JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 48, NO. 2, PP. 149–169 (2011)

Examining the Effect of Teachers' Adaptations of a Middle School Science Inquiry-Oriented Curriculum Unit on Student Learning

Jay Fogleman,[1] Katherine L. McNeill,[2] Joseph Krajcik[3]

[1]*School of Education, University of Rhode Island, 711 Chafee Building, Kingston, Rhode Island 02881*
[2]*Lynch School of Education, Boston College, Boston, Massachusetts*
[3]*School of Education, University of Michigan, Ann Arbor, Michigan*

Received 30 January 2009; Accepted 21 October 2010

Abstract: Reform based curriculum offer a promising avenue to support greater student achievement in science. Yet teachers frequently adapt innovative curriculum when they use them in their own classrooms. In this study, we examine how 19 teachers adapted an inquiry-oriented middle school science curriculum. Specifically, we investigate how teachers' curricular adaptations (amount of time, level of completion, and activity structures), teacher self-efficacy (teacher comfort and student understanding), and teacher experience enacting the unit influenced student learning. Data sources included curriculum surveys, videotape observations of focal teachers, and pre- and post-tests from 1,234 students. Our analyses using hierarchical linear modeling found that 38% of the variance in student gain scores occurred between teachers. Two variables significantly predicted student learning: teacher experience and activity structure. Teachers who had previously taught the inquiry-oriented curriculum had greater student gains. For activity structure, students who completed investigations themselves had greater learning gains compared to students in classrooms who observed their teacher completing the investigations as demonstrations. These findings suggest that it can take time for teachers to effectively use innovative science curriculum. Furthermore, this study provides evidence for the importance of having students actively engaging in inquiry investigations to develop understandings of key science concepts. J Res Sci Teach 48: 149–169, 2011

Keywords: curriculum development; inquiry; teacher beliefs; science education

The science learning goals specified in national standards documents (American Association for the Advancement of Science, 1993; National Research Council, 1996) have provided an opportunity for researchers to focus their efforts to develop classroom resources that enhance student learning on key learning goals. In addition to establishing a coherent framework for the science topics at the different grade levels, these documents suggest that students should learn science by engaging in inquiry processes that allow them an active role in their own learning and reflect how knowledge is constructed within the various scientific communities.

Reviews of traditional textbooks have called into question the degree that these textbooks support students developing deep understandings of the learning goals identified in the national standards (Kesidou & Roseman, 2002). To provide more effective classroom materials, we have developed the Investigating and Questioning Our World Through Science and Technology (IQWST) curriculum units (Krajcik, McNeill, & Reiser, 2008). One of the first two units designed for IQWST is a middle school chemistry unit, ''How Can I Make New Stuff from Old Stuff?'' or the *Stuff* unit.

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: National Science Foundation; Contract grant Numbers: ESI 0101780. ESI 0227557.

Correspondence to: J. Fogleman; E-mail: fogleman@mail.uri.edu

DOI 10.1002/tea.20399

Published online 3 December 2010 in Wiley Online Library (wileyonlinelibrary.com).

Early enactments of the *Stuff* unit in urban, suburban, and rural settings indicated that the curriculum helped teachers address their target learning goals successfully and supported student learning (Krajcik et al., 2008). During these enactments, we observed teachers choosing to enact the unit's activities in different ways. This process of teacher adaptation, or transformation, is a common occurrence when teachers use innovative materials (Pinto, 2005), and an essential step if the materials are to be used long term in these classrooms (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000; Brown & Edelson, 2001; Fullan, 1991). Consequently, it is important to better understand how teachers' curricular adaptations affect student learning. In this study, we investigated how middle school teachers' self-efficacy, their experience using the curriculum materials, and curricular adaptations of an inquiry-oriented science curriculum impacted student learning of key science learning goals.

Theoretical Framework

Designing Innovative Curriculum

Over the last decade, researchers have worked to incorporate what we currently know about teaching and learning into curriculum materials they believe will prove effective. Effective curriculum materials must meet the follow criteria: (1) Their content primarily focuses on a coherent set of important, age-appropriate student learning goals (Roseman, Linn, & Koppal, 2008); (2) their instructional design effectively supports the attainment of the specified student learning goals; and (3) the teacher's guides support teachers in helping students attain these goals (Kesidou & Roseman, 2002). The first requirement reflects the need to focus on key learning goals that help students learn fundamental scientific concepts. The second requirement specifies that the curriculum must provide support for instructional strategies that are consistent with what we know about how people learn, such as helping students to make sense of new experiences in light of what they already know, to share and refine their understandings, and to assume responsibility for their own learning (Bransford, 2000). The third requirement is that resources be provided for the teacher so that he or she can facilitate an effective learning environment and develop knowledge of students' commonly held ideas and expertise in assessing students' understanding and adapting instruction accordingly. Because of the deficiencies of the texts used in most classrooms, there is a dire need for more supportive science curricula (Kesidou & Roseman).

The Role of the Enacted Curriculum in Science Education Reform. Though curriculum materials provide critical support for teachers implementing reforms in their classrooms, students' experiences with reform-based materials depend on how teachers choose to use these resources. A tension exists in how researchers have conceptualized teachers' use of innovative curriculum materials, ranging from acknowledging teachers' role in adapting curriculum materials to stressing the need for teachers to implement new materials with fidelity to how they were designed (O'Donnell, 2008; Remillard, 2005).

In this study, we focus on understanding the teacher's role in adapting curriculum materials to meet the needs of her students and the conditions she perceives in her classroom. For Cohen and Ball (1999), curriculum materials are one element of an instructional context that the teacher must mediate while managing a learning environment. Remillard (2005) argues that while curriculum materials represent a formal curriculum that expresses learning goals and activities sanctioned by school policies or textbooks, teachers use available materials to design the enacted curriculum that is experienced by students. Teachers' adaptations of available curriculum materials are a persistent element of the cultural systems that have been used to explain the relative homogeneity of American classrooms (Squire, MaKinster, Barnett, Luehmann, & Barab, 2003; Stigler & Hiebert, 2009).

National standards documents explicitly recognize the need for teachers to have the capacity to select and adapt curriculum materials to tailor their instruction to meet their students' needs (NRC, 1996). The premise of providing curriculum materials that embody the teaching practices called for in the national standards is grounded in the reality that many teachers can benefit from supports that help them enact challenging new practices in their classrooms (Powell & Anderson, 2002; Schneider, Krajcik, & Blumenfeld, 2005). Teachers' ability to adapt materials in ways that preserve their intent while at the same time meet their

students' individual needs has been called pedagogical design capacity (Brown & Edelson, 2001) or instructional capacity (Cohen & Ball).

Teachers' decisions about how they will use curriculum materials can have both positive and negative effects on how their students experience reforms as well as what their students learn. Squire et al. (2003) found that teachers played a critical role in contextualizing an innovative science unit when students did not find the unit's central question relevant to their lives. Pinto (2005) found that teachers' transformations of reform-based innovations often demote the designers' intentions. Songer and Gotwals (2005) found that students' ability to apply scientific concepts was enhanced in classes where teachers used more of their reform-based curriculum materials. Schneider, Krajcik, and Blumenfeld (2005) found that with extensive supports, teachers could use the materials to activate students' understandings, conduct investigations, and facilitate small group discussions.

We focus in this study on discerning the effects on student learning of teachers' decisions about enacting reform-rich curriculum materials. O'Donnell (2008) argues that in order to make valid judgments about the effects of curricular innovations, researchers must be able to determine the fidelity of the teachers' use of the materials. Fidelity in this context has been defined as the extent to which a delivery of an intervention adheres to the protocol or program model originally developed (Mowbray, Holter, Teague, & Bybee, 2003). Schneider and coworkers noted that given the complexity of what was expected from teachers during the enactment of inquiry-rich curriculum materials, it was more useful for them to focus on congruence of the enactment with the designers' intentions rather than the stricter standard of fidelity.

We acknowledge that the complexity of the classroom requires each teacher to adapt materials to their setting, and are interested in discerning the effects of these decisions across several teachers. If curriculum materials are to serve as tools that embody knowledge about science education reforms that have been developed by researchers and designers to support teachers' instruction (Putnam & Borko, 1997), then enacting the materials serves a critical role in teachers deepening their understanding of reform-based teaching (Kubitskey, 2006). One way researchers can support teachers during this process is to provide evidence of the effects of their decisions on what their students learn (Fishman, Best, Marx, & Tal, 2001; Pinto, 2005).

