2017

Investigating Purposeful Science Curriculum Adaptation as a Strategy to Improve Teaching and Learning

Angela Haydel DeBarger

William R. Penuel
University of Colorado Boulder, william.penuel@colorado.edu

Savitha Moorthy

Yves Beauvineau

Cathleen A. Kennedy

See next page for additional authors

Recommended Citation
DEBARGER, A. H., PENUEL, W. R., MOORTHY, S. , BEAUVINEAU, Y. , KENNEDY, C. A. and BOSCARDIN, C. K. (2017), Investigating Purposeful Science Curriculum Adaptation as a Strategy to Improve Teaching and Learning. Sci. Ed., 101: 66-98. doi:10.1002/sce.21249

Authors
Angela Haydel DeBarger, William R. Penuel, Savitha Moorthy, Yves Beauvineau, Cathleen A. Kennedy, and Christy Kim Boscardin

Running Head: SCIENCE CURRICULUM ADAPTATION

Investigating Purposeful Science Curriculum Adaptation as a Strategy to Improve Teaching and Learning

Angela Haydel DeBarger*

George Lucas Educational Foundation

William R. Penuel*

University of Colorado Boulder

Savitha Moorthy

SRI International

Yves Beauvineau

Culturally Responsive Science Pedagogies, LLC

Cathleen A. Kennedy

KAC Group Consulting

Christy Kim Boscardin

University of California San Francisco

*Co-equal first authors

Abstract

In this paper, we investigate the potential and conditions for using curriculum adaptation to support reform of science teaching and learning. With each wave of reform in science education, curriculum has played a central role, and the contemporary wave focused on implementation of the principles and vision of the *Framework for K-12 Science Education* (National Research Council, 2012) is no exception. Curriculum adaptation—whereby existing curriculum materials are purposefully modified—may provide an important strategy for teacher leaders in schools and districts to support changes to teacher practice aligned with the vision of the *Framework*. Our study provides empirical evidence that under supportive district conditions and within a research-practice partnership, purposefully adapted curriculum materials can improve student understanding of science and that these are linked to shifts teachers make in classroom culture facilitated by augmented curriculum materials.

Investigating Science Curriculum Adaptation to Improve Teaching and Learning

Improving the quality of curriculum materials has long been and remains an important strategy for changing science education practices. Ever since the Sputnik era, policy makers and educational leaders have embraced the idea that high-quality curriculum materials are needed to advance new visions for science education (Pea & Collins, 2008). Policy changes and major investments in new curriculum materials have gone hand in hand. Each major shift in aims for student learning in science often is followed by the development of multiple sets of curriculum materials intended to support those new aims (Ogborn, 2005). New curriculum materials are thought to have potential to provide teachers with models for instruction and to provide students with opportunities to learn science needed to master new goals for proficiency (Atkin & Black, 2003; Lagemann, 2002).

For local education agencies, the need for new materials that embody new visions of science teaching typically presents many challenges. It is not always feasible for these agencies to develop new curriculum materials on their own: curriculum development and revision requires specialized skills that may not be part of the repertoire of central office administrators and requires significant time and collaboration among teachers (Conley, 2003; Massell, Kirst, & Hoppe, 1997). In addition, the complex tasks of culling teacher-developed resources, validating their alignment to new policies or standards, and designing professional development to support effective use of new materials are challenging to coordinate across different district subunits (Jackson & Cobb, 2013). As a consequence, efforts to develop curriculum from scratch to meet new standards can prove disappointing, including to the district leaders who initiated those efforts (Zubrzycki, 2014).

Even when new materials are available, it may not be possible for districts to adopt them immediately. Schools, districts, and some states in the United States choose curriculum materials through formal and often highly political adoption processes that happen on specified timescales (see, e.g., Bianchini & Kelly, 2003). In the context of new standards, districts also must be cautious of investing significant resources in publisher materials that have not been adequately vetted in demonstrating student growth toward learning goals (National Governors Association, Council of Chief State School Officers, Achieve, Council of Great City Schools, & Education, 2013). Thus, even if local leaders believe curriculum is an important lynchpin of reform, they may not be able to use new materials to guide and support improvements to teaching and learning.

