COMPARING THREE APPROACHES TO PREPARING TEACHERS TO TEACH FOR DEEP UNDERSTANDING IN EARTH SCIENCE: SHORT-TERM IMPACTS ON TEACHERS AND TEACHING PRACTICE

William R. Penuel

Lawrence P. Gallagher

SRI International

COMPARING THREE APPROACHES TO PREPARING TEACHERS TO TEACH FOR DEEP UNDERSTANDING IN EARTH SCIENCE: SHORT-TERM IMPACTS ON TEACHERS AND TEACHING PRACTICE

Introduction

It is widely believed that teachers need high-quality curriculum materials to improve teaching and learning. Professional development designs differ, however, in whether they emphasize preparing teachers to use expert-designed curricula or preparing teachers with the tools needed to design and implement high-quality science units themselves. Evidence exists for the effectiveness of providing teachers with training in how to implement expert-designed curricula (Bredderman, 1983; Shymansky, Hedges, & Woodworth, 1990; Weinstein, Boulanger, & Walberg, 1982) and for providing teachers with professional development aimed at preparing teachers to design instruction and assessments (Black & Harrison, 2001; Shepard, 1997; Sneider, Adams, Ibanez, Templeton, & Porter, 1996). No studies, however, have compared explicitly these different approaches to preparing teachers to plan and enact instruction in science.

This study used an experimental design to compare the efficacy of three different approaches to professional development for preparing teachers to implement curriculum units aimed at teaching for deep conceptual understanding. All three approaches were organized around the principles of *Understanding by Design* (Wiggins & McTighe, 1998), a highly-specified process for curriculum development that focuses on teaching for and assessing conceptual understanding. In the *Investigating Earth Systems* condition, teachers learned how to *adopt* an expert-designed investigation-based curriculum unit; in the *Earth Science by Design* condition, teachers learned how to *design* high quality curriculum units; and in the *Hybrid* condition, teachers learned how to *adapt* the IES curriculum materials using a principled approach to curriculum design. In a *Control* condition, teachers received no special professional development, but were expected to teach to the same standards as teachers in the other three conditions.

In this paper, we focus on the impacts of each of the approaches to professional development on instructional planning and on the quality of assignments and assessments they give to students. We measured impacts on instructional planning using an end-of-unit questionnaire that focused on changes to teachers' overall approach to planning units of instruction, their strategies for organizing assignment, and materials they use in class. We measured impacts on overall assignment quality using a combination of Many-Facet Rasch analysis and hierarchical linear modeling techniques, using scores from rubrics focused on the nature of scientific communication, construction of knowledge, quality of Earth science content, presentation of the nature of science, and opportunities for inquiry. Scores from a holistic rubric were used to rate performance assessment quality.

The Roles of Teachers in Curriculum Design and Implementation

Since the 1950s, the National Science Foundation has been investing in the development of science curriculum materials for teachers. From the beginning, these curriculum development projects had the aim of producing deep conceptual understanding of science (Atkin & Black, 2003; Carlson, 1967). Then, as now, students had limited time in which to study particular topics, and leading figures in early curriculum development projects thought what was most important was to give students a sense of the fundamental concepts of science:

> Students, perforce, have a limited exposure to the materials they are to learn. How can this exposure be made to count in their thinking for the rest of their lives? The dominant view among men [sic] who have been engaged in preparing and engaging new curricula

> is that the answer to this question lies in giving students an understanding of the fundamental structure of whatever subjects we choose to teach. (Bruner, 1960, p. 11)

Nearly all of the early curriculum development projects funded by the National Science Foundation were led by scientists working in collaboration with educational researchers. The dominant view then, as expressed here by Jerome Bruner, a leading figure in the early curriculum development projects, was that "the best minds in any particular discipline must be put to work on the task"(1960, p. 19) by which he meant scientists who were eminent in their fields. Teachers' roles were as testers of the curricula, and often when scientists observed how teachers enacted their materials, they reacted strongly and negatively to what they saw (Atkin & Black, 2003). There was a sense then that it was "a great deal to ask" (Bruner, 1960, p. 68) of a teacher to have the requisite knowledge of subject matter to design curricula, much less implement them in a way that reflected the ideals of the designers.

In recent years, science educators and researchers have called for much more significant roles for teachers in curriculum design. Some scholars argue that teachers need to have a significant role in curriculum design, because teachers, not outside experts, are the actors in the school system with primary responsibility for student learning (Keys & Bryan, 2001; Parke & Coble, 1997). Others argue that by engaging in the activities of planning, enacting, and revising curricular units, teachers can come to understand more fully the principles of effective curriculum (Gess-Newsome, 1999; Spillane, 1999; Wiggins & McTighe, 1998). Still others argue that teachers must be central partners in design of curricular materials, because the process of adopting a curriculum inevitably involves teacher-led design and adaptation to local circumstances (Barab & Luehmann, 2003; Ogborn, 2005; Squire, MaKinster, Barnett, Luehmann, & Barab, 2003). In adapting curricula, other teachers are an important resource, since they have a deep knowledge of the constraints and opportunities of particular school contexts (Judson & Lawson, 2007; Lang, Drake, & Olson, 2006).

Policymakers and reformers today also expect more of teachers in terms of the knowledge they are expected to hold. The federal *No Child Left Behind* law requires all new secondary teachers to be "highly qualified" in the subjects they teach, meaning that they must hold a degree in that subject. In addition to subject matter knowledge, many scholars also believe it is important for teachers to possess knowledge of strategies for teaching particular content to diverse groups of students (Gess-Newsome & Lederman 1999; Loughran, Mulhall, & Berry, 2004; van Driel, Beijaard, & Verloop, 2001). Teachers' knowledge of curricular purposes and standards, materials and technologies for teaching science, and the difficulties students are likely to face when they encounter particular concepts for the first time are also considered important (Neiss, 2005; Shulman, 1986).

Even with these heightened expectations of teachers, questions remain about the extent to which teachers can design curriculum materials in ways that lead to instruction that focuses on teaching for deep understanding of subject matter. Teachers are often more heavily involved in NSF-funded curriculum projects, but accountability for those projects still rests with the Principal Investigators, who are typically educational researchers and scientists (Penuel, Roschelle, & Shechtman, 2007). Further, in projects where teachers have been heavily involved in curriculum design or co-design, researchers have noted that teachers often have trouble thinking as designers do (Reiser et al., 2000), and adaptations of materials that are inconsistent with design teams' goals for those materials are still possible (Penuel & Yarnall, 2005). Finally, although expectations for teacher knowledge have increased, there is at least an implicit acknowledgement in the field that many teachers do not yet possess sufficient content knowledge, pedagogical content knowledge, or knowledge of tools of particular science disciplines and that such knowledge needs to be the focus of professional development (Cohen & Hill, 2001) and embedded in curriculum materials themselves (Davis & Krajcik, 2005; Schneider & Krajcik, 2002).

The Challenge of Preparing Teachers to Teach for Understanding

The widely-cited finding from the curriculum analyses linked to the Third International Mathematics and Science Study (TIMSS) that U.S. science curricula are "a mile wide and an inch deep" has led to increasing attention to the need for materials that promote deep understanding (Schmidt, Raizen, Britton, Bianchi, & Wolfe, 1997). Calls for the development of materials that help teachers *teach for understanding* preceded the critical TIMSS analyses (Cohen, McLaughlin, & Talbert, 1993; Gardner & Dyson, 1994; Wiske, 1997), but widespread appeal of the notion that curriculum needs to be focused and deep among policy makers, professional developers, and school leaders is more recent. That appeal is reflected in the popularity of the *Understanding by Design* approach to curriculum planning, which in 2003 was being used by some 55,000 educators (McTighe & Seif, 2003). More recently, the creators of this program claim to have reached over 250,000 educators (http://www.grantwiggins.org/ubd.html).