Factors Influencing How Teachers Utilize Classroom Innovation

Analysis of past reform efforts indicated that in order for innovations to be sustained, teachers had to adapt them to meet local needs and conditions. In their review of past studies of teachers' adaptations of innovations, Pinto (2005) found that teachers' adaptations to innovations were influenced by their knowledge and beliefs about the subject they were teaching, their beliefs about their own identity and about teaching and learning, and the degree that the innovation was supported within their local contexts.

Teachers' beliefs about teaching and learning influence their use of new curriculum materials. Implementing classroom innovations often requires a teacher to change his or her practice and take on the unpleasant role of "novice" again (Fullan, 1991). A strong predictor of whether teachers can successfully meet this challenge is their sense of self-efficacy. Tschannen-Moran, Hoy, and Hoy (1998) define a teacher's self-efficacy as his or her belief in their ability to act in ways that successfully accomplish specific teaching goals. In their review of the use of the teacher efficacy construct, they found that it has been correlated with teachers' willingness to implement innovations. In other words, teachers who believe they are able to achieve specific teaching goals are more willing to try new innovations in their classroom.

Another factor that influences how innovations are enacted in classrooms is teachers' experience with the innovation. In past studies, we have seen that teachers continue to strengthen their use of reform-based curriculum materials through their second and third years using the units (Geier, 2005). Each time a teacher uses a particular innovation, we would expect an increase in both their understanding about how to use the innovation in his or her class, as well the effectiveness of their use of the innovative materials.

In addition, the local context can also have a significant impact on teachers' use of the curriculum. Our work with teachers occurs predominately in urban schools. Impoverished urban schools have been characterized as relying on a "pedagogy of poverty" in which students predominately engage in low level tasks (Haberman, 1991). Instructional innovations aimed at supporting complex scientific inquiry can be

difficult to implement in such impoverished settings due to inadequate resources, insufficient time, large class sizes, teachers' low levels of science and computer knowledge, lack of training opportunities, high levels of teacher and student mobility, limited instructional freedom, lack of administration support, and unreliable internet connectivity (Songer, Lee, & Kam, 2002). Consequently, these contextual challenges can also impact the adaptations that teachers make to the curriculum.

Teachers' Curricular Adaptations

Although we see adaptation as essential for enacting units, some adaptations can diminish the intended function of the curricular unit. Pinto (2005) identified common themes from concurrent implementation studies of four classroom innovations. Each team of researchers saw their innovations being transformed by teachers. Sometimes these adaptations were benign and sometimes problematic. Teachers tended to adapt the innovation so its use more closely resembled familiar classroom practices. In order to provide opportunities for teachers to reflect on and refine their uses of innovations in subsequent professional development workshops, it is important to be able to share with them how specific transformations affect student learning. The transformations we are concerned with in this study include how much time teachers spend on the unit, the level of completion of the unit's activities, and whether the teachers had students actively experience the unit's investigations first hand or presented them as whole-class demonstrations.

When implementing a new curriculum or other classroom innovation, the teacher must decide how much time can be spent on the new unit. There has been considerable research on how time is spent in classrooms, and the effect of these practices on student learning. When using curriculum units designed to facilitate deep conceptual understanding, students need sufficient instructional time, that is, time spent actively engaged in learning activities, to integrate their understandings. Reducing the amount of instructional time originally called for by the unit can reduce students' depth of understanding (Clark & Linn, 2003). However, previous research on the effects of the amount of time that teachers allocate for particular classroom activities on student learning has produced mixed results. Allocated time is not always spent on learning activities. Consequently, some studies suggested that while allocating more time for particular activities may have a small positive effect for low ability students, there is no overall effect on what students learn (Cotton, 1989). We are interested in whether the quantity of time teachers' spent on the *Stuff* unit affected student learning.

Teachers have to continually seek a balance between ''covering'' the topics they feel are important and ensuring that students' experiences are sufficient to develop deep understanding (Van den Akker, 1998). Teachers sometimes scale back student investigations, or decide to omit particular activities or portions of activities in the unit because of a lack of time, resources, because they are unsure of how to enact the activity, or because they do not see an activity's value. Adaptations such as this might limit students' opportunities to engage in inquiry practices, such as asking questions and talking with classmates to solve problems, or affect the coherence of unit overall (Shwartz, Weizman, Fortus, Krajcik, & Reiser, 2008). Previous research has shown that teachers who frequently use inquiry-oriented teaching practices have a positive impact on students' science achievement (Kahle, Meece, & Scantlebury, 2000). Kahle and her coworkers found teachers who had their students solve problems with their peers, learn from their classmates, and repeat experiments to check results saw greater science achievement from their students. Consequently, we are interested in how the level of completion of a unit by a teacher influences student learning. Though this measure aligns with traditional measures of fidelity, we asked teachers how long they spent on each activity as an indication of how they alotted time to topics across the unit in comparison to the times suggested by the curriculum materials. In addition to considering how much of the curriculum they complete, it is also important to consider how they use the curriculum materials.

The tendency for teachers to transform innovative curriculum so that they resemble more traditional classroom practices suggests that how teachers choose to enact the unit might affect what students learn. The different ways that teachers manage classroom discourse have been called participation structures (Cazden, 1986) or activity structures (Fuson & Smith, 1998). These patterns of classroom discourse can vary in time scale and purpose, ranging from simple routines such as ''initiation-reply-evaluation'' (I-R-E) (Mehan, 1978, 1979) exchanges where students answer questions and receive immediate feedback to a sequence of project

milestones used to facilitate open-ended classroom inquiry (Polman, 2004). The tendency toward transmissive classroom routines despite accepted evidence for the need for students to take a more active role in their learning is well known (Bean, 2001). In other words, whole class teacher-centered instruction often dominates classroom practice. We are interested in the relationship between teacher adaptations of the activity structures, such as completing the activities as teacher-centered demonstrations versus student-centered investigations, on student learning.

When teachers try to implement innovations such as standards-based curriculum units, there are many challenges. Teacher support structures are necessary for teachers as they implement reforms and refine their understandings (Fullan, 1991). Our own efforts at supporting systemic reform acknowledge and support teachers adapting innovative curriculum materials as they address the needs of their students, time constraints, and limitations in resources (Blumenfeld et al., 2000). One way that designers can support the adaptation process is by providing teachers with feedback on the effect their adaptations have on student learning and to provide opportunities in subsequent professional development efforts to reflect upon their practice and discuss enactment issues with coworkers and designers (Pinto, 2005). To do this, we need ways of determining how teachers' curricular adaptations influence what their students learn. In this study, we ask the following research questions:

1. How do teachers' responses on a post-enactment survey align with their enactment of curriculum materials?
2. How do teachers' curricular adaptations (the amount of time on the unit, the level of completion of the unit, and the activity structures), teacher self-efficacy, and teacher experience enacting the unit influence student learning of target science learning goals?

Method

In order to address our research questions, we used data from the enactment of the *Stuff* unit during the 2003–2004 school year. In this section, we begin by describing the *Stuff* unit in more detail and the professional development the teachers received who enacted the curriculum. Then we discuss the participants and data sources that we used to address our research questions. Finally, we describe our procedure for analyzing the videotapes, test data, and teacher survey data.

Description of Stuff Unit

The IQWST curriculum units were designed to address the need for curriculum materials that support learning goals expressed in the national standards documents and to support classroom inquiry (Krajcik et al., 2008; McNeill et al., 2003). The units' activities engage students in inquiry activities with relevant phenomena and support teachers in facilitating discussions that allow students opportunities to understand how their experiences relate to the units' learning goals. Each unit also includes supports for inquiry practices such as using evidence to construct scientific explanations and creating representations or models of phenomena.

The *Stuff* unit introduces students to the concepts of characteristic properties, substances, chemical reactions, the conservation of mass, as well as how the particulate nature of matter explains these macroscopic phenomena (McNeill, Harris, Heitzman, Lizotte, & Sutherland, 2004). The unit consists of 16 lessons, some of which contain several different activities. Each lesson includes activities designed to engage students, including investigations, using models to explore concepts, and teacher-led class discussions. Some of the activities are identified as ''optional,'' in order to provide teachers guidance in their adaptations of the completion of the unit. We felt that if teachers did need to cut activities in the unit because of time limitations that the optional activities could be removed and the students would still have opportunities to adequately support their learning of each of the target learning goals. For example, Lesson 13, ''Does mass change in a chemical reaction?'', includes three activities. Activity 13A is an optional activity that has students investigate whether the mass changes when they create ''gloop.'' Activity 13.1 has students observing the reaction of Alka Seltzer in water in open and closed systems. Activity 13.2 has student redesign the

13.1 experiments so that mass will stay the same during the reaction. If all the ''optional'' activities are used, the unit is designed to take 33–35 school days, but if only the ''core'' activities are used, the unit should take only 26–28 school days.