In the current study, we explored an alternative to curriculum design and adoption for districts, curriculum adaptation. As we use this term, curriculum adaptation refers to a purposeful effort to bring existing materials into alignment with new visions for science learning by adding to, adapting, or transforming those materials. Whether curriculum adaptation can change teacher practice and improve student learning in the diverse classrooms of a large urban district is the unanswered question that we sought to answer in this research study. Our study involved first augmenting inquiry-oriented curriculum materials for middle-grades Earth science to include enhanced supports to help teachers address a problem of practice they identified – becoming better at incorporating student ideas in instruction and linking these ideas to investigations. These supports were intended to help teachers link disciplinary core ideas to investigations they had conducted as called for in the original materials. The supports included questions teachers could ask to elicit commonly held ideas related to disciplinary content, moves for teachers to use to orchestrate discussion of those ideas, and a set of activities teachers could use when neither investigations nor discussion led to conceptual understanding. Our research investigated effects of the adapted materials on both teaching and learning outcomes.

Contemporary Science Education Reform: Why New and Redesigned Materials Now?

Since the 1960s, curricular reforms have sought to improve materials by providing direct encounters with phenomena that "help students learn to think and act like scientists" (National Research Council, 2007b, p. 13). Early reforms were based on the idea that students should gain experience with doing scientific investigations and develop a sense of the structure of scientific disciplines as conceived by practicing scientists (Bruner, 1960). In the 1990s, new wave of science education reform in the United States produced two sets of standards—the *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993) and the *National Science Education Standards* (National Research Council, 1996, 2000). This new wave emphasized the role of standards as guidelines to inform systemic reform of curriculum, instruction, assessment, and professional development (Atkin & Black, 2003; Pea & Collins, 2008). These sets of standards were adapted by individual states, and their content maintained the emphasis on inquiry and on the structure of and interconnections among disciplinary ideas of early curriculum-based reforms, but in contrast to the reforms of the 1960s that were focused on improving the preparation of future scientists, the new reform documents emphasized that scientific literacy is a goal that all students should attain (Eisenhart, Finkel, & Marion, 1996).

Science education research has advanced our understanding of ways that students' prior knowledge and initial ideas shape student learning from encounters with curriculum (National Research Council, 1999, 2005, 2007b). As a consequence, there have been strong calls by scholars and policymakers for building into curriculum materials better supports that help teachers elicit and make use of students' initial ideas to build more coherent, scientific understandings of core ideas over time (e.g., Duschl, Maeng, & Sezen, 2011; Erduran, Simon, & Osborne, 2004). In addition, more recent reform documents have called for curriculum materials that better support students' grasp of how and when to pose questions, challenge claims of others, and revise their own thinking, in order to support greater metacognition of principles of scientific reasoning (National Research Council, 2007b, p. 19). Reviews of curricula developed to support the first wave of standards-based reforms found that few sets of materials provided these kinds of supports (Stern & Ahlgren, 2002).

Today's reform efforts also emphasize the integration of core ideas and science practices in ways that most contemporary curricula do not yet reflect. That is, today's reforms present an image of science "as both a body of knowledge and an evidence-based, model-building enterprise that continually extends, refines, and revises knowledge" (National Research Council, 2007b, p. 2). Prior generations of reforms in the United States treated inquiry as separate from content and often presented principally as a means to develop conceptual understanding (Eisenhart et al., 1996). As a consequence, few curriculum materials offer students opportunities to draw on the experience of planning and conducting investigations to build and use theories, explanations, and models (Kesidou & Roseman, 2002; Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003). Contemporary reform documents therefore conclude that today's curricula will require "substantial redesign" to provide multiple and "increasingly sophisticated" opportunities to develop students' facility with the scientific practices and proficiency in using core ideas and crosscutting concepts of science to describe and explain their world (National Research Council, 2012, p. 247).

Purposeful Curriculum Adaptation as a Strategy for Materials Redesign

In this paper, we investigate whether purposeful curriculum adaptation can be an effective strategy for school districts to engage in the "substantial redesign" of curriculum. As we define it here, curriculum adaptation refers to a purposeful effort to bring existing materials into alignment with new visions for science learning by adding to, adapting, or transforming those materials. It contrasts with curriculum development, which involves the creation of entirely new sets of materials designed from their inception to align with new policy aims and visions. Using adaptation as a strategy for change presumes that curriculum is central to any reform effort and that existing curricula are likely to include some elements that align with new directions in science learning. It presumes that changes in curriculum and instruction are at least partly evolutionary rather than revolutionary, that is, they do not represent complete departures from earlier policies and visions for science education (Tyack & Cuban, 1995). To the extent that teachers are involved in the process of adaptation, it acknowledges another historical reality, namely that curricular reforms have largely failed when they have not given teachers some say in the design of curricula (Cuban, 1993).