The central premise behind this particular approach and behind teaching for understanding more broadly is that teachers should plan and enact instruction in which students have opportunities to learn about, experience, relate, and apply core disciplinary ideas (Gardner & Dyson, 1994; Wiggins & McTighe, 1998). There may be fewer distinct topics taught in a classroom where instruction is designed to promote deep understanding, since students may need to encounter core disciplinary ideas in different contexts to appreciate their role in scientific explanation and inquiry and their relationship to other ideas, including their own (Porter, Floden, Freeman, Schmidt, & Schwille, 1988; Smith, di Sessa, & Roschelle, 1993-1994; Wiggins & McTighe, 1998). The approach also gives a central role to assessment, particularly assessment of higher-order thinking and application of knowledge, rather than rote memorization (Wiggins & McTighe, 1998).

Teaching for understanding relies on potentially unfamiliar strategies for enacting instruction, including inquiry-oriented pedagogies. Hands-on, student investigations are important because they provide opportunities for students to experience and learn about the nature of science (National Research Council, 2000). As part of those investigations and in other assignments, students also need opportunities to develop explanations of scientific phenomena, using such warrants as evidence from investigations or references to big ideas in the discipline (Driver, Newton, & Osborne, 2000; Jimenez-Alexandre, Rodriguez, & Duschl, 2000; Kuhn, 1993; Zohar & Nemet, 2002). Memorizing facts, scientific laws, or even explanations is not sufficient for understanding; teaching for understanding requires that teachers provide opportunities for students to interpret, synthesize, evaluate, and apply knowledge (National Research Council, 1999; Oshima, Scardamalia, & Bereiter, 1996; Pressley et al., 1992; Wiggins & McTighe, 1998). To motivate students who may be used to succeeding in science class by memorizing facts, understanding requires assessments that are aligned to the goal of developing students' understanding of concepts (Ruiz-Primo & Furtak, 2007; Treagust, Jacobowitz, Gallagher, & Parker, 2001).

Past research suggests that teachers need professional development to learn how to teach for understanding. Teachers' own understanding of the nature of science and of particular science concepts is often limited, especially when compared to subject matter experts' understanding

(Gess-Newsome, 1999). Developing teachers' understanding of some science concepts may require fairly intensive exposure to those concepts (Harlen & Holroyd, 1997). Furthermore, many studies have documented the barriers to implementing inquiry pedagogies in the classroom, which are difficult in part because inquiry requires significant new roles for students and teachers in the classroom (Brown & Edelson, 1998; Crawford, 2000; Songer, Lee, & Kam, 2002). Professional development may be required for teachers to be able to use inquiry-oriented curriculum materials that give significant roles to students and present them with challenging assignments only with professional development (Spillane & Jennings, 1997; Tushnet et al., 2000). There is also evidence that teachers have a limited repertoire of strategies for assessing students' conceptual understanding: past studies have found that most teachers' assessments in science focus on measuring rote knowledge (Brady, Taylor, & Hamilton, 1989; Haertel, 1986). Even where professional development yields changes to teachers' instructional planning and classroom practice, researchers have observed that changes in assessment practice are difficult to achieve (Spicer, Stark, & Hatch, 2006).

An emerging consensus on what constitutes effective professional development in science exists (Hawley & Valli, 1999; Loucks-Horsley, Hewson, Love, & Stiles, 1998; Loucks-Horsley & Stiles, 2001). That consensus is now supported by evidence from a number of large-scale correlational studies, which point to the importance of professional development that is of an extended duration, focused on science content, linked to classroom practice, aligned to local standards, and designed to provide teachers with hands-on learning and feedback (Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001; Supovitz & Turner, 2000). At the same time, there is still a lack of empirical evidence from experimental studies that test the efficacy of particular professional development designs. The large-scale survey research that has been conducted in the past is limited in its ability to inform policy about effective designs, because it has not examined specific professional development interventions, but rather at common features of professional development *across* specific activities (Penuel, Fishman, Yamaguchi, & Gallagher, 2007). Furthermore, on particular aspects of professional development where the field has yet to reach consensus, experimental studies of interventions can provide data that inform debates about how best to prepare teachers to teach for understanding in science. The experimental study that is the focus of this paper is intended to inform policy in its focus on an enduring debate within the field of science education: the role of the teacher in curriculum design, adoption, and adaptation in which the aim is teaching for deep understanding.

The Current Study

The current paper reports results from an efficacy study that took place in a single, large urban district in the United States Southeast. In an efficacy study, an intervention or set of interventions is tested under relatively good or even ideal conditions, to investigate its potential impact and key implementation characteristics (Flay et al., 2005; Glascow, Lichtenstein, & Marcus, 2003). The conditions in the participating district were ideal for a study of professional development focused on teaching for understanding for two reasons: (1) the district had recently adopted an Understanding by Design (Wiggins & McTighe, 1998) approach for its teachers to use in guiding the development of four 9-week units throughout the school year focused on developing deep understanding of a few concepts, and (2) the district had not yet invested significant professional development in preparing teachers to use the approach. Such a context allowed us to compare the efficacy of different approaches to preparing teachers to provide instruction designed to foster deep understanding and also to test the efficacy of these approaches with a comparison group of teachers who, though they would not be exposed to the same professional development, would nonetheless be expected to teach to the same set of standards.

Each of the approaches tested in the efficacy study drew on an approach to curriculum design called Understanding by Design, which promotes instruction driven by goals for student learning tied to the "big ideas" of a discipline (Wiggins & McTighe, 1998). A central feature of Understanding by Design (UbD) is the encouragement of the use of formative assessments of student learning to provide students and teachers with feedback throughout all phases of instruction. Curriculum activities are developed only after a few key ideas have been targeted for students to learn and assessments have been developed to test whether students understand those ideas. UbD has gained tremendous popularity with educators and policy-makers across the United States in recent years; some 55,000 copies of the Understanding by Design handbook are in use by educators engaged in UbD-based initiatives (McTighe & Seif, 2003).

In the *curriculum adoption* condition, professional development activities prepared teachers to implement units of study that are part of the *Investigating Earth Systems (IES)* curriculum materials developed by the American Geological Institute in association with It's About Time Publishing. Experts on Earth science and curriculum development designed these materials, and teachers field-tested them with support from grants from the National Science Foundation. The materials are aligned to the National Science Education Standards for grades 6 through 8, and each module is organized around a set of five to seven investigations. In the investigations, students work in groups to form and/or answer questions to investigate on core topics in Earth science. Key questions drive each investigation, consistent with the UbD model, and there are pre- and post-assessments for each module teachers can use to test preconceptions and growth in student understanding over time. For the current study, each grade level used one or two of the modules judged to be most closely aligned to the district's standards. Those modules were *Investigating our Dynamic Planet, Rocks and Landforms, Investigating Water as a Resource,* and *Investigating Astronomy.*

The professional development for implementing the *IES* materials in the current study consisted of a two-week summer institute and four days of follow-up workshops spread throughout the school year. The leaders of the professional development were AGI staff (including one of the curriculum writers) and a teacher from outside the district who was experienced in training teachers to use the curriculum. The first part of the training covered topics that underpin the curriculum: typical module structure; nature of inquiry-based science and the Earth systems approach; managing materials and students working in collaborative groups; teacher support; *IES* Web site and assessment components used in *IES*. In the second part of the summer training, teachers worked in grade-level groups to develop pacing guides and practice implementing investigations they would be expected to use with their students. During the five follow-up training sessions throughout the academic year, AGI staff met with teachers to discuss issues and successes they have experienced during the implementation.

In the *curriculum design* condition, professional development activities prepared teachers to design units of study aligned to district standards using the *Earth Science by Design* (*ESBD*) program for middle school teachers. The development of *ESBD* was funded by the National Science Foundation and developed with support from AGI and TERC. During the professional development program implemented for the current study, teachers learned how to design units of instruction using principles from UbD. Core aspects of the *ESBD* program included (1) becoming aware of research on misconceptions in science (2) developing assessment strategies and instruments to measure student understanding, (3) using reflection to understand and improve teaching, and (4) learning to evaluate and incorporate scientific visualizations into the teaching of Earth science. In *ESBD*, teachers can use any materials they wish, including materials from the adopted textbook and materials they or their colleagues have developed.