The IQWST Approach to Professional Development

The IQWST approach to providing professional development has evolved over working with teachers for an extended time period on a variety of curriculum units. At the heart of these experiences are opportunities for discussion between teachers enacting the units and researchers. We call the conceptual framework that guides these activities (Krajcik, Blumenfeld, Marx, & Soloway, 1994): Collaborative construction of understanding; Enactment of new practices in classrooms; Reflection on practice; and Adaptation of materials and practices. The professional development activities for the Stuff unit include a 1-week summer institute and monthly Saturday workshops during teachers' enactment of the unit. Researchers used a design approach (Simon, 1996) to plan workshop activities based on feedback on teachers' enactments and student assessments (Fishman, Best, Foster, & Marx, 2000). Efforts to document how these professional development strategies influence teachers' enactments of the previous units and subsequent student learning have been described elsewhere (Fishman, Marx, Best, Stephen, & Tal, 2003; Kubitskey, 2006; Kubitskey, Fishman, & Marx, 2003; Margerum-Leys, Fishman, & Peek-Brown, 2004).

Participants

The 2003–2004 enactment of the *Stuff* unit included five school districts and 24 teachers from different areas of the country. We only included those teachers in the study from whom we received data from the required sources, student pre- and post-test data and the teacher curriculum survey. This limited our analysis to 19 teachers (see Table 1). Four of the teachers who we were unable to obtain complete data sets were located in other states. Because of the distance, we were unable to drive to the schools to obtain the requisite data and the teachers never mailed the data to us. The fifth teacher was located in the same state as the authors. This teacher changed schools at the end of the school year and because of communication difficulties we were unable to receive her curriculum survey.

Eight of the teachers were in public middle schools in a large urban area in the Midwest (Urban A). The majority of students in this school district were African American and come from lower to lower-middle income families. Three of the teachers taught in an independent school in a large college town in the Midwest (Town B). The majority of these students were Caucasian and from middle to upper-middle income families. Two of the teachers taught in a second large urban area in the Midwest (Urban C). The student population in this school district was 49.8% African American, 38% Hispanic, 8.8% Caucasian, and 3.2% Asian. Three of the teachers taught in a suburb of the second large urban area (Suburb D). The student population in this school district was ethnically diverse ($\sim$42% Caucasian, 44% African American, 10% Hispanic, and 4% Asian). Finally the last three teachers taught in a rural area in the south (Rural E). These schools had diverse populations each with a majority of African American students.[1]

Measures

To answer our research questions, we needed to examine both teachers' enactment as well as their responses to our survey. To determine how teachers' survey responses might relate to classroom practice, we examined a selection of videotaped lessons from a subset of the respondents. To determine how teachers'

Table 1
Participants from the 2003 to 2004 school year

Site	Urban A	Town B	Urban C	Suburb D	Rural E	Total
Schools	7	1	2	2	3	15
Teachers	8	3	2	3	3	19
Classrooms	30	5	4	13	13	65
Students	983	79	105	280	269	1,716

practices might influence student learning, we measured student learning using pre-/post-tests and related these results to their teachers' responses to a survey about their enactment. In this section, we describe our use of video to characterize teachers' use of curriculum materials, our conceptual model, and our measures of each of the variables included in our model.

Description of Video. To understand how teachers' survey responses corresponded to their actual classroom practice, we compared teacher responses with our own observations for a subset of lessons and teachers where video recordings were available. Due to the limited number of videotaped lessons available, our selection of teachers was neither representative of the larger group nor a random sample of the larger group. The four teachers we videotaped taught in three different schools in the Urban A school district. These four teachers were selected to be observed, because of their proximity to the researchers and their willingness to be videotaped. We reviewed their enactments of five *Stuff* activities to determine the duration, activity structure, and level of completion, and compared our observations with the teachers' survey responses. Table 2 summarizes the number of hours of videotape reviewed for the four teachers.

Our Conceptual Model. To investigate the influence of teachers' adaptations on students' learning during the *Stuff* unit, we compared measures of student learning with factors that may influence teachers' adaptations of the materials and the adaptation practices themselves. A conceptual model of our study including all of the measures that we investigated is shown in Figure 1. We describe each of these measures in more detail below.

Description of Pre-/Post-Test. To measure student learning for all teachers, the same test was administered to students before and after the *Stuff* unit. The test consisted of 15 multiple-choice items and 4 open-ended items for a total of 30 points. See Figure S1 for example test items. The test items were aligned with the unit's learning goals and learning tasks (Krajcik et al., 2008). All open-ended items were scored using specific rubrics created to address the particular inquiry practice and content area (see McNeill & Krajcik, 2007 for a description of rubrics and coding). One rater scored the students' open-ended responses. We then randomly sampled 20% of the tests, which were scored by a second independent rater. Our estimates of inter-rater

Table 2
Hours of video examined for each activity

Activity	Description	Teacher D (hours)	Teacher B (hours)	Teacher H (hours)	Teacher E (hours)
Lesson 8	Does acid rain make new substances?				
8.1	After reading about the discoloration of the Statue of Liberty, students see a demonstration of burning magnesium and use the properties of the reactants and products to explain whether a chemical reaction has occurred	1	2	2	2
8.2	Students study a model of the Statue of Liberty by investigating the effect of vinegar vapor on pennies	2	1	3	1
Lesson 10	Do I always make new substances?				
10.1	After hypothesizing whether or not the bubbles always indicate that a chemical reaction is occurring, students investigate whether boiling and condensing water is a chemical reaction	2	2	1	1
10.2	Students investigate whether creating a mixture such as ''Kool-Aid'' involves a chemical reaction	2	1	0	0
Lesson 12	How can I make soap from fat?				
12.1	Students return to materials they described in the first learning set to create soap from fat. After they create their cake of soap, they read about the history of soap making to discuss the following day	3	2	2	1

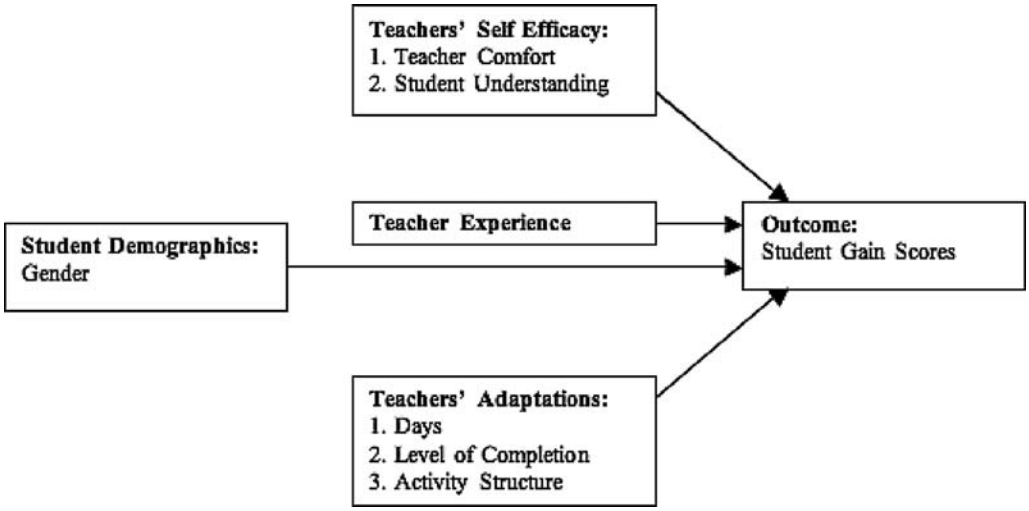

Figure 1. Conceptual model.

reliability were calculated by percent agreements. Our inter-rater agreement was above 96% for each of the four open-ended test items.

Only students who completed both the pre-test and post-test were included in our analysis, because we were interested in how students' science achievement changed over time. Due to the high absenteeism in the urban schools, only 1,234 students completed both the pre- and post-test. In order to examine whether the students who only completed the pretest or only completed the post-test were different than the students who completed both the pre- and post-test, we conducted a missing data analysis. We compared the students' pretest scores for those students who also completed the post-test to those students who did not complete the post-test for each of the 19 teachers. Sixteen of the 19 teachers did not have a significant difference for the two groups. For the three teachers that did have a significant difference (Teachers B, O, and P), their students who were not missing the post-test had significantly higher pretest scores than those students who were missing the post-test. We also compared the students' post-test scores for those students who completed the pretest to those students who did not complete the pretest for each of the 19 teachers. Fifteen of the 19 teachers did not have a significant difference for the two groups. For the four teachers who did have a significant difference (Teachers N, R, P, and Q), the students who were not missing the pretest had higher post-test scores than those students who were missing the pretest. This suggests that the students who were in school for both the pre- and the post-test were higher science achievers than those students who were absent on one of the test administration days for some of the teachers. Yet for the majority of the teachers those students who were absent one test day were not significantly different. Nonetheless, one limitation of this study is that we were unable to collect pre- and post-test data from all of the students.