The term "curriculum adaptation" has been in the lexicon of educational researchers for decades. One of its earliest uses was to refer to the process of adapting the course of study for young students and for immigrant and low-income students (Wrightstone, Parke, & Bressler, 1944). In science education, scholars proposed adapting materials to developmental levels of thinking hypothesized by Piaget (Shayer, 1978). Others in science education have studied how local educators have adapted materials to students by translating them into students' native language and simplifying them (Williams, 1978). More recently, a focus has been on how curriculum can be adapted to link to students' interests and experiences, so as to enhance students' perceptions of the relevance of science to their everyday lives (Tzou & Bell, 2010).

Another focus of research has been on how teachers and educational leaders adapt or tailor curriculum to local conditions. Local adaptation of materials has long been seen as a necessity (Cofer, 1952) and also as something that cannot be avoided (Berman & McLaughlin, 1975). What has shifted over time and differs by setting are the particular conditions to which curriculum materials are adapted and the degree to which adaptations are viewed as potentially benefiting (e.g., Squire, MaKinster, Barnett, Luehmann, & Barab, 2003) or as limiting student learning opportunities (e.g., A. L. Brown & Campione, 1996; Zangori, Forbes, & Biggers, 2013). At different points in time, policymakers and researchers have emphasized either the need to promote fidelity of implementation (e.g., Century, Rudnick, & Freeman, 2010; O'Donnell, 2008) or the need for principled adaptation of materials (e.g., Borko & Klingner, 2013; Singer, Krajcik, Marx, & Clay-Chambers, 2000).

The term, as we are defining it, does not appear in science education research until the late 1990s, when researchers used it to characterize a phase of a professional development cycle focused on preparing teachers to use new project-based units in science (Marx, Freeman, & Krajcik, 1998). In that context, teachers collaborated with researchers to develop units that were then spread—through professional development—throughout a large urban district (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000). Recognizing that new materials implied major shifts in teachers' beliefs and practices, researchers organized professional development for teachers new to the project in a cycle that began with collaborative analysis of the materials with peers, followed by enacting materials with students and reflecting on their enactment with colleagues in a follow-up workshop. Also as part of the follow up workshop, researchers guided teachers through a process of purposeful adaptation of materials for the next time they enacted them with students.

More recent studies also provide evidence for the promise of curriculum adaptation. In a study of augmented elementary science curriculum materials, Shutt and colleagues (Shutt, Phillips, Vye, Van Horne, & Bransford, 2010) enhanced a science kit, the Isopod Habitat Challenge, to include more opportunities for student questioning and reflection on their ideas. In the curriculum adaptation students had greater agency and choice in the questions they would investigate than in the original kit. Researchers compared the performance of students in seven classrooms where teachers implemented the augmented units with the performance of students in six comparison classrooms that implemented the traditional kits, using a quasi-experimental design. Students' scores on a content assessment were comparable across the two groups, and students who took part in the enhanced activities generated better questions to guide their investigations and were more skilled at planning a follow-up investigation to the one they did in class (Shutt et al., 2010). In addition, a study of small group discussion revealed that students were more likely to engage in the science practice of argumentation in the classrooms where the augmented materials were in use (Shutt, Vye, & Bransford, 2011). Studies of other adapted curriculum units within the same district show that students in classrooms where teachers are implementing the units exhibited more positive affective responses toward science than teachers implementing the original units (Morozov et al., 2014).

There is also promising evidence for the approach at the middle school level. Scholars from the Twenty-First Century Center for Research and Development in Cognition and Science Instruction led the adaptation of investigation-based science curriculum materials, focusing on ways to use contrasting cases to set up a sequence of learning activity and to support students' interpret visualizations in text (Cromley et al., 2011; Schunn et al., 2010). Results of a large-scale cluster randomized controlled trial found positive but non-significant impacts of the modifications on student learning (Schunn, Richey, & Alfieri, 2011).