The *ESBD* professional development program began with a 2-week summer institute in which teachers learn and apply the basics of UbD to the design of an Earth science unit that they will be teaching during the school year. The leaders of the summer institute were district science staff and teachers who had been part of an *ESBD* pilot study in the district. During the summer institute, teachers made decisions about what understandings they wanted students to develop (aligned to standards), developed performance-based assessments, created pre- and posttests of students' conceptual knowledge, and evaluated and identified visualizations appropriate for conveying Earth science content. To support the teachers during the school year, the *ESBD* program provided two follow-up conferences. Follow-up conferences provided teachers with opportunities to give and receive feedback on their fully developed units and to present their experiences with implementing their units.

In the *curriculum adaptation* condition, professional development activities prepared teachers to adapt units of study aligned to district standards using the *ESBD* approach in conjunction with *IES* materials. The development of this *Hybrid* condition was funded by the U.S. Department of Education as part of the current study. The basic idea behind the *Hybrid* program of professional development is to provide teachers with a principled approach to adapting *IES* units and incorporating additional materials into their teaching of Earth science units. Teachers are expected to implement at least half of the investigations in the modules selected for their grade level. Teachers can incorporate materials from other sources, including their textbook or materials they designed on their own or with colleagues, into their units, but they are required to use the ESBD planning process for selecting and aligning materials to their instructional objectives.

During the *Hybrid* group's professional development implemented for the current study, teachers learned the principles of unit design according to UbD and also gained practice with the *IES* investigations. As with the other groups, the professional development began with a two-week summer institute, which was followed up by two conferences. The two conferences followed the *ESBD* format, but the summer institute incorporated elements from both the *IES* and *ESBD* workshops. One unique feature of the *Hybrid* workshop as implemented was the opportunity for teachers to analyze the *IES* units from the standpoint of UbD, with guidance from professional developers from TERC and AGI curriculum writers who were leading the institute. Teachers in the *Hybrid* group, with support of the professional developers, also decided to work in grade-level teams to share ideas for adapting units.

In the *Control* condition, teachers did not receive any professional development related to UbD, but they were free to take any professional development, including in science, during the project. The district did expect these teachers to adopt the UbD approach when designing units, however.

To ensure a fair test of the efficacy of the different conditions, three key factors were held constant. Measures of instructional planning, classroom practice, and student achievement (not reported here), were the same for all conditions, including the control condition. For the three professional development conditions, the amount of time spent in summer institutes and follow-up was the same. Teachers spent a total of 80 hours in summer institutes, and 32 hours in follow-up activities throughout the school year. Finally, all teachers, including teachers in the control condition had a $200 supply budget, which *IES* and *Hybrid* teachers used to purchase materials needed for *IES* investigations. Other teachers could use the funds to purchase science equipment or materials.

Research Questions

In this paper, we answer a subset of four research questions being investigated by the research team related to impacts of the professional development conditions on instructional planning and practice:

1. What are the impacts of each approach on teachers' instructional planning process?
2. What are the impacts of each approach on the quality of teachers' assignments?
3. What are the impacts of each approach on the quality of teachers' culminating performance assessments of student learning?
4. How do the impacts on instructional planning and enactment compare across condition?

Methods

Study Design

This study used an experimental design, in which teachers were randomly assigned either to one of the three treatment conditions or to the control condition. Random assignment studies have the fewest threats to internal validity, and are thus more likely to yield unbiased estimates of potential impact compared with other designs (Shadish, Cook, & Campbell, 2002). The random assignment process took place after teachers volunteered to be in the study; therefore, it is important to note that the findings of this particular study cannot be generalized beyond groups of teachers who volunteer for professional development. Other efficacy studies that study the impact of the interventions when teachers were compelled to participate would be needed to establish the potential under conditions where teachers were compelled to participate.

Statistically, ideal random assignment would result in every teacher having an equal probability of being assigned to any particular condition (Shadish et al., 2002). At the same time, to minimize possible school effects we sought to balance (to the degree possible) conditions within schools and grade levels. We employed a randomization scheme with the three key features. First, each teacher had exactly a 1 in 4 chance of being assigned to a particular condition, thus meeting a fundamental condition of random assignment. Second, the number of teachers in each condition balanced across the entire sample. Third, within a school and also within a single grade level within a school, we had the widest possible dispersion of assignments to condition.

Participants

A total of 53 6th, 7th, and 8th grade teachers from 19 middle schools in a large urban district participated. Of these, 13 teachers were assigned to the IES condition, 13 to the ESBD condition, 13 to the IES condition, and 14 to the control group. There were no significant differences among groups on any of the characteristics of teachers presented below in Table 1. On average, teachers in the ESBD condition did have less experience teaching than teachers in the other conditions; the differences among groups were not statistically significant for overall years of teaching, and were marginally non-significant (χ^2=7.59, p = 0.06)[1] for years teaching science. As can be seen in Table 1, teachers in the Hybrid condition had on average less experience in teaching science than teachers in the other three conditions.

[1] χ^2 tests with 3 degrees of freedom based on null hypothesis that expected outcome is equal across all 4 conditions when a 2-level random intercepts hierarchical linear model is fit (teachers nested within schools).

Table 1. Characteristics of Faculty Respondents to Questionnaire

	Condition			
	IES	ESBD	Hybrid	Control
Gender				
Percent Male	23%	38%	23%	29%
Percent Female	77%	62%	77%	71%
*Race/Ethnicity**				
White	77%	46%	46%	79%
African American	15%	46%	46%	14%
Hispanic/Latino	8%	0%	15%	7%
Asian	8%	0%	15%	0%
Other/Unknown	0%	8%	8%	0%
Teaching Experience				
Years Teaching	*M* = 13.2 yrs *SD* = 11.8 yrs	*M* = 14.9 yrs *SD* = 11.8 yrs	*M* = 8.4 yrs *SD* = 8.4 yrs	*M* = 12.8 yrs *SD* = 9.0 yrs
Years Teaching Science	*M* = 11.5 yrs *SD* = 9.6 yrs	*M* = 8.8 yrs *SD* = 6.1 yrs	*M* = 4.3 yrs *SD* = 3.3 yrs	*M* = 10.4 yrs *SD* = 8.1 yrs
Highest Degree†				
Bachelor's	69%	77%	85%	64%
Master's	23%	23%	8%	36%
Educational specialist's	8%	0%	8%	0%
Missing	0%	0%	0%	0%
Teaching Assignment				
6	5	7	6	3
7	3	2	4	4
8	5	4	3	7

Note: table is based on 53 teachers who submitted examples of assignments for analysis
***** Teachers could select multiple categories.
† Totals may not sum to 100% due to round off error

As Table 2 shows, although members of every group had received some professional development in the *Understanding by Design* approach, there were no significant differences among groups in overall exposure. Similarly, there were no significant differences among groups in professional development related to the development of classroom assessments or unit planning.

Table 2. Participation in Professional Development by Topic and Condition

Topic/Condition	0 hours	1-8 hours	9-16 hours	17-25 hours	25+ hours
Understanding by Design					
IES	31%	38%	8%	8%	15%
ESBD	23%	69%	0%	8%	0%
Hybrid	31%	54%	8%	0%	8%
Control	36%	29%	14%	14%	7%
TOTALS	30%	47%	8%	8%	8%
Classroom assessment					
IES	15%	31%	23%	0%	31%
ESBD	8%	38%	15%	15%	23%
Hybrid	15%	62%	8%	8%	8%
Control	36%	29%	7%	21%	7%
TOTALS	19%	40%	13%	11%	17%
Unit Planning					
IES	46%	23%	15%	0%	15%
ESBD	15%	38%	15%	23%	8%
Hybrid	31%	46%	8%	0%	15%
Control	50%	36%	0%	7%	7%
TOTALS	36%	36%	9%	8%	11%

Note: Percentages may not sum to 100% across columns due to round off error

Sources of Data

Teacher Questionnaire

Data on teachers' instructional planning were collected through a questionnaire. The questionnaire focused on a range of topics related to teachers' instructional planning process and unit implementation. The analysis presented in this paper focuses on the three aspects of the planning process described below.

