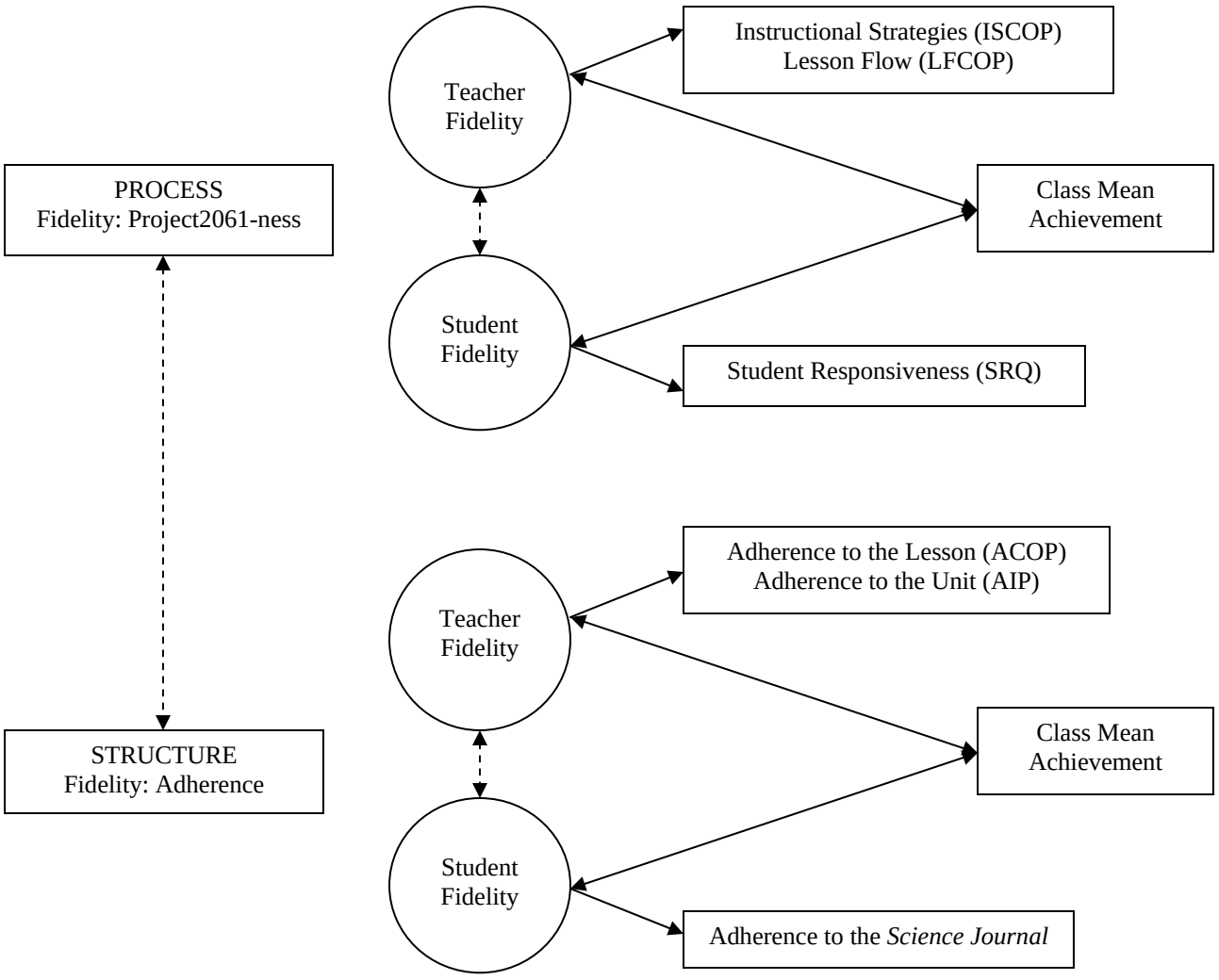
Figure 2. Two views of curriculum implementation represented in the science education literature.

Figure 3. Conceptual model for studying fidelity of implementation in SCALE-uP, reflecting both process and structure criteria and informed by both teacher and student participation.

Instructional Analysis Ratings Excellent=●; Very Good=◕; Satisfactory=◒; Fair=◔; Poor=○; N/R=No Rating **Instructional Categories**		*Chemistry That Applies**	*Exploring Motion and Forces***	*The Real Reasons for Seasons***	*Macmillan/McGraw-Hill Science**
I. Identifying a Sense of Purpose					
Conveying unit purpose		○	○	N/R	◔
Conveying lesson/activity purpose		◕	◒	●	◔
Justifying lesson/activity sequence		◒	◒	◒	○
II. Taking Account of Student Ideas					○
Attending to prerequisite knowledge and skills		◒	○	○	○
Alerting teacher to commonly held student ideas		◒	○	N/R	○
Assisting teacher in identifying own students' ideas		◒	◒	○	○
Addressing commonly held ideas		◒	○	◒	○
III. Engaging Students with Relevant Phenomena					
Providing variety of phenomena		●	●	○	○
Providing vivid experiences		●	●	◒	○
IV. Developing and Using Scientific Ideas					
Introducing terms meaningfully		●	◒	◒	◔
Representing ideas effectively		◒	◒	●	○
Demonstrating use of knowledge		◕	○	◒	○
Providing practice		●	○	○	○
V. Promoting Student Thinking about Phenomena, Experiences, and Knowledge.					
Encouraging students to explain their ideas		●	◒	○	○
Guiding student interpretation and reasoning		●	○	●	○
Encouraging students to think about what they've learned		○	○	○	○

*Ratings conducted by Project 2061.
**Ratings conducted by SCALE-uP.


Curriculum Unit
Teacher
Student
Outcomes
Teacher
Curriculum Unit
Student
Outcomes


PROCESS
Fidelity: Project2061-ness
STRUCTURE
Fidelity: Adherence
Teacher Fidelity
Student Fidelity
Instructional Strategies (ISCOP)
Lesson Flow (LFCOP)
Class Mean Achievement
Student Responsiveness (SRQ)
Teacher Fidelity
Student Fidelity
Adherence to the Lesson (ACOP)
Adherence to the Unit (AIP)
Class Mean Achievement
Adherence to the *Science Journal*