In order to assess student learning over the unit, we used students' gain scores. We calculated the gain scores by subtracting the pretest score from the post-test score. We used this measure as the outcome for our model. On the test, students also indicated their gender, which we also included in the model. Unfortunately, our agreement with the schools did not allow us to collect other demography data from the students so were not able to include race or other measures in our study.

Description of Survey. To gauge how teachers assessed and adapted the *Stuff* unit, each teacher was asked to complete a survey after they finished their enactment. The survey consisted of 16 pages, one for each of the unit's lessons, which could include more than one activity (for a sample survey page, see Figure S2). Since we were interested in the teachers' appraisals of their efficacy using the unit, they were asked to indicate their comfort-level with each activity and their students' understanding of each activity. To get feedback on

their adaptation strategies, teachers were asked to indicate whether each activity was done by students or as a teacher demonstration, its level of completion, and how many days were spent on each lesson. We asked teachers about how they enacted each activity to determine whether students were provided with opportunities to engage in inquiry and experience phenomena as suggested by the *Stuff* materials. We asked about teachers' level of completion and the length of time they spent on each lesson to gauge how they allotted time across the unit. To determine each teacher's experience with the unit, we used our records of previous enactments.

To analyze the survey responses, we first converted each teacher's checkmarks on the survey form to numerical codes and transferred them to a cumulative table. Table 3 summarizes how numbers were assigned to the teachers' responses.

After tabulating teachers' responses, we reduced each teacher's responses to a single number for each of the six variables listed above. For the teachers' self-efficacy, we calculated two variables. We averaged their responses for their own comfort level and their responses for students' understanding for each activity across the entire unit. Each teacher's experience with the unit was coded as either the first or second use of the materials. For teacher adaptations, we calculated three variables: activity structure, level of completion and days spent on curriculum. We view level of completion as a measure of fidelity, because we were specifically interested in alignment with the intended curriculum. We view activity structure and days spent on the curriculum as other adaptations. These two measures do not specifically measure fidelity, because we did not compare the teachers' reports to the recommendations in the curriculum. Rather, we were interested in more generally how teachers' decisions around time and activity structure (regardless of the intent of the curriculum) impacted student learning. First, in order to summarize the activity structures teachers used during the unit, we averaged their scores across all of the activities in the unit. For their level of activity completion, we totaled their scores across the unit and divided this total by the number of "core" or not optional activities so that teachers who enacted the core activities along with one or more optional activities would have a score greater than one. The total number of days each teacher allocated to the unit was found by adding the days he or she indicated were spent on each lesson.

Analytic Method

We analyzed both the videos teachers' enactments to determine how the their practice related to their reports of their practice on our survey as well as how their survey results related to their students' achievement. Each of these steps is described below.

Enactment Analysis. To answer how teachers' survey responses represented how they enacted the *Stuff* materials, a small sample of videotaped lessons were reviewed. Teachers' survey responses for activity structure and level of completion were compared with their videotaped enactments for the four teachers where videotape data were available.

HLM Analysis. Determining the impact of teacher adaptations on student learning is a complex issue. Because each teacher's efforts affect each of his or her students, learning by individual students in the same class is not independent. On the other hand, considering the class mean as the outcome variable loses the

Table 3
Numerical assignments for teachers' survey responses

Categories	Variables	Numerical assignment
Self-efficacy	Teacher comfort level	1, low; 2, medium; 3, high
Self-efficacy	Student understanding	1, low; 2, medium; 3, high
Experience	Experience	0, first use of unit; 1, second use of unit
Teacher adaptation	Activity structure	1, teacher demo; 2, student investigation; 3, both
Teacher adaptation	Level completion	0, not used; 0.5, partially completed; 1, completed
Teacher adaptation	Days spent on lesson	Total number of days spent teaching the unit

individual variability of student learning. Neither approach would allow us to disentangle individual and group effects on student learning. In our analysis of the survey and test data, we needed to consider this grouping or nesting of students and any differential effects across teachers. Multi-level modeling recognizes the dependence and grouping of data leading to more correct estimation of effects and variance. We used Hierarchical Linear Modeling (HLM) in a two-level format to investigate the effect of factors that affect teachers' adaptations and teachers' adaptation strategies on student learning (Raudenbush & Bryk, 2001). Our use of HLM consisted of three steps. First, we created a fully unconditional model (FUM), then we created a level 1 or within-teacher model to examine the effect of student level variables, and finally we created a level 2 or between-teacher model to examine the effect of teacher level variables.

Fully Unconditional Model. HLM analysis begins with a fully unconditional model, which consists only of the outcome variable and no independent variables. The fully unconditional model provides the results of partitioning the outcome variance into within-group (σ_2) and between-group (τ_{00}) components, testing whether the between group component is significantly different from zero. In our model we used student gain scores, to determine whether there were differences in student learning across the 19 teachers. From these measures we computed the intraclass correlation coefficient (ICC), ρ, which is the proportion of variation in the student gain scores that is due to differences between teachers.

Within-Teacher Model. Next, we investigated the student-level measures that could account for the variation within teachers. We entered gender as a fixed effect. This meant that the effect of gender did not vary depending on what teacher a student had. The following is the equation for our level-1 model:

$$\text{Gain Score}_{ij} = \beta_{0j} + \beta_{1j}(\text{Gender}_{ij} - \text{Gender}\ldots) + r_{ij}$$

In this equation, β_{0j} represents the intercept or the gain score when all other variables equal zero, β_{1j} represents the effect of gender on student gain scores and, and r_{ij} represents the error term. After running the within-teacher model, we determined how much of the total unexplained individual-level variance for student gain scores was explained by the addition of our level-1 variable.

Between-Teacher Model. Lastly, we ran a between-teacher model. This allowed us to model student learning with our teacher-level measures to explain the between-teacher variation in our outcome variable. More specifically, we determined if student learning was impacted by teacher self-efficacy, experience, and curricular adaptations. We tested the six teacher level variables that we described above: teacher comfort level, teacher evaluation of student understanding, teacher experience enacting the unit, the number of days allocated to the unit, the level of completion of the unit's activities, and the teachers' activity structure (i.e., whole-class demonstration vs. student investigation). We removed any variables that were not significant. We report the final model below in the results section. As with the within-teacher model, we can determine how much of the total unexplained individual-level and teacher-level variance of our outcome has been explained by the addition of our level-2 variables.

Results

In this section, we begin by presenting the descriptive statistics for the results of the teacher survey and the results from the students' pre- and post-test to provide an overview of the data. Then we present the results from comparing the teacher survey data with the video analysis. Finally, we present the results from the HLM model.

Descriptive Statistics

Before creating our HLM model, we first examined whether there were differences in student learning and teacher adaptations across the 19 teachers. Table 4 displays the descriptive statistics for all of the variables included in our study.

Fifty percent of the students in the sample are male and 50% are female. On average, students gained 7.49 points from the pre- to post-test though the gain scores ranged from −13.36 to 22.80.

Table 4
Descriptive statistics (n = 1,234)

	Mean % (standard deviation)
Student variables	
Gender[a]	50.00
Test gain score	7.49 (5.23)
Teacher variables	
Self-efficacy—teacher comfort level	2.55 (0.34)
Self-efficacy—student understanding	2.39 (0.44)
Experience[b]	27.00
Teacher adaptation—days	31.17 (6.97)
Teacher adaptation—activity structure	1.93 (0.14)
Teacher adaptation—level completion	0.94 (0.16)

[a]Percentage of female compared to males.

[b]Percentage of teachers who have done the unit before compared to those who have not.