Influence on Instructional Planning Process. This measure was comprised of a single item related to the influence of the project. Teachers answered the question, "Has your TIDES experience influenced how you plan instruction in science?" There were three ordinal response options: (1) not at all; (2) somewhat; and (3) very much.

Reported Changes to How Teachers' Organized Assignments. This measure was comprised of four items related to teachers' instructional planning process. The four items were related to aspects of the instructional planning process targeted by professional development in the Understanding by Design model: (1) use of activities aligned to big ideas of the discipline; (2) number of topics covered; (3) selection of materials to develop deep understanding; and (4) selection of materials intended to hook students' interest. Teachers rated whether each of these had changed since the beginning of the research project began on a scale from 1 (not at all) to 3 (very much). A scale representing the raw sum of responses was constructed from the items, with a reliability of $\alpha = 0.81$.

Use of Instructional Materials. This measure was comprised of five items, each analyzed separately, focused on changes to teachers' use of materials when developing and teaching their Earth science units. Teachers responded to questions about their use of the district-selected textbook, the Internet, visualizations (a focus of the ESBD and Hybrid professional development), materials they created, and materials created by colleagues. For each, teachers responded on a six-point scale ranging from "used much less" to "used a lot more."

Teacher Assignments

A number of evaluation studies in recent years have employed teacher assignments as indicators of the quality of instruction in classrooms (e.g., Matsumara & Steinberg, 2002; Shkolnik et al., 2007). Validation studies have found that teacher assignments can be rated reliably and consistency, and that the quality of assignments is significantly correlated with direct observations of instructional quality (Clare & Aschbacher, 2001). Validation studies have also demonstrated positive associations between the quality of assignments and the work students produce in class (Matsumara, Patthey-Chavez, Valdez, & Garnier, 2002; Shkolnik et al.,

2007) and standardized test scores (Matsumara, Garnier, Pascal, & Valdes, 2002; Newmann, Bryk, & Nagaoka, 2001). Ours is one of the first studies to examine assignment quality in science; earlier studies have examined the quality of language arts and mathematics assignments.

The study team developed five four-point rubrics to analyze teacher assignments. The rubrics for assignments addressed each of the following dimensions of assignments:

Scientific Communication. Pertains to the extent to which the assignment requires students to engage in scientific communication to demonstrate their understanding of science content.

Indicators: Students are asked to engage with and communicate about science content by posing an investigable question, making a claim, forming a hypothesis, or drawing a conclusion. Requirements for students to draw on supporting evidence (e.g. observations, examples, details, illustrations, facts, data, and/or logical reasoning) and incorporate it in to their communication are apparent. Students are asked to make connections with and refer to larger "big ideas" in science in their communication.

Construction of Knowledge about Earth Science Content. Pertains to the extent to which the assignment facilitates students' construction of new knowledge about Earth science content, and to what extent does it ask students to apply that knowledge to a different situation.

Indicators: Students are asked to construct new knowledge about Earth science content (e.g. concept, phenomena, idea, process, etc.) by investigating, interpreting, analyzing, synthesizing, or evaluating information. Students are also asked to apply the knowledge they have gained within the assignment to a different situation.

Quality of Earth Science Content. Pertains to the quality of the Earth science content students encounter in the assignment.

Indicators: Assignments that address or include high quality Earth science content are those in which students encounter content that is aligned with a “big idea” that is significant and fundamental to the field, appropriate & challenging, and accurate.

Approach to the Nature of Science. Pertains to the extent to which the assignment frames science as a dynamic body of knowledge developed through investigation and calls for students to understand or experience how science is done, including the tools and processes involved.

Indicators: The assignment presents or frames science as a dynamic body of knowledge developed through investigation, rather than as an isolated set of facts to be memorized.

Scientific Inquiry. Pertains to the extent to which the assignment engages students in all aspects of the inquiry process

Indicators: Students are asked to pose questions, select methods to use to answer those questions, carry out an investigation, and analyze and communicate results. The quality of teachers’ performance assessments was coded using a holistic rubric.

Performance Assessment Quality. Pertains to the extent to which the performance assessment requires students to apply the content they have learned, and to demonstrate the understanding and skill they have gained

Indicators: The performance assessment asks students to complete an open-ended project or solve an open-ended problem. The assessment requires students to apply the content they have learned and to demonstrate the skills and /or communicate the understanding they have gained. The performance assessment involves providing students with clear expectations and evaluation criteria before students begin the task.

Procedures

Questionnaire Administration

The research team administered questionnaires to teachers online, using a commercial Web-based survey program. Teachers completed the questionnaires within 3 to 4 weeks of completing teaching of their Earth science unit. Teachers completed units at different times; the study team used information obtained from teachers through an earlier survey to identify each teacher's ideal window for questionnaire completion. For 6^{th} and 8^{th} grade teachers in the study, the questionnaire windows were in January and February 2007. For the 7^{th} grade teachers, the questionnaire windows were all in May. We used the online survey program to monitor response rates and followed up with teachers, resulting in a 91 percent response rate (n = 53).

Teacher Assignment Data Collection and Scoring

The research team collected three assignments and one performance assessment from each teacher in the study during and after their Earth science units. We provided teachers in all four conditions with a list of broad criteria that would be used in coding assignments, and asked teachers to submit assignments they believed would best represent the ideal assignments we outlined from three different time points in their unit: one assignment from weeks 1 to 3, a second from weeks 4 to 6, and a third from weeks 7 to 9. In addition, we asked teachers to submit one culminating performance assessment and indicated in general terms how these would be coded. This approach certainly means that we did not necessarily receive typical assignments from teachers; thus, the findings from the study cannot be said to capture the full range of assignments teachers gave students. At the same time, the bias we introduced could not be expected to differentially benefit any one group, since the criteria for coding assignments were established independently from the interventions being studied. Thus, the impact estimates are fair judgments of the differences between intervention groups with respect to teachers' submissions of what they saw as their best assignments and performance assessment.

A member of the study team visited each classroom to assist teachers with preparing packets for submitting assignments. The purpose of the visit was to ensure that each teacher submitted a cover sheet describing the context of the assignment, as well as all relevant worksheets, instructional materials, and samples of student work that would allow coders to get as comprehensive a view of what the assignment entailed. For any curriculum materials employed as part of assignments, including IES materials, teachers were instructed to make photocopies of the relevant pages of the student or teacher guide that outlined what teachers and students did in class as part of the assignment.

Teachers across conditions submitted a total of 197 assignments. As Table 3 below shows, the assignments submitted reflect a high level of consistency with what would be expected from teachers in each condition. Among IES teachers, 42 of 52 assignments submitted were from the curriculum. ESBD included no assignments from the IES curriculum, and the distribution of types of materials submitted was similar to the control group's distribution. Finally, the Hybrid condition teachers submitted a mix of assignments from the IES curriculum and from materials they had developed on their own.

Table 3. Types of Assignments Submitted by Condition

	Condition				
	IES	ESBD	Hybrid	Control	TOTAL
Activities from the IES Curriculum	42	0	15	0	57
Teacher-Modified IES Activities	5	0	6	0	11
Textbook/Publisher-Developed Activities	0	10	3	12	25
Activities Obtained from Other Professional Development	0	1	2	5	8
District Performance Task	0	1	0	7	8
Teacher-Developed Assignment (Developed On Own)	1	18	5	16	40
Teacher-Developed Assignment (As Part of Project Workshop)	0	3	10	0	13
Other	0	2	2	2	6
TOTAL	52	47	48	50	197

A coding team analyzed the teacher assignments and performance assessments according to the rubrics described above. A science education researcher with a background in teaching led the coding session. The coding team consisted of five members, all experienced middle school science teachers. The coding team members were all blind as to the condition to which teachers in the study had been assigned; in addition, teacher identifiers of all kinds were stripped from assignments, so coders were blind to the teacher whose assignment they were coding.