Penuel and colleagues took a different approach to studying curriculum adaptation in a large urban district. In their study, teachers made adaptations to curriculum materials using principles developed by researchers at TERC (McWilliams et al., 2006). The impetus for their study was to explore conditions under which teachers might play a role in the design of coherent instructional experiences for students. The researchers randomly assigned teachers to one of four conditions: (1) a curriculum design condition, in which teachers learned how to develop their own units of instruction according to principles of Understanding by Design (Wiggins & McTighe, 1998) as applied to science teaching; (2) a curriculum implementation condition, in which teachers were expected to implement an inquiry-based science unit with fidelity; (3) a principled adaptation condition, in which teachers applied the principles of unit design to adapt materials from the unit; and (4) a comparison condition. The researchers found that teachers in the principled adaptation condition were able to plan more coherent, rigorous sequences of instruction and engage students in activities that were meaningful to students (Penuel & Gallagher, 2009; Penuel et al., 2009). Students whose teachers were assigned to that condition also learned more than those in the curriculum design and comparison conditions (Penuel, Gallagher, & Moorthy, 2011).

Toward a Partnership Strategy for Curriculum Adaptation in Science

In the studies that showed positive and significant effects of curriculum adaptation on student learning, participating districts shared some common characteristics. For one, the districts had more than just adopted a curriculum: they devoted resources to supporting its implementation, either through extensive professional development, a system of coaching and instructional support, or both. In addition, during the time when these studies took place, the districts had strong leadership in science that could guide the efforts at the district level. At the same time, these districts wanted to improve their adopted materials in some way, either to align better with learning sciences research (in the case of the Isopod Habitat Challenge project) or to align better with the district's approach to unit design and organization (in the case of the TIDES study).

In addition, the interventions were *co-designed* (Penuel, Roschelle, & Shechtman, 2007) within researcher-practitioner partnerships. Rather than researchers, teachers and district leaders working separately, the co-design processes made productive use of expertise in key science ideas, strategies for eliciting and making use of student thinking in science, approaches to design coherent curriculum sequences, and local concerns of teachers and district leaders. These interventions were premised on the idea that access to these forms of expertise is critical to successful implementation of new visions for science learning across a local education agency, such as a large urban school district or county education office (National Research Council, 2015).

At the same time, in each of these contexts, the projects fell short of one key goal, namely helping teachers make use of student ideas to adjust their instruction. In the Isopod Habitat Challenge study, while teachers were successful in eliciting student ideas in discussion, they varied widely in how much they took up student questions and helped students develop them into investigable ones in the context of the overall challenge posed in the unit: What kind of environment do isopods prefer? (Harris, Phillips, & Penuel, 2012). In the TIDES study, though significant improvements to the quality of teachers' assignments were observed in the principled adaptation condition, there were no observed changes to the quality of assessments as judged by independent coders (Penuel & Gallagher, 2009). Notably, these shortcomings reflected a common finding in science education research, that many teachers have difficulty making connections among student ideas, science ideas, and the investigations that students conduct to support their learning (Roth et al., 2009; Roth et al., 2006; Ruiz-Primo & Furtak, 2007).

The Project 2061 synthesis of their reviews of middle school curricula summarize well the need for greater supports for eliciting and making use of student thinking in instruction. Specifically, that review found that questions that teachers were expected to pose to students "focused on trivial misconceptions" (Kesidou & Roseman, 2002, p. 532). Of the nine curricula reviewed, only one included questions that asked students to make predictions or explain their thinking, and none provided meaningful help for teachers to interpret student responses or specific guidance on how to make use of the information. These gaps were of concern to reviewers, because the team at Project 2061 was convinced that curriculum materials that incorporated strategies to take better account of student ideas were more likely to be effective in promoting student understanding in science (Kesidou & Roseman).

In the study we describe here, the purpose of our design efforts was to work closely with a district partner to augment its adopted curriculum in middle school Earth science, *Investigating Earth Systems*, specifically to address weaknesses identified in past research and by the district with respect to opportunities for eliciting and making use of student ideas in science teaching and connecting them to big ideas in science. Our aim was to co-design adaptations with teachers and address three key questions:

Can adapted curriculum materials support teachers in making shifts to their instructional practice to elicit and make use of student thinking?

Can adapted curriculum materials improve students' conceptual understanding of disciplinary core ideas in Earth science?

Are variations in outcomes linked to teacher use of different aspects of the augmented materials?