For the teacher variables, we see a range of scores for both the teacher adaptation variables and the efficacy variables. Teachers' average comfort level was a 2.55, which is between medium and high. Teachers' perception of student understanding was 2.39, which is also between medium and high. Twenty-seven percent of the teachers previously enacted the unit. On average, teachers spent 31.17 days on the unit. For teachers' activity structure the average score was 1.93. Remember a score of 1 means that a teacher completed all activities as a demonstration, a score of 2 means that students completed all activities, and a score of 3 means that all activities were both demonstrated by the teacher and completed by the students. This suggests that for most lessons teachers had students complete the activities, but some were on average only completed as demonstrations. The average level of completion was 0.94 suggesting that typically teachers were completing a little less than the recommended core activities within the unit. Remember that we coded teachers' completion as 0 for not using the activity, 0.5 for partially completing the activity, and 1 for fully completion of the activity.

Student Assessment Data

Since we are interested in whether there is differential learning by teacher, we examined the effect size of student learning by teacher.[2] Figure 2 shows the effect sizes for the 19 teachers.

Across the 19 teachers, there is a wide range of effect sizes from 0.47 to 5.27. We tested whether there was a significant teacher effect by performing an ANCOVA on students' post-test scores with the pretest scores as the covariate. There was a significant teacher effect with the learning gains of some teachers being greater than other teachers, $F(18, 1215) = 9.062$, $p < 0.001$. There was also a significant interaction between the teacher and students' pretest scores, $F(18, 1215) = 2.868$, $p < 0.001$, suggesting that the effect of a students' pretest on their post-test varied by teacher.

This analysis suggests that something is occurring in each of these classrooms that is influencing student learning. These differences could be caused by a variety of factors such as the school culture, parental influence, or different resources. We also believe that the differences in teachers' enactments are influencing student learning based on prior research on teacher practices (Kahle et al., 2000). Our hypothesis is that some of this difference in student learning is the result of teacher adaptations, experience using the materials, and efficacy.

Enactment Analysis

In order to evaluate the validity of the self-report survey data, teachers' adaptations of lessons 8, 10, and 12 were reviewed from videotapes. The results, sorted by each teacher's effect size, are summarized in Table 5. For each activity, an objective measure of the degree that the activity was completed was computed by dividing the number of activity elements observed (AES_{OB}) divided by the total number called for in the

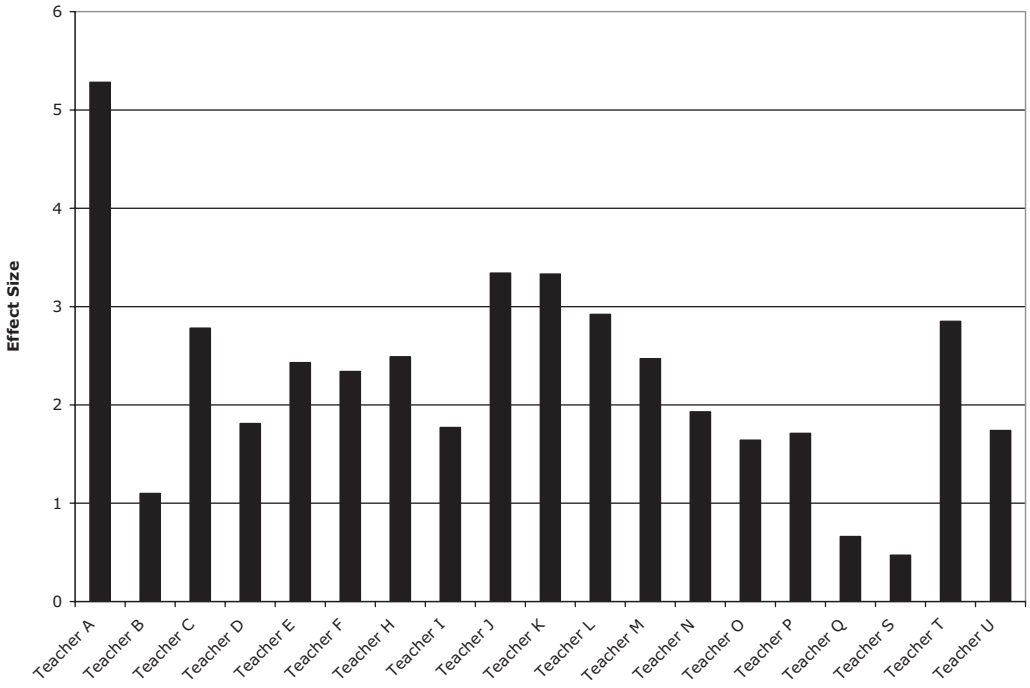

Figure 2. Effect size by teacher.

Stuff teacher's guide (AEs_{TG}). The teachers' activity structure for each activity (Demonstration, Student Investigation, or Both) observed in the videotapes was also recorded, along with teachers' survey responses for how they adapted each activity.

To determine how teachers' responses on the survey might represent their enactment of the *Stuff* curriculum materials, we compared the fraction of AEs observed to the level of completion indicated by each teacher on the survey. To compare the activity structures observed and the activity structures indicated by teachers on the survey, we averaged the values for each teacher across the reviewed activities. Table 5 illustrates that teachers reported their activity structure in different ways. For example, Teach B reported a combination of 2s and 3s, while Teacher E reported a combination of 1s and 2s. We used these averages for the observed enactment to characterize each teacher's adaptation preferences, or teaching style across the lessons we reviewed. Figure 3 compares four teachers' activity structures observed in videotapes of selected activities with their survey responses about how they enacted the same lessons.

These comparisons are arranged in order of effect sizes, with the leftmost teacher having a lowest effect size (1.10), the middle two teachers having similar effect sizes (1.77, 1.81), and the rightmost teacher having a higher effect size (2.43). The graph shows that the survey responses belonging to the three teachers with the highest effect sizes were very similar to the assessments of their enactments from the videotapes. These teachers' average activity structure scores were between 1.75 and 2.00, suggesting they generally provided students opportunities to conduct the investigations called for in the unit. Though the fourth teacher's survey suggested that students also had these opportunities, examination of the videotaped activities indicated that the teacher relied more on demonstration and direct instruction to address the unit's topics. This suggests that the majority of the teachers' self report did align with their enactment, though for one teacher there was clearly a difference.

Figure 4 compares the proportion of the activity elements observed in the video compared to the proportion of the activity the teacher reported completing on the survey. As in the previous graph, the comparisons are arranged from lowest to highest effect size. All four teachers overestimated their completion

Table 5
Observed activity elements

Observed activity elements (AE)	Teacher B	Teacher H	Teacher D	Teacher E
Teacher effect sizes	1.10	1.77	1.81	2.43
Lesson 8: Does acid rain make new substances?				
Activity 8.1: Students observe a demonstration of burning magnesium and use the properties of the reactants and products to explain whether a chemical reaction has occurred				
Level of completion from video (AE_{OB}/AE_{TG})[a]	0.6	0.7	0.7	0.8
Level of completion from survey (1 = fully completed, 0.5 = partially used, 0 = not used)	1	1	1	1
Activity structure from video (1 = teacher demo, 2 = student investigation, 3 = both)	1	2	2	1
Activity structure from survey (1 = teacher demo, 2 = student investigation, 3 = both)	3	3	2	2
Activity 8.2: Students study acid rain by investigating the effect of vinegar vapor on pennies				
Level of completion from video (AE_{OB}/AE_{TG})	0.3	0.6	0.6	0.4
Level of completion from survey (1 = fully completed, 0.5 = partially used, 0 = not used)	0.5	1	1	1
Activity structure from video (1 = teacher demo, 2 = student investigation, 3 = both)	1	3	2	3
Activity structure from survey (1 = teacher demo, 2 = student investigation, 3 = both)	2	3	2	2
Lesson 10: Do I always make new substances?				
Activity 10.1: Students investigate whether boiling and condensing water is a chemical reaction				
Level of completion from video (AE_{OB}/AE_{TG})	0.3	0.4	0.7	0.6
Level of completion from survey (1 = fully completed, 0.5 = partially used, 0 = not used)	0.5	0.5	1	0.5
Activity structure from video (1 = teacher demo, 2 = student investigation, 3 = both)	1	1	2	1
Activity structure from survey (1 = teacher demo, 2 = student investigation, 3 = both)	3	1	2	1
Lesson 12: How can I make soap from fat?				
Activity 12.1: Students return to materials they described in the first learning set to create soap from fat				
Level of completion from video (AE_{OB}/AE_{TG})	0.8	0.6	0.8	0.6
Level of completion from survey (1 = fully completed, 0.5 = partially used, 0 = not used)	1	1	1	1
Activity structure from video (1 = teacher demo, 2 = student investigation, 3 = both)	3	2	3	2
Activity structure from survey (1 = teacher demo, 2 = student investigation, 3 = both)	3	2	2	2

[a]AEs_{OB} is the number of activity elements observed. AEs_{TG} is the total number called for in the teacher guide.

of the unit. Yet teachers who completed more of the unit did report completing more (Teachers D and E), while those teachers completing less of the unit also reported completing less (Teacher B and H).