The coding leader trained and calibrated the coders on each rubric independently, and the coders coded for that rubric only, before moving on to the next rubric. She used actual examples and examples from other sources as training papers. During the first 25% of coding on a rubric, the coders' decisions were entered into a spreadsheet, and reliability was calculated. If there was a discrepancy of more than one code point on the scale, the coding leader reconvened the coders for recalibration. Overall, inter-coder agreement within one point of the rubric scales was 87 percent. This level of reliability is comparable to other studies of teacher assignments (Shkolnik et al., 2007).

We randomly assigned all assignments to the members of the coding team. To accomplish this task, we constructed a matrix that randomly assigned the assignment and assessment artifacts to members of the coding team such that no coder saw the same artifact twice for the same rubric, and each coder saw the artifacts no more than three times across the five or six rubrics. This was done to ensure that the codes for each rubric on each set of artifacts were independent of an overall impression of the assignment by a coder that might be formed if the coder saw the same assignment multiple times. Further, the coders never saw an assignment twice in a row on two different rubrics. All assignments and performance assessments were coded twice by two separate coders.

Approach to Data Analysis

Survey Results

We fit a 2-level hierarchical linear model for two of the three instructional planning variables analyzed in the study using the Stata statistical package (version 9). Our primary interest was not the omnibus test of overall equality among all 4 experimental conditions, however. Rather, we were interested in pair-wise comparisons between each treatment group and the control group first. Secondarily, we were interested in pair-wise comparisons among the treatment group. In any cases where the omnibus test was statistically significant at $p < .05$, we examined all pair-wise comparisons, employing a false discovery rate correction for family-wise Type I error rate (Benjamini & Hochberg, 1995). For the third instructional planning variable, overall influence on their instructional planning process, we analyzed the categorical data using an ordered logistic regression model with correction for clustering of teachers within schools.

Teacher Assignments and Performance Assessments

The estimation procedures for analyzing the quality of assignments are based on earlier procedures first developed by Newmann and colleagues in their studies conducted in Chicago Public Schools (Newmann et al., 2001; Newmann, Lopez, & Bryk, 1998) and as elaborated by study teams at the American Institutes for Research and SRI International (Shkolnik et al., 2007). The analysis first involves using a Many-Facet Rasch Model (MFRM) (Lineacre, 1989) to combine scores from individual rubrics for assignments into a single score for quality, and then conducting a hierarchical linear modeling (HLM) analysis to analyze treatment effects.

The Many-Facet Rasch Model is a technique to adjust statistically for the differences between scorers and rubrics. The model uses data from both scorers to estimate the severity of raters, so that raw scores can be adjusted for how harsh or easy a particular rater is. In addition, the model uses data across rubrics to adjust for the difficulty of different rubrics. On some rubrics, it may be much easier to score a "4" than on others, and the model takes the difficulty into account in estimating a score for each assignment. For purposes of our analyses, we analyzed the holistic rubric scores for the performance assessment quality rubric using the Many-Facet Rasch Model, since we had data available from two scorers and other rubrics to estimate parameters for rater severity and rubric difficulty. At the end of the analysis, the modeling procedure re-scales the raw scores from 0 to 10.

Hierarchical linear modeling is the appropriate statistical procedure for estimating impacts when data are nested as they were in this study (Raudenbush & Bryk, 2002). In the models we tested, assignments are nested within teachers (who are part of treatment conditions), and teachers are nested within schools. We tested separate models for seven dependent variables: overall assignment quality, each of the individual rubrics, and overall performance assessment quality. We first fit unconditional models, to test the significance of variance at the assignment, teacher, and school levels. For all seven dependent variables, the unconditional variance component at the school level was nearly zero (less than 1% of the total variance). Therefore we adopted a 2-level model (assignment within teacher) for the subsequent analyses. [2]

For each model, we included three teacher-level (Level 2) predictors. The first predictor, treatment condition, was entered as a set of 3 indicator variables (where the Control condition was the omitted category). The models also incorporated grade level taught (as 2 indicator variables for 3 grade levels) and years teaching science (as a continuous predictor). We found

[2] Note that for the overall performance assessment score, with 1 performance assessment per teacher, a simpler one-level regression model was used.

significant interaction effects between treatment condition and grade level taught, and therefore incorporated these interactions into the model. The full model is:

Level 1 (Assignment)

$$
\begin{aligned}
Y_{ij} = {} & \beta_{0j} + \beta_1 IES + \beta_2 ESBD + \beta_3 Hybrid \\
& + \beta_4 G6 + \beta_5 G7 \\
& + \beta_6 IES * G6 + \beta_7 IES * G7 + \beta_8 ESBD * G6 + \beta_9 ESBD * G7 + \beta_{10} Hybrid * G6 + \beta_{11} Hybrid * G7 \\
& + \beta_{12} Years + \varepsilon_{ij}
\end{aligned}
$$

Level 2 (Teacher)

$$\beta_0 = \gamma_{00} + u_j$$

Where the intercept β_{0j} is modeled as a random effect of teachers, and all other predictors are fixed effects.

Corrections for Multiple Pair-wise Comparisons

Because in both the survey and HLM analyses we are conducting multiple comparisons within the same dataset, the risk of family-wise error is increased. In particular, the likelihood of obtaining a significant result increases with multiple tests, even when the probabilities are low for detecting statistical significance. Throughout this paper, we report significance levels for treatment impacts and for comparisons among the treatment groups. For both, we employed a false discovery rate correction (Benjamini & Hochberg, 1995) to establish a more conservative criterion for statistical significance. Pairwise comparisons among four experimental conditions yields 6 statistical tests; the false discovery rate correction ensures that fewer than 5% of the reported significant results in any one family of tests will be the result of a Type I error.

Results

Impacts on Instructional Planning

There were statistically significant differences among the four conditions with respect to teachers' reports of the project's influence on how they plan instruction in science. Table 4 shows the distribution of responses to the implementation questionnaire item related to overall impacts on planning. All of the ESBD and Hybrid teachers reported changes to their practice; by contrast, at least some teachers in the IES and Control conditions reported no changes.

Table 4. Reported Influences of Project on Instructional Planning in Science

	Condition			
	IES	ESBD	Hybrid	Control
"Not at all"	3	0	0	5
"Somewhat"	7	2	5	6
"Very much"	3	9	8	2

Model F (3,16) = 8.23, p<.01

The kinds of changes to instructional planning reported were consistent with the Understanding by Design model, where there was also an overall effect of condition on planning, F (3,16) = 8.23, $p < .01$). Post-hoc tests revealed that the ESBD and Hybrid intervention conditions outscored both the IES and Control conditions.

With respect to uses of instructional materials, there were significant overall effects of condition for all the materials queried in the implementation survey (see Table 5). Post-hoc pair-wise comparisons indicated that teachers in the IES condition made significantly less use of the district textbook than teachers in all other conditions. Teachers in each of the intervention conditions made significantly more use of Internet-based resources. Teachers in the ESBD and Hybrid conditions made more use of visualizations than did teachers in the IES and control conditions. Teachers in the ESBD condition also made more use of materials they themselves had created than teachers in the Hybrid condition, and teachers in the Hybrid condition in turn made more use of self-created materials than teachers in the condition. Finally, teachers in the

Hybrid condition made more use of materials colleagues had developed than did teachers in the IES and control conditions. All statistically significant pairwise comparisons within each type of instructional materials were controlled for a false discovery rate of $p < .05$.