Approach to Curriculum Adaptation

Our efforts focused on augmenting materials with respect to support they provided to teachers for eliciting and making use of student thinking. This was a carefully considered choice that met both the district's larger goals for science reform and addressed gaps identified by earlier research on curriculum and assessment in science education. (For a more extensive description of the partnership, see Penuel & DeBarger, in press.)

The two units we adapted came from the *Investigating Earth Systems* curriculum materials. *Investigating Earth Systems* is a 10-unit middle school curriculum, funded by the National Science Foundation and developed by the American Geological Institute (AGI). All of the units are organized around 6-7 student investigations, and they include a teacher's edition with relevant science background, teaching tips, advice about how to manage materials and investigations, assessments, and online materials. We modified two of the units, replacing the embedded assessments with our own activities, as elaborated below.

We focused on improving supports for teachers in three key areas: (1) eliciting and interpreting students' ideas at the beginning and end of investigations; (2) creating a classroom culture for academically productive talk; and (3) adjusting teaching when students' difficulties in understanding could not be easily overcome. Below, we describe these augmentations and the theoretical rationale for them.

Eliciting and Interpreting Student Ideas at the Beginning and End of Investigations

A large body of research underscores the importance of eliciting student thinking related to disciplinary core ideas in all phases of instruction (National Research Council, 1999). However, without a framework for interpreting student responses in relation to learning goals, it is difficult for teachers to make use of what they learn to guide instruction (National Research Council, 2014). In our project, we adopted a *facet-based* approach to developing questions teachers could pose to students. A facet is a construction of one or more pieces of knowledge by a learner in order to solve a problem or explain an event (diSessa & Minstrell, 1998b; Minstrell & Kraus, 2005). The facets perspective assumes that, in addition to problematic thinking, students also possess insights and understandings about the core disciplinary idea that can be deepened and revised through additional learning opportunities (Minstrell & van Zee, 2003). Facet-based approaches to eliciting student thinking use research on student thinking to generate potentially productive lines of questioning, investigations, and sensemaking (Minstrell & Kraus, 2005). Facets can be used on a day-to-day basis as a framework for interpreting what students say or do in response to teacher questions and tasks.

Through empirical research conducted at the beginning of our research and development project, we identified facets related to six different topics linked to the content of two Earth science units of the curriculum we sought to augment: (1) weathering; (2) erosion and deposition; (3) patterns with the location of volcanoes, mountains, and earthquakes, (4) causes of earthquakes, mountains, and volcanoes, (5) why plates move, and (6) how plate movement affects the shape of continents and species of life on continents. For each topic, we identified goal facets associated with a disciplinary core idea in the new *Framework for K-12 Science Education* (National Research Council, 2012) and problematic facets that reflect non-normative ideas or ways of reasoning. The *Causes of Earthquakes, Volcanoes and Mountain Building* facet cluster is shown in Table 1.

Insert Table 1 about here

Drawing on the facets framework, researchers and teachers collaboratively designed two kinds of questions. Diagnostic *elicitation* questions were intended to identify aspects of disciplinary core ideas students understood at the beginning of the lesson. We also developed *reflect-and-revise* questions to check student understanding of disciplinary core ideas at the conclusion of an investigation.

As in other forms of facet-based instruction, technology played a facilitative role in collecting and aggregating student responses to these questions. In a typical facet-based approach, a teacher (or computer) may present students with a series of questions focused on students' reasoning (e.g., Levidow, Hunt, & McKee, 1991). In our project, students each had clickers or student response pads that communicated wirelessly with the teacher's computer. The teacher posed the question, and all students could respond. The clicker technology then aggregated the responses; the teacher displays the distribution of responses to the class and the shared display could be used as a launching point for discussion.

Following each facets-based clicker question, teachers posed a "spark discussion question" that demanded students to construct the reasoning behind each response option. The teacher's role was to elicit reasoning behind each answer and to position each answer as a plausible response option. An example of such a question is, "Why might someone think response option A is a reasonable answer?" Posed in this way, students who may not have chosen "A" could still volunteer an answer, and students have the opportunity to hear different perspectives on the question and debate them. Figure 1 illustrates elicitation and reflect-and-revise questions followed by spark discussion questions for the *Causes of Earthquakes, Volcanoes and Mountain Building* facet cluster.