HLM Analysis

Fully Unconditional Model. We began our HLM analysis by examining the fully unconditional model, which partitions the total variance in students' gain scores into its within- and between-teacher components. Table 6 provides the results from the unconditional model.

Lambda is the pooled reliability estimate across all the teachers for estimating our outcome variable, student gain scores. Since the reliability estimate is high, 0.967, we are comfortable using the adjusted intraclass correlation (ICC). The adjusted ICC tells us that 38% of the variance in student gain scores lies between teachers. There was a significant difference in student gains between teachers, $\chi^2 = 693.85$, df = 18, $p < 0.001$. This means that 38% of the variance in student learning can be attributed to the role of the teacher. Because the results of the fully unconditional model were significant, this supports our decision to use multilevel methods.

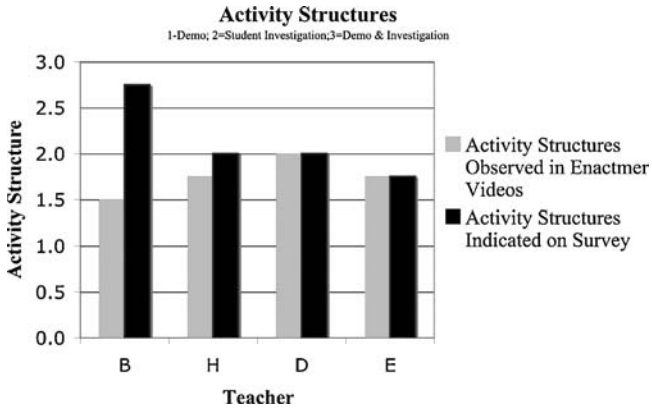

Figure 3. Activity structures.

Within-Teacher Model. The within-teacher model explores whether student gender is associated with student learning. We included gender as fixed effect, which means that the effect of gender did not vary depending on the teacher. Table 7 provides the results from our within-teacher model.

A student's gender does significantly influence their gain scores. On average, a female's gain score increases 0.104 standard deviations more than a male. Although adding gender does significantly influence student learning, it explains a very small percentage of the individual-level variance in student learning, less than 1%.[3] Unfortunately, we do not have access to other student level variables to include in the model. The intercept variance at the bottom of Table 7 suggests that there is still significant between-teacher variability. This provides support that there are contextual factors or characteristics of the teachers that influence student learning. In order to further unpack the role of teacher characteristics, we need to add level 2 predictors to our HLM model.

Between-Teacher Model. Table 8 presents the results from our complete HLM model including Level 1, student level predictors, and Level 2, teacher level predictors. Although we tested numerous teacher level characteristics in our model, we only kept in the model those measurements that were significant. The relatively small number of teachers in the study limited our model. As a general rule, you need 10 cases at a level (either level 1 or level 2) for each significant variable included in a model (Raudenbush & Bryk, 2001). Since we only had 19 teachers in our study to include in the level 2 model, it was not surprising that we ended

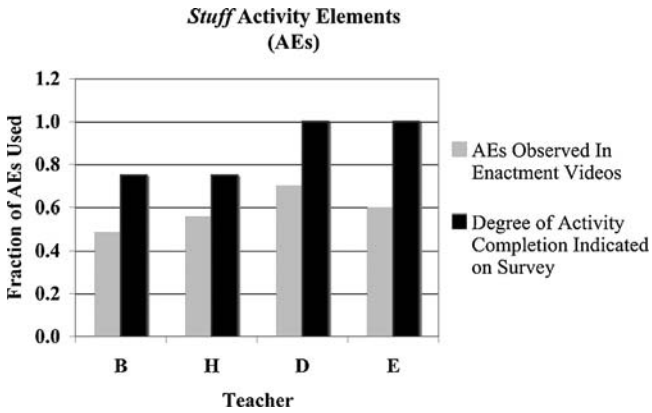

Figure 4. Level of completion.

Table 6
Unconditional model of student learning (n = 1,234 students, N = 19 teachers)

	Student gain scores
Tau (τ)	0.384
Sigma-squared (σ^2)	0.646
Lambda-reliability (λ)	0.967
Intraclass correlation (ICC)[a]	0.373
Adjusted-ICC[b]	0.380

[a]ICC = $\tau/(\tau + \sigma^2)$.
[b]Adjusted ICC = $\tau/(\tau + (\lambda\sigma^2))$.

Table 7
Within-teacher model of student gain scores (n = 1,234 students, N = 19 teachers)

	Student gain scores
Random effects	
Intercept (β_0)	$-0.012^{\sim}$
Fixed effects	
Gender[a]	0.104^{*}
Variance components for random effects	
Intercept variance (β_0)	0.383^{***}

$^{\sim}p < 0.1$.
$^{*}p < 0.05$.
$^{***}p < 0.001$.
[a]Females compared to males.

up with a model that included only two significant teacher practices. In our testing of the various models, we found two models that each included two significant variables. One model included teacher experience and level of completion and the second model included teacher experience and activity structure. Since the second model including teacher experience and activity structure had lower significant levels, we used it as our final model. We hypothesize that if we had a larger sample of teachers, all three variables, teacher experience, level of completion and activity structure, would significantly influence student learning. The following is our

Table 8
Between-teacher model of student learning (n = 1,234 students, N = 19 teachers)

	Student gain scores
Random effects	
Intercept (β_0)	
Base	−0.234
Activity structure	$1.869^{\sim}$
Experience with unit	0.715^{**}
Fixed effects	
Gender[a]	0.105^{*}
Variance components for random effects	
Intercept variance (β_0)	0.258^{***}

$^{\sim}p < 0.1$.
$^{*}p < 0.05$.
$^{**}p < 0.01$.
$^{***}p < 0.001$.
[a]Females compared to males.

equation for the level-2 model for student gain scores:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Activity structure}) + \gamma_{02}(\text{Teacher Experience}) + \mu_{0j}$$

In this equation, γ_{00} represents the intercept, γ_{01} represents the effect of activity structure, γ_{02} represents the effect of teacher experience, and μ_{0j} represents the error term. Teachers' activity structures (i.e., whole-class demonstration versus student investigation) and level of experience were used to model the intercept. None of the other teacher-level measures, days spent on the unit, student understanding, teacher comfort level or level of completion, were significant. Consequently, we removed them from our final level-2 models.

The first set of results under intercept in Table 8 is for our model in terms of the intercept as the outcome. These results tell us whether any of the teacher characteristics influence student learning. Teachers' activity structures (i.e., demo vs. student investigation) have a marginally significant effect and a teacher's experience with the unit has a significant effect on student learning. Holding all other variables constant, as a teacher's activity structure increases by 1 point (i.e., goes from all lessons completed as demonstrations to all lessons completed by students), students gain scores increase by 1.869 standard deviations.

This is a very large increase in students' gain scores and suggests that having students actively complete the activities is important for students understanding of the key learning goals. On average, a teacher with experience teaching the unit has student gain scores of 0.715 standard deviations higher than a teacher who is completing the unit for the first time. This suggests the importance of having experience with reform based curriculum units. Consequently, the results of the HLM model suggest that (1) Teachers who had previously taught the inquiry-oriented curriculum had greater student gains and (2) Students who completed investigations themselves had greater learning gains compared to students in classrooms who observed their teacher completing the investigations as demonstrations.

Neither the number of days spent on the unit, teacher comfort level, teachers' report of their students understanding, nor level of completion significantly influenced student learning. As we mentioned before, since our data includes only 19 teachers we would expect to only have at most two significant variables in our model. Our model is not powerful enough to detect significant effects of more variables. Other variables, particularly level of completion, which was significant by itself or in combination with teacher experience, could be important predictors of student learning if we had a more powerful model. Our model does not suggest that these other variables are not important; rather it provides support that both teacher experience and activity structure are particularly important for student learning.

For average student learning between teachers, our model explains 33% of the variance between teachers.[4] This suggests that 33% of impact of the role of the teacher can be attributed to activity structure and the role of the teacher. By including only two variables in our model about teacher adaptations and experience, we explained a considerable percentage of the between teacher variation. Furthermore, we obtained the measure of teacher adaptation through a simple teacher survey of how they enacted the curriculum. Yet the variance component at the bottom of Table 8 shows that the between teacher variances is still significant. This means that we have not explained away all of the between teacher variance for student learning.