Table 5. Changes to Instructional Materials Used in Earth Science Units

Type of Material	Condition	M	SD	Omnibus χ^2
District-selected textbook	IES[abc]	1.15	0.38	42.95***
	ESBD[a]	3.18	1.60	
	Hybrid[b]	3.23	0.93	
	Control[c]	3.38	0.87	
The Internet	IES[a]	5.08	0.76	23.32***
	ESBD[b]	5.64	0.50	
	Hybrid[c]	4.92	0.76	
	Control[abc]	3.62	1.76	
Visualizations	IES[ab]	4.31	0.85	33.89***
	ESBD[ac]	5.73	0.47	
	Hybrid[bd]	5.31	0.85	
	Control[cd]	4.54	0.77	
Materials I created	IES[abc]	2.85	1.72	27.23***
	ESBD[ad]	5.18	0.87	
	Hybrid[bd]	3.85	0.99	
	Control[c]	4.46	0.66	
Materials colleagues created	IES[a]	1.77	1.30	15.58**
	ESBD	2.82	1.60	
	Hybrid[ab]	3.77	1.30	
	Control[b]	2.15	1.41	

Scale: 1 = did not use; 2 = used much less; 3 = used a little less; 4 = used about the same; 5 = used a little more; 6 = used a lot more

** $p < .01$ *** $p < .001$

Within a type of material, experimental conditions sharing a superscript letter in common are statistically significantly different from one another when controlling for a false discovery rate of p<.05.

Impacts on Assignment Quality

Table 6 below presents descriptive statistics for mean assignment quality scores by condition for each dimension of the rubric and for overall assignment quality. Overall means by rubric dimensions suggests that quality of content was the "easiest" rubric dimension, while it was hardest to earn a high rating on science inquiry.

Table 6. Mean Scale Scores by Condition (combining grades)

	Condition			
	IES	ESBD	Hybrid	Control
Scientific Communication	*M* = 6.45 *SD* = 1.68	*M* = 5.33 *SD* = 1.85	*M* = 6.14 *SD* = 1.41	*M* = 5.59 *SD* = 1.44
Construction of Knowledge	*M* = 5.48 *SD* = 2.18	*M* = 3.99 *SD* = 1.50	*M* = 5.14 *SD* = 1.71	*M* = 4.42 *SD* = 1.74
Quality of Content	*M* = 6.73 *SD* = 1.67	*M* = 6.33 *SD* = 2.11	*M* = 7.20 *SD* = 1.68	*M* = 6.30 *SD* = 1.79
Nature of Science	*M* = 5.35 *SD* = 1.76	*M* = 3.32 *SD* = 1.54	*M* = 5.19 *SD* = 2.36	*M* = 3.55 *SD* = 1.58
Scientific Inquiry	*M* = 3.47 *SD* = 2.48	*M* = 1.46 *SD* = 1.36	*M* = 2.67 *SD* = 2.22	*M* = 1.76 *SD* = 2.21
Overall Assignment Quality	***M* = 5.64** ***SD* = 0.83**	***M* = 4.79** ***SD* = 0.86**	***M* = 5.53** ***SD* = 0.84**	***M* = 4.99** ***SD* = 0.77**
Performance Assessment	*M* = 3.17 *SD* = 1.50	*M* = 3.85 *SD* = 1.33	*M* = 5.27 *SD* = 2.22	*M* = 4.16 *SD* = 2.44

Table 7 shows effect sizes and significance levels for pair-wise comparisons across conditions. Overall assignment quality was significantly higher in both the IES and Hybrid conditions than in the control condition. Among the treatment conditions, assignment quality was significantly higher in the IES and Hybrid conditions than in the ESBD condition.

Table 7. Effect Sizes, Overall Assignment Quality

Condition	ESBD	Hybrid	Control
IES	1.10*	0.15	0.85*
ESBD		0.95*	0.25
Hybrid			0.70*

*statistically significant difference controlling for false discovery rate of .05 among 6 comparisons

Assignment quality in the IES condition was higher with respect to the extent to which assignments called for elaborated written or verbal scientific communication when compared to assignments in the control conditions. The quality of scientific communication was also higher in the IES and Hybrid conditions (which adapted IES materials) than in the ESBD condition..

Table 8. Effect Sizes, Scientific Communication

Condition	ESBD	Hybrid	Control
IES	0.78*	0.21	0.60*
ESBD		0.56*	0.18
Hybrid			0.38

*statistically significant difference controlling for false discovery rate of .05 among 6 comparisons

The IES condition significantly outscored the control condition with respect to how much assignments called on students to construct and synthesize scientific knowledge. Among the intervention conditions, comparisons revealed that assignments were rated of a higher quality on this dimension in the IES and Hybrid conditions when compared with assignments in the ESBD condition (Table 9).

Table 9. Effect Sizes, Construction of Knowledge

Condition	ESBD	Hybrid	Control
IES	0.85*	0.19	0.61*
ESBD		0.66*	0.24
Hybrid			0.42

*statistically significant difference controlling for false discovery rate of .05 among 6 comparisons

With respect to judgments about the accuracy and significance of Earth science content included in the submitted assignments, there were no statistically significant differences among conditions (Table 10).

Table 10. Effect Sizes, Quality of Earth Science Content

Condition	ESBD	Hybrid	Control
IES	0.23	0.26	0.24
ESBD		0.49	0.02
Hybrid			0.51

No effect sizes are statistically significant controlling for false discovery rate of .05 among 6 comparisons

Assignments submitted by teachers in the IES and Hybrid condition were rated higher than assignments submitted by teachers in the control condition with respect to the nature of science reflected in the work they demanded of students (Table 11). In other words, students in these conditions were more likely to be exposed to science as a body of disciplinary knowledge and tools that grows through ongoing, collaborative work of scientists engaged in inquiry. Among the intervention conditions, IES and Hybrid condition assignments scored significantly higher than ESBD assignments as well.

Table 11. Effect Sizes, Nature of Science

Condition	ESBD	Hybrid	Control
IES	1.28*	0.10	1.14*
ESBD		1.18*	0.15
Hybrid			1.04*

*statistically significant difference controlling for false discovery rate of .05 among 6 comparisons

Teachers in the IES condition submitted assignments that were more likely than teachers in the control condition to provide students with opportunities to engage in scientific inquiry (Table 12). Raters scored assignments in this condition significantly higher than assignments in the ESBD condition as well.

Table 12. Effect Sizes, Scientific Inquiry

Condition	ESBD	Hybrid	Control
IES	0.91*	0.36	0.77*
ESBD		0.55	0.14
Hybrid			0.41

*statistically significant difference controlling for false discovery rate of .05 among 6 comparisons

Impacts on Performance Assessment Quality

Table 6 above presents descriptive statistics for overall performance assessment quality by condition. This holistic rubric considered the extent to which performance assessments submitted presented students with an open-ended project or problem that gave them an opportunity to apply what they had learned from their Earth science unit and to demonstrate through that application their understanding of the unit's content. As Table 13 shows, none of the intervention conditions outperformed the control condition with respect to performance assessment quality, nor were the mean ratings in any one condition significantly different than those of any other condition

Table 13. Mean Scale Scores by Grade and Condition, Performance Assessment Quality

	Condition			
	IES	ESBD	Hybrid	Control
Total Score	*M* = 3.17 *SD* = 1.50	*M* = 3.85 *SD* = 1.33	*M* = 5.27 *SD* = 2.22	*M* = 4.16 *SD* = 2.44

Discussion

The findings from this study point to the promise of professional development focused on instructional planning, but also to the central importance of curriculum in improving instructional quality. On the one hand, the interventions designed to influence instructional planning processes used by teachers did so, and all of the interventions influenced the kinds of materials used by teachers in ways consistent with the models of instructional quality promoted by each intervention. At the same time, the assignments submitted by teachers in the conditions where teachers either used their regular textbook or materials they created as part of their professional development experience were not judged by independent raters to be as consistent with the idea of teaching for deep understanding as were assignments submitted by teachers who had access to expert-designed curriculum materials in Earth science.

The two interventions that most influenced teachers' instructional planning process were ones that explicitly taught those skills. In the ESBD and Hybrid conditions, professional development activities focused on preparing teachers to design or use activities aligned with big ideas in Earth science, to reduce the number of topics covered to focus on depth rather than breadth, and to select materials that could foster both deep understanding and student engagement. Implementation data suggested that both these interventions achieved this goal; furthermore, teachers in the ESBD condition, which required the most planning and design on the part of teachers, were significantly more likely than teachers in the curriculum adoption condition (IES) to report significant changes to how they planned instruction.