Insert Figure 1 about here

Creating a Classroom Culture for Academically Productive Talk

Even with strong guidance as to the kinds of questions to pose students and a framework for interpreting student responses, teachers may struggle with how to orchestrate *academically productive talk* to develop students' ideas. In academically productive talk students exert effort to explain their thinking with evidence and to listen and respond constructively to others' ideas, in order to make progress in solving a challenging problem, interpreting a text, or conducting an investigation (Michaels, O'Connor, & Resnick, 2008). Studies indicate a strong link between engagement in academically productive talk and learning in a range of disciplines, including literacy, mathematics, and science (Resnick, Michaels, & O'Connor, 2010).

We augmented the district-adopted *Investigating Earth Systems* curriculum by constructing with teachers *talk moves* to employ when discussing student responses to elicitation and reflect-and-revise questions. These talk moves served a range of purposes, such as encouraging students to explain their thinking so others can understand (Furtak, Thompson, Braaten, & Windschitl, 2012; Michaels & O'Connor, 2011; Thompson, Braaten, & Windschitl, 2009). Other talk moves supported teachers inviting students to add on to an earlier contribution from other students, agree or dispute claims made by another student and provide reasons for their position, and restate what another student has just said in their own words (for a more extensive list and treatment, see Michaels & O'Connor, 2011; National Research Council, 2007a). Our conjecture was that when integrated into a coherent sequence of instruction, these talk moves could help teachers cultivate a culture in which the class takes collective responsibility for advancing understanding of focal ideas (Michaels & O'Connor, 2011; van Zee, Iwasyk, Kurose, Simpson, & Wild, 2001).

To advance learning in science, the norms of talk need to align with particular norms of scientific practice (Driver, Newton, & Osborne, 2000; Duschl & Osborne, 2002; Osborne, Erduran, & Simon, 2004). To do so requires cultivation of explicit, shared norms for talk that reflect some of the ways scientists articulate, evaluate, critique, and revise knowledge claims (Bell, 2002; Bricker & Bell, 2008; A. Brown & Campione, 1994; Duschl & Osborne, 2002; Magnusson, Palincsar, & Templin, 2004; Sandoval & Reiser, 2004). In addition, students need to believe they have a stake in the argument and a strong desire to contribute to collective knowledge building (Calabrese Barton & Tan, 2009; Oliveira, 2010; Oliveira, Akerson, & Oldfield, 2012; Radinsky, Oliva, & Alamar, 2010). Our conjecture was that presenting, referring regularly to, and embodying norms of scientific practice would support student learning.

To facilitate the development of supportive norms and practices for orchestrating academically productive talk, we augmented the curriculum by introducing explicit social norms (Yackel & Cobb, 1996). These norms were co-developed with teachers and introduced in the form of a poster that teachers periodically reviewed with students that focused on “thinking, talking, and acting like a scientist.” The norms articulate a set of expectations for students to follow, and teachers use the poster as a means to periodically remind students of norms, especially during class discussions. The first norm, “everyone participates,” expresses the expectation that all students would respond to questions posed by the teacher and that students should take a risk in exposing their thinking to the class, even if they are unsure of an answer.

The second norm, “support claims with evidence,” was intended to support the development of a culture of scientific argumentation. Students were warned they could expect follow-up questions from the teacher or fellow students to justify any claim or answer they gave to a question. A third norm, “challenge ideas but respect the person,” called on students to challenge each other without putting other students down, so that argumentation takes place in a climate of mutual respect. A fourth norm, “revise and rethink often,” expressed the idea that it is all right to be wrong and that in science, students should expect to need to revise their thinking about a particular topic.

Adjusting Teaching in the Face of Persistent Student Difficulties in Understanding

A persistent challenge for teachers is making adjustments to teaching when students’ difficulties persist, even after being presented with opportunities to engage with science phenomena. Many curricula simply fall short in offering students opportunities to make sense of their investigations (Kesidou & Roseman, 2002). In Earth science, scholars have found that even when students build physical models that are rich and accurate analogs to the Earth systems they are intended to represent, curriculum materials often provide limited support for teachers as to how to help students connect their models to the processes they are intended to model (Rivet & Kastens, 2012). Thus, curriculum may need to be adapted in ways that help them to map different components and processes of physical models to the components and processes of conceptual models of the natural systems (Kastens & Rivet, 2010; Kastens, Rivet, Lyons, & Miller, 2011).