Discussion

Despite the numerous limitations of existing science curriculum (Kesidou & Roseman, 2002), teachers often rely heavily on them in their science instruction. Reform oriented science curricula provide a potential avenue for changing classroom practice. Teachers often need to adapt these materials to meet their students' needs as well as their own teaching style (Davis & Varma, 2008) and these adaptations can have either a positive or negative effect on what students experience in the classroom. Consequently, in this research we were interested in how middle school science teachers' adaptations of science curriculum impacted student learning.

Alignment of Teachers' Survey Responses With Enactment

In addition to modeling the effect of teachers' adaptations on what students learned during the *Stuff* unit, we were also interested in how teachers' responses to a survey about their enactments compared to our observations of videotapes of their lessons. We examined two teacher adaptations that were addressed in the survey, activity structure and level of lesson completion, and found mixed results.

In the past, teacher self-reports of curriculum enactment have been questioned (Snyder, Bolin, & Zumwalt, 1992). Mayer (1999) conducted a study with math teachers in which he found that survey data was a reliable and valid method to measure whether or not teachers were engaging in reform based instructional practices, but that the surveys were less effective in measuring either the quality of the instructional practice or the amount of time teachers spent using one practice or another. Our results comparing the video enactments to teacher surveys are similar to this previous work. We found that our survey had some validity in measuring the type of activity structure that occurred in the classroom. Most of the teachers' identification aligned with our own interpretation of whether the activity was a teacher demonstration, student led activity or both teacher and student directed. In terms of level of completion of the activities, teachers responded that they completed more of the activities than our interpretation from analyzing the video indicated. Level of completion is actually a more abstract construct than activity structure. The teachers may have felt that they completed the entire activity if they had students conduct the investigation, while we were also looking for the teacher to engage students in discussion before and after the investigation. This may be one reason why the teachers were more likely to over estimate their completion. This suggests that surveys may be more appropriate for asking teachers concrete questions about curriculum enactment, like activity structure and number of days, while more abstract ideas, such as level of completion and quality of teacher instructional practices, may be more difficult to validly measure through surveys. Consequently, the survey item measuring level of completion may not have been as accurate a representation of what occurred in the teachers' classrooms. If this item had been more reliable, it could have influenced the results from the HLM analysis.

Effect of Teachers' Curricular Adaptations on Student Learning

Across all of the teachers, the use of the inquiry-oriented middle school science curriculum resulted in considerable student learning, with effect sizes ranging from 0.47 to 5.27. Yet this incredible variation in effect size suggests that it is not just the quality of the curriculum that is important, but also the way the curriculum is used by teachers in the science classroom. Thirty-eight percent of the variation in students' gain scores occurred between teachers, suggesting the role of the teacher is incredibly important. Understanding what factors impacted this difference can offer important insights for future curriculum development and associated professional development. We examined two teacher characteristics and four ways that they could adapt the materials during their enactment to look for their effect on student learning. We found that one teacher characteristic, teachers' experience with the materials, and one adaptation, teachers' choice of activity structure, were found to be significant.

In terms of activity structure, we found that students who completed the activities themselves had greater student gains than students in classrooms where the teacher completed the activities as demonstrations. The recent National Research Council publication *Taking Science To School* (Duschl, Schweingruber, & Shouse, 2007) argues for the importance of actively engaging students in science practice where "...students carry out investigations and talk and write their observations of phenomena, their emerging understanding of scientific ideas, and ways to test them" (p. 6). This image of science instruction aligns with the scientific inquiry approach advocated by the National Science Education Standards (1996), which argues that "Students at all grade levels and in every domain of science should have the opportunity to use scientific inquiry and develop the ability to think and act in ways associated with inquiry, including asking questions, planning and conducting investigations, using appropriate tools and techniques to gather data, thinking critically and logically about relationships between evidence and explanations, constructing and analyzing alternative explanations, and communicating scientific arguments." (p. 105). Although this importance of actively engaging students in science is prevalent throughout the literature, there is little empirical support that shows that having students complete investigations themselves results in greater student learning than students observing their teachers conduct the same experiment. The results from this study suggest that having the students conduct the activities and investigations themselves is a key factor in determining the successful implementation of the inquiry-oriented curriculum. This is an important finding to let teachers know how this adaptation (to change student-directed activities to become teacher-directed activities) can have a negative impact on student learning.

We also found that teacher experience with the curriculum materials significantly impacted student learning. Teachers who had previously enacted the reform based curriculum had larger student test gains than teachers who were using the curriculum for the first time. Our finding is consistent with previous work documenting the importance of teacher experience in enacting reform-based curriculum (Geier, 2005). This is important to keep in mind particularly because the implementation of science inquiry in classrooms can present significant challenges for teachers such as having sufficient science background knowledge and managing extended activities (Edelson, Gordin, & Pea, 1999). The first time teachers enact an inquiry-oriented curriculum they may become frustrated using the materials. By letting teachers know that usually teachers are more effective using the curriculum a second and third time, may encourage teachers to try using the innovative curriculum for more than 1 year. This is also an important finding in terms of evaluating the effectiveness of a new curriculum or other instructional tool. The first year teachers enact a reform the students' learning gains may under-represent the potential of the curriculum. Teachers may need to enact a curriculum multiple times before they are able to effectively use it in their own classroom.

There are limitations to our study. First, we were unable to obtain complete data sets from all of the teachers that participated and consequently this lowered our sample size by five teachers. Including these other teachers may have impacted the results of our analysis. Furthermore, the relatively low number of teachers in our study for HLM limited the power of our model. We would expect other measures to be strong predictors of student learning, but their effects were not significant here. Specifically, we would expect that the level of completion of the unit and measures of teacher efficacy to influence student learning. Our model does not suggest that these measures are unimportant; rather it just suggests that teacher experience with the curriculum materials and activity structure have a greater impact on student learning. In terms of level of completion, as we mentioned earlier, there were limitations in our use of survey data to measure the level of completion and observational data may have provided a more accurate measure of this construct. In terms of self-efficacy, we define self-efficacy as a teacher's belief in his or her ability to act in ways that successfully accomplish specific teaching goals (Tschannen-Moran et al., 1998). The teachers in our study were all using the inquiry curriculum for either the first or second time. Our measure of self-efficacy, specifically asked them about their comfort level around the curriculum. Consequently, it is not surprising that teachers' experience with the curriculum materials had a significant impact on student learning while teachers' self-efficacy did not. Since our study only looked at teachers during their first or second enactments of the materials, it may be too soon for a reliable sense of efficacy to take shape. In order to more effectively examine the role of self-efficacy, future studies should include teachers with a wider range of experiences with curriculum as well as include a more diverse measure of self-efficacy beyond curriculum comfort level.

Furthermore, as we discussed previously there are limitations in using survey data to measure teacher enactment. The survey did not provide a variety of details about the enactments such as a measure of the quality of instructional practices being used in the classroom and teachers struggled with estimating their level of completion of the curriculum. Observational or videotape analyses for all of the teachers in the study would provide a more detailed measure of teachers' enactments and instructional practices. Future research needs to continue to explore what other characteristics in teachers' enactment cause the significant variation in student learning between teachers.

Yet the importance of both activity structure and teacher experience are essential fundamental findings from this study that should be kept in mind during future curriculum development and professional development support of teachers. Actively engaging students in science investigation is important for students to successfully learn key science concepts. Furthermore, teachers need experience using reform oriented curriculum in order to reap the benefits in terms of greater student learning.

Notes

[1]For this study we were not able to directly collect ethnicity data from the students. Instead, we used the information available online for the school or district in terms of self report of student ethnicity data. Unfortunately, at the time (2003–2004) the data were reported in different ways by the respective schools and districts. We included the most specific data that we had available to us, which did not include the specific percentages for Urban A, Town B, and Rural C.

[2]Effect Size: Calculated by dividing the difference between post-test and pre-test mean scores by the pre-test standard deviation.

[3]From the Fully Unconditional Model, we found that the amount of variance at the individual level was 0.64575. After taking into account our predictor variables in our within-teacher model, the within teacher variance is 0.64359. Therefore, the proportion of the individual-level variance that has been explained by our individual-level predictors is (0.64575 − 0.64359)/0.64575, which equals 0.0033. This means that our within teacher model explains 0.33% of the variance in student learning.

[4]To calculate the proportion of the between-level variance that we explained in our model we used the following equation: $(\tau_{\text{withinmodel}} - \tau_{\text{betweenmodel}})/\tau_{\text{withinmodel}}$. In this case (0.38257 − 0.25814)/0.38257 = 0.3252.