It is notable that teachers assigned to the other two conditions did report changes to their instructional planning. The curriculum adoption process appeared to shape teachers' process for planning standards-based units in Earth science, even in the control condition, where teachers had to make adjustments to lessons used in the past to fit within the new science textbook. For their part, IES teachers reported making adjustments to their planning process, including making much less use of the district textbook in planning and coordinating instruction.

Together, these two sets of findings regarding instructional planning suggest that adaptation is an ongoing feature of teachers' instructional planning. All of the teachers to some degree adjusted the way they approached planning instruction in Earth science, and the materials they used also changed. This particular finding is not surprising, in that it is consistent with prior policy and curriculum implementation research and with theories of implementation that view curriculum "adoption" as necessarily always involving some "adaptation" of materials.

What is significant is that our study found that planned adaptation can yield desired changes to instruction. Teachers in the ESBD and Hybrid condition all learned to use strategies for planning instruction that were part of the Understanding by Design model, and they reported using these strategies more than teachers in the control condition. Perhaps because the IES materials embedded this model of planning into the curriculum materials, those teachers, too, reported changes to planning consistent with the model.

With respect to overall assignment quality, our findings suggest that curriculum does in fact matter. Overall quality was highest in the IES and Hybrid conditions, where teachers were permitted and often did submit expert-designed IES materials as examples of assignments they gave students. Quality was lower in the ESBD condition, where teachers did not have access to the IES materials and where teachers reported using many more materials they themselves had designed than in the other conditions.

The fact that IES materials were designed in such a way that scores on the independent rubric might be expected to be higher than materials developed by teachers does not diminish the significance of the finding. Instead, they point to the potential benefit of expert-designed curricula, namely that such curricula provide enhanced opportunities for students to learn. In this particular study, the IES materials were particularly powerful with respect to how much they exposed students to the nature of science and to the process of scientific inquiry, something that teachers were not able to do as well either on their own or by using the district-adopted textbook in science.

The findings do not support the claim, however, that these interventions were effective in improving the quality of performance assessments students encountered at the end of their Earth science units. None of the average ratings of performance assessments from the intervention conditions were significantly higher than the control condition. Even in the IES condition, where teachers could submit assessments from the curriculum, performance assessment scores were no higher than in the control condition. This finding may be attributable to a mismatch between IES assessments and the rubric, which was based on definitions of assessment quality presented by Wiggins and McTighe (1998), but it may also point to the challenge of providing professional development to teachers that can significantly influence their assessment design skills.

Implications for Policy and Practice

This particular study was intended to shed empirical light on an enduring debate within the policy, professional development, and curriculum communities about the role of teachers in curriculum. The use of random assignment limits the threats to internal validity for the study, making the study a potentially powerful source of evidence in this debate. Further, the use of objective, reliable measures of instructional quality linked to a valued educational end—teaching for understanding—that were applied to naturally occurring teacher assignments contributes to the external validity of the findings.

Consistent with calls for greater attention to curriculum that followed international comparison studies of mathematics and science achievement (e.g., Schmidt, McKnight, & Raizen, 1997), this study's findings suggest that curriculum does matter for teacher practice. Those teachers with access to curriculum materials in science that present science as a discipline that grows through inquiry and that provide students with the opportunity to experience the inquiry materials directly are more likely to be able to provide students with the opportunity to develop an understanding for the nature of science and skills of science inquiry. Conversely, if teachers rely on their own materials or select materials with little guidance as to how to do so to foster deep understanding, students' may have fewer opportunities to develop these particular understandings and skills.

The study also provides evidence for the perspective held by a number of researchers who study implementation (e.g., Barab & Luehmann, 2003; Reiser et al., 2000) that "adaptation" is part of even the most straightforward curriculum adoption process. Because mandated texts and materials are ever changing, teachers adjust the materials they use and their planning process, even without much outside intervention. Furthermore, when given a set of materials to use, at a minimum, teachers must fit those materials into a schedule that fits within their school's schedule and that follows any pacing guidance given by their district and state.

The findings are also consistent with earlier findings on teacher professional development, which suggest that when planning for implementation is incorporated to professional development, the nature and quality of implementation is more likely to be consistent with curriculum developers' intent (Cohen & Hill, 2000; Penuel, Fishman et al., 2007). The policy implication is that professional development for "planned adaptation" should accompany any curriculum adoption and that, for middle school Earth science at least, there is some evidence to suggest the Understanding by Design approach is a viable approach to instructional planning.

We note that our analyses and discussion have focused on instructional change, not direct assessment of student learning. Evidence from our study indicates teachers' instructional planning as well as qualities of their assignments were clearly influenced by professional development and materials. But whether these experimental conditions differentially predict student learning is still an open question. The assignments we collected from teachers were a sample of assignments, and therefore may not have represented the domain of constructs measured on tests of achievement. Furthermore, achieving impacts on instruction does not guarantee impacts on achievement (Rowan & Miller, 2007). We are still developing and testing models that examine the direct impacts of these conditions on student achievement and analyzing the relationships between assignment quality and achievement.

Limitations of the Study and Directions for Future Research

One limitation of this study is the absence of a measure of prior instructional quality; some variance may be due to initial differences in quality among teachers in the study. This particular experiment was relatively small, so that even with random assignment, there is the possibility that prior differences in instructional quality could have shaped the outcomes measured here. Future studies might employ a longitudinal design with a baseline year after initial recruitment, to collect data on prior instruction; this solution would provide a direct prior measure, but risks of attrition increase with such a design.

Second, our assignments represent a sample of those students encountered in their Earth science units, and that sample is no doubt biased such that our study team received the most challenging and interesting assignments from teachers. Teachers, not researchers, selected from all possible assignments to be given, and the selection process was not random. The guidance we provided to teachers, to be sure, was the same across condition; however, to the extent that the professional development activities made it more likely that teachers would be sensitive to particular dimensions of assignment quality on which their assignments were ultimately scored, there may have been some interaction of treatment condition with selection of assignments submitted. Alternate methods that could increase the external validity of findings might be to randomly select from all assignments or to incorporate more observations into the study. The latter option is potentially much more expensive, as many observations may be required to achieve sufficient confidence in teacher ratings of instructional quality.

More evidence about how best to prepare teachers to adapt curriculum materials in a principled manner and about how to improve assessment practice is needed to shed further light on the debate about the role of teachers in curriculum. For example, studies in other domains of science education may yield evidence about the role of scientific domain in curriculum planning and enactment. Similarly, studies in domains outside science may yield either convergent or conflicting evidence about the value of curriculum or of planning for principled adaptation of curriculum. In the end, no single study, no matter how well designed, can settle such an important policy matter (National Research Council, 2002). No doubt, even with more evidence, different policymakers and reformers will continue to advocate for their particular positions. But by accumulating evidence from multiple studies in different domains, that debate can be informed to a much greater extent by evidence from research.

References

Atkin, J. M., & Black, P. (2003). *Inside science education reform: A history of curricular and policy change*. New York: Teachers College Press.

Barab, S., & Luehmann, A. L. (2003). Building sustainable science curriculum: Acknowledging and accommodating local adaptation. *Science Education, 87*(4), 454-467.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289-300.

Black, P., & Harrison, C. (2001). Feedback in questioning and marking: The science teacher's role in formative assessment. *School Science Review, 82*(301), 55-61.

Brady, M. P., Taylor, R. D., & Hamilton, R. (1989). Differential measures of teachers' questioning in mainstreamed classes: Individual and classwide patterns. *Journal of Research and Development in Education, 23*(1), 10-17.

Bredderman, T. (1983). Effects of activity-based elementary science on student outcomes: A quantitative analysis. *Review of Educational Research, 53*(4), 499-518.

Brown, M., & Edelson, D. C. (1998). Matching the design of activities to the affordances of software to support inquiry-based learning. In A. S. Bruckman, M. Guzdial, J. L. Kolodner & A. Ram (Eds.), *Proceedings of ICLS 98:International Conference on the Learning Sciennces, Atlanta, GA, December 16-19* (pp. 77-83). Charlottesville, VA: AACE.

Bruner, J. S. (1960). *The process of education*. Cambridge, MA: Harvard University Press.

Carlson, J. S. (1967). Science and the curriculum. *Science Education, 51*(3), 251-254.

Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*(1), 39-59.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The rathematics reform in California. *Teachers College Record, 102*(2), 294-343.

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

Cohen, D. K., McLaughlin, M. W., & Talbert, J. E. (1993). *Teaching for understanding: Challenges for policy and practice*. San Francisco, CA: Jossey-Bass.

Crawford, B. (2000). Embracing the essence of inquiry: New roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916-937.

Davis, E. A., & Krajcik, J. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34*(3), 3-14.

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

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Science, 6*(3), 151-175.

Gardner, H., & Dyson, V. (1994). Teaching for understanding in the disciplines and beyond. *Teachers College Record, 96*(2), 198-218.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Gess-Newsome, J. (1999). Secondary teachers' knowledge and beliefs about subject matter and their impact on instruction. In J. Gess-Newsome & L. M. Lederman (Eds.), *Pedagogical content knowledge and science education* (pp. 51-94). Boston: Kluwer.

Gess-Newsome, J., & Lederman , L. M. (Eds.). (1999). *Examining pedagogical content knowledge*. Boston, MA: Kluwer.

Glascow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research into practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health, 93*(8), 1261-1267.

Haertel, E. (1986, 1986, April). *Choosing and using classroom tests: Teachers' perspectives on assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Harlen, W., & Holroyd, C. (1997). Primary teachers' understanding of concepts of science: Impact on confidence and teaching. *International Journal of Science Education, 19*(1), 93-105.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 127-150). San Francisco: Jossey-Bass.

Jimenez-Alexandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education, 84*(6), 757-792.

Judson, E., & Lawson, A. E. (2007). What is the role of constructivist teachers within faculty communication networks? *Journal of Research in Science Teaching, 44*(3), 490-505.

Keys, C. W., & Bryan, L. A. (2001). Co-constructing inquiry-based science with teachers: Essential research for lasting reform. *Journal of Research in Science Teaching, 38*(6), 631-645.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77*(3), 319-337.

Lang, M., Drake, S., & Olson, J. (2006). Discourse and the new didactics of scientific literacy. *Journal of Curriculum Studies, 38*(2), 177-188.

Lineacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.

Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin Press.

Loucks-Horsley, S., & Stiles, K. E. (2001). Professional development designed to change teaching and learning. In J. Rhoton & P. Bowers (Eds.), *Professional development planning and design* (pp. 13-24). Arlington, VA: NSTA Press.

Loughran, J., Mulhall, P., & Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching, 41*(4), 370-391.

Matsumara, L. C., Garnier, H. E., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8*(3), 207-229.

Matsumara, L. C., Patthey-Chavez, G. G., Valdez, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*(1), 3-25.

Matsumara, L. C., & Steinberg, J. R. (2002). *Collaborative, school-based professional development settings for teachers: Implementation and links to improve the quality of classroom practice and student learning*. Los Angeles, CA: Center for Research

McTighe, J., & Seif, E. (2003). *A summary of underlying theory and research base for understanding by design*.Unpublished manuscript.

National Research Council. (1999). *How people learn: Brain, mind, experience*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Neiss, M. L. (2005). Preparing teachers to teach science and mathematics with technology: Developing a technology pedagogical content knowledge. *Teaching and Teacher Education, 21*(5), 509-523.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

Newmann, F. M., Lopez, F., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago Schools: A Baseline Report*. Chicago, IL: Consortium on Chicago School Research.

Ogborn, J. (2005). 40 years of curriculum development. In K. Boersma, M. Goedhart, O. de Jong & H. Eijkelhof (Eds.), *Research and the quality of science education*. Dordrecht, Netherlands: Springer-Verlag.

Oshima, J., Scardamalia, M., & Bereiter, C. (1996). Collaborative learning processes associated with high and low conceptual progress. *Instructional Science, 24*(2), 125-155.

Parke, H. M., & Coble, C. R. (1997). Teachers designing curriculum as professional development: A model for transformational science teaching. *Journal of Research in Science Teaching, 34*(8), 773-789.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), 921-958.

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). The WHIRL co-design process: Participant experiences. *Research and Practice in Technology Enhanced Learning, 2*(1), 51-74.

Penuel, W. R., & Yarnall, L. (2005). Designing handheld software to support classroom assessment: An analysis of conditions for teacher adoption. *Journal of Technology, Learning, and Assessment, 3*(5), Available from http://www.jtla.org.

Porter, A. C., Floden, R. E., Freeman, D. J., Schmidt, W. H., & Schwille, J. R. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (Vol. 1). Reston, VA: National Council of Teachers of Mathematics.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist, 27*(1), 91-109.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reiser, B. J., Spillane, J. P., Steinmuler, F., Sorsa, D., Carney, K., & Kyza, E. (2000). Investigating the mutual adaptation process in teachers' design of technology-infused curricula. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences.* (pp. 342-349). Mahwah, NJ: Erlbaum.

Rowan, B., & Miller, R. J. (2007). Organizational strategies for promoting instructional change: Implementation dynamics in schools working with comprehensive school reform providers. *American Educational Research Journal, 44*(2), 252-297.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.

Schmidt, W. H., McKnight, C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Boston: Kluwer.

Schmidt, W. H., Raizen, S., Britton, E. D., Bianchi, L. J., & Wolfe, R. G. (1997). *Many visions, many aims: Volume II: A cross-national investigation of curricular intentions in school science*. London: Kluwer.

Schneider, R. M., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. *Journal of Science Teacher Education, 13*(3), 221-245.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

Shepard, L. (1997). *Insights gained from a classroom-based assessment project*. CSE Technical Report 451. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shkolnik, J., Song, M., Mitchell, K. J., Uekawa, K., Knudson, J., & Murphy, R. (2007). *Changes in rigor, relevance, and student learning in redesigned high schools*. Palo Alto, CA and Menlo Park, CA: American Institutes for Research and SRI International.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance. *Journal of Research on Science Teaching, 27*, 127-144.

Smith, J. P., III, di Sessa, A. A., & Roschelle, J. (1993-1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Sneider, C., Adams, S., Ibanez, F. G., Templeton, B., & Porter, J. (1996). *Evaluation study of Global Systems Science Part 1: Changing Climate*. Berkeley, CA: Lawrence Hall of Science, University of California-Berkeley.

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.

Spicer, D. E., Stark, R. B., & Hatch, D. K. (2006, 2006, April). *Detecting shifts in classroom instruction after online professional development.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies, 31*, 143-175.

Spillane, J. P., & Jennings, N. E. (1997). Aligned instructional policy and ambitious pedagogy: Exploring instructional reform from the classroom perspective. *Teachers College Record, 98*, 449-481.

Squire, K. D., MaKinster, J. G., Barnett, M., Luehmann, A. L., & Barab, S. (2003). Designed curriculum and local culture: Acknowledging the primacy of classroom culture. *Science Education, 87*(4), 467-489.

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching, 37*(2), 963-980.

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. *Science Education, 85*(2), 137-157.

Tushnet, N. C., Millsap, M. A., Abdullah-Welsh, N., Brigham, N., Cooley, E., Elliott, J., et al. (2000). *Final report on the evaluation of the National Science Foundation's Instructional Materials Development Program*. San Francisco: WestEd.

van Driel, J. H., Beijaard, D., & Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. *Journal of Research in Science Teaching, 38*(2), 137-158.

Weinstein, T., Boulanger, D., & Walberg, H. (1982). Science curriculum effects in high school: A quantitative synthesis. *Journal of Research in Science Teaching, 19*, 511-522.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: ASCD.

Wiske, S. (1997). *Teaching for understanding: Linking research with practice*. San Francisco, CA: Jossey-Bass.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*(1), 35-62.