This research was conducted as part of the Investigating and Questioning our World through Science and Technology (IQWST) project and the Center for Curriculum Materials in Science (CCMS), supported in part by the National Science Foundation grants ESI 0101780 and ESI 0227557 respectively. Any opinions expressed in this work are those of the authors and do not necessarily represent either those of the funding agency, the University of Rhode Island, Boston College or the University of Michigan. We would like to thank all of the researchers involved with IQWST and CCMS, especially Brian Reiser, and Jonathan Singer.

References

American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. USA: Oxford University Press.

Bean, J.A. (2001). Teaching in middle schools. In V. Richardson (Ed.), Handbook of research on teaching (4th ed., pp. 432–463). Washington: AERA.

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

Bransford, J. (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Brown, M., & Edelson, D.C. (2001). Teaching by design: Curriculum design as a lens on instructional practice. In Annual meeting of the American Educational Research Association, Seattle, WA. Presented at the Annual meeting of the American Educational Research Association, Seattle, WA.

Cazden, C.B. (1986). Classroom discourse. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd ed., pp. 432–463). New York: Macmillan.

Clark, D., & Linn, M.C. (2003). Designing for knowledge integration: The impact of instructional time. Journal of the Learning Sciences, 12(4), 451. DOI: 10.1207/S15327809JLS1204_1.

Cohen, D.K., & Ball, D.L. (1999). Instruction, capacity, and improvement (CPRE Research Report No. RR-043) (No. RR-43). CPRE Research Report Series. Philadelphia, PA: Consortium for Policy Research in Education.

Cotton, K., (1989). School improvement research series: Educational time factors. Northwest Regional Educational Laboratory. Retrieved August 12, 2008, from http://www.nwrel.org/scpd/sirs/4/cu8.html.

Davis, E.A., & Varma, K. (2008). Supporting teachers in productive adaptation. In Y. Kali, M.C. Linn, & J.E. Roseman (Eds.), Designing coherent science education: Implications for curriculum, instruction, and policy (pp. 94–122). New York: Teachers College, Columbia University.

Duschl, R.A., Schweingruber, H.A., & Shouse, A.W. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=11625&page=213.

Edelson, D.C., Gordin, D.N., & Pea, R.D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. The Journal of the Learning Sciences, 8(3&4), 391–450.

Fishman, B.J., Best, S., Foster, J., & Marx, R.W. (2000). Fostering teacher learning in systemic reform: A design proposal for developing professional development. Presented at the Annual meeting of the National Association of Research on Science Teaching.

Fishman, B.J., Best, S., Marx, R.W., & Tal, R. (2001). Fostering teacher learning in systemic reform: Linking professional development to teacher and student learning. Presented at the Annual Meeting of the National Association of Research on Science Teaching, St. Louis, MO.

Fishman, B.J., Marx, R.W., Best, S., & Tal, R. (2003). Linking teacher and student learning to improve professional development in systemic reform. Teaching and Teacher Education, 19(6), 643–658.

Fullan, M. (1991). The new meaning of educational change (2nd ed.). New York: Teachers College Press.

Fuson, K., & Smith, S. (1998). The chalkboard activity structure as a facilitator of helping, understanding, discussing, and reflecting. Annual Meeting of the American Educational Researchers Association presented at the Annual Meeting of the American Educational Researchers Association, San Diego, CA.

Geier, R.R. (2005). Student achievement outcomes in a scaling urban standards-based science reform. Ann Arbor, MI: University of Michigan.

Haberman, M. (1991). The pedagogy of poverty. Phi Delta Kappan, 73(4), 290–294.

Kahle, J.B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? Journal of Research in Science Teaching, 37(9), 1019–1041.

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), 522–549.

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

Krajcik, J.S., McNeill, K.L., & Reiser, B.L. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. Science Education, 92(1), 1–32.

Kubitskey, M.E. (2006). Extended professional development for systemic curriculum reform. Ann Arbor, MI: University of Michigan.

Kubitskey, M.E., Fishman, B.J., & Marx, R.W. (2003). The Relationship Between Professional Development and Student Learning: Exploring the Link through Design Research. Presented at the Annual meeting of the American Education Research Association, New Orleans.

Margerum-Leys, J., Fishman, B.J., & Peek-Brown, D. (2004). Lab partners: Research University and Urban District Join Forces to Promote Standards-Based Student Learning in Science. Journal of Staff Development, 25(4), 5.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? Educational Evaluation and Policy Analysis, 21(1), 29–45.

McNeill, K.L., Harris, C.J., Heitzman, M., Lizotte, D.J., & Sutherland, L.M. (2004). How can I make new stuff from old stuff? In J.S. Krajcik, B.J. Reiser, & J.S. Krajcik (Eds.), IQWST: Investigating and questioning our world through science and technology. Ann Arbor, MI: University of Michigan.

McNeill, K.L., & Krajcik, J.S. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett, & P. Shah (Eds.), Thinking with data (pp. 233–266). New York: Routledge.

McNeill, K.L., Lizotte, D.J., Harris, C.J., Scott, L.A., Krajcik, J.S., & Marx, R.W. (2003). Using backward design to create standards-based middle-school inquiry-oriented chemistry curriculum and assessment materials. Presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Mehan, H. (1978). Structuring school structure. Harvard Educational Review, 48(1), 32–64.

Mehan, H. (1979). Learning lessons. Massachusetts: Harvard University Press Cambridge.

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

National Research Council (NRC). (1996). National science education standards. Washington, DC: National Academies Press.

O'Donnell, C.L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. Review of Educational Research, 78(1), 33.

Pinto, R. (2005). Introducing curriculum innovations in science: Identifying teachers' transformations and the design of related teacher education. Science Education, 89(1), 1–12.

Polman, J.L. (2004). Dialogic activity structures for project-based learning environments. Cognition and Instruction, 22(4), 431–466.

Powell, J.C., & Anderson, R.D. (2002). Changing teachers' practice: Curriculum materials and science education reform in the USA. Studies in Science Education, 37, 107–135. DOI: Article.

Putnam, R.T., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B.J. Biddle, T. Good, & I. Goodson (Eds.), International handbook of teachers and teaching. (1st ed). Norwell, MA: Kluwer Academic Publishers.

Raudenbush, S.W., & Bryk, A.S. (2001). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage. Publications, Inc.

Remillard, J.T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. Review of Educational Research, 75(2), 211.

Roseman, J.E., Linn, M.C., & Koppal, M. (2008). Characterizing curriculum coherence. In Y. Kali, M.C. Linn, & J.E. Roseman (Eds.), Designing coherent science education: Implications for curriculum, instruction, and policy (pp. 13–38). New York: Teachers College Press.

Schneider, R.M., Krajcik, J.S., & Blumenfeld, P.C. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. Journal of Research in Science Teaching, 42(3), 283–312. DOI: 10.1002/tea.20055.

Shwartz, Y., Weizman, A., Fortus, D., Krajcik, J., & Reiser, B. (2008). The IQWST experience: Using coherence as a design principle for a middle school science curriculum. The Elementary School Journal, 109(2), 199–219.

Simon, H.A. (1996). The sciences of the artificial (3rd ed.). The MIT Press.

Snyder, J., Bolin, F., & Zumwalt, K. (1992). Curriculum implementation. In P.W. Jackson (Ed.), Handbook of research on curriculum (pp. 402–435). New York: Macmillan.

Songer, N.B., & Gotwals, A.W. (2005). Fidelity of implementation in three sequential curricular units. In Unpublished manuscript. Presented at the Annual Meeting of the American Educational Researchers Association, Montreal, CA.

Songer, N.B., Lee, H.S., & Kam, R. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? Journal of Research in Science Teaching, 39(2), 128–150.

Squire, K.D., MaKinster, J.G., Barnett, M., Luehmann, A.L., & Barab, S.L. (2003). Designed curriculum and local culture: Acknowledging the primacy of classroom culture. Science Education, 87(4), 468–489. DOI: 10.1002/sce.10084.

Stigler, J.W., & Hiebert, J. (2009). The teaching gap: Best ideas from the world's teachers for improving education in the classroom (reprint). New York: Free Press.

Tschannen-Moran, M., Hoy, A.W., & Hoy, W.K. (1998). Teacher efficacy: Its meaning and measure. Review of Educational Research, 68(2), 202.

Van den Akker, J. (1998). The science curriculum: Between ideals and outcomes. In B.J. Fraser & G. Tobin (Eds.), International handbook of science education (Vol. 1, pp. 421–447). Great Britain: Springer.