In our project, co-design teachers resonated with this concern, and so we augmented the curriculum to address student difficulties that persist even after the completion of investigations and discussions using revise-and-reflect questions. We developed, as part of the project, a total of 19 additional activities for the units for this purpose, focusing primarily on addressing problematic facets identified in the earliest stages of our research. For example, the *Earthquake Weather* activity was designed to challenge students' belief that there is a correlation between weather and earthquakes. Other activities targeted an aspect of a goal facet that was covered in a more limited way in the curriculum. For instance, the *Constructive and Destructive Forces* activity requires students to integrate their understanding of erosion and weathering with their understanding of mountain building at convergent continental plate boundaries. The *Mantle Convection* activity engages students in further discussion about the complex process of convection, which requires an integrated understanding of concepts such as density, heat transfer, and movement of solids to make sense of why plates move.

Each activity was organized into the same sequence of activities intended to engage students in constructing and using models. The models are focused on explanations of phenomena that are presented to students, inspired by the idea that model-based reasoning should be focused on puzzling phenomena (Windschitl, Thompson, & Braaten, 2008). In Step 1, students were prompted to recall and discuss the key components of and functions of models in Earth science. In Step 2, the teacher presented a model of a phenomenon (e.g., a diagram of convection currents in a cutaway of the interior of the Earth) or image and guides the students' review of it by facilitating a discussion about what is represented in the model. This step was intended to elicit and challenge students' problematic ways of reasoning about phenomena. In Step 3, students worked in small groups to use the model to explain and interpret the phenomenon further. Groups were asked to describe their observations about the model and explain the phenomenon represented in the model or the processes that might have resulted in the outcome or event that is represented in the model. In Step 4, groups shared and discuss their interpretations with the class.

Finally, in Step 5, students extend their understanding of the model by representing the process in the model in a new way (e.g., in a cartoon or storyboard) or applying the model in a new context to make a prediction.

Testing the Potential of Adaptation: Study Design

After two years of iteratively refining the adaptations to the curriculum described above through design research, we conducted a field trial to estimate the impact of these adaptations on student learning. Conditions in the district were not optimal for random assignment, so we conducted a quasi-experimental study. We recruited teachers with the help of district curriculum leaders, and teachers volunteered to be part of a treatment condition or a comparison condition.

Teachers in both conditions implemented the same units of *Investigating Earth Systems* curriculum, but teachers in the treatment condition also implemented the augmentations in the focal unit (*Dynamic Planet*) of the curriculum we chose to study more intensively. This unit is comprised of seven different investigations into how Earth’s interior shapes Earth’s surface. In the district pacing guides, teachers are allocated a range of 35 to 41 class periods to teach the unit. Teachers in both groups were allocated the same amount of time to implement their units.

Professional development was an integral part of the curriculum adaptation. Teachers participated in a three-day workshop co-led by researchers and teachers on the design team that introduced new teachers to the changes to the curriculum materials. In addition, teachers participated in teleconferences every two weeks while they were teaching the unit, where they shared implementation challenges and strategies for overcoming them with peers and the research team. During these calls, teachers received feedback from researchers and design team teachers related to the integrity of their implementation (as could be inferred from their reports on the teleconferences).

Sample

Below, we describe characteristics of teachers who volunteered and for students who completed assessments as part of the impact study.

Teachers. A total of 19 teachers were in the study, which included 12 teachers who implemented the adapted components, and 7 comparison teachers. All teachers volunteered to be in either the treatment or comparison conditions. The intention had been to recruit equal numbers of treatment and comparison conditions; however, fewer teachers volunteered to be part of the comparison condition.

The teachers overall were an experienced group, with an average of 17.2 years teaching, including 12.4 years as science teachers. In addition, all teachers had taken at least one course as undergraduates or graduate students in which they covered the topics covered in the target units. Only one teacher in the study had used clicker technology prior to becoming part of the project. There were no significant differences in background and preparation of treatment and comparison teachers.

Students. A total of 577 students completed both pre- and post-unit assessments of student learning for the *Dynamic Planet* unit. For each teacher, the research team selected one classroom of students was selected to score assessment data, the same class period for all teachers in the study. Scores of a total of 418 students in the treatment teacher group were compared to 159 for the comparison group. The treatment group had significantly higher proportion of White (32% vs. 8%) and lower proportion of Hispanic students (46% vs. 61%) compared to the comparison group. Overall, the proportions were similar between treatment (55% girls) and comparison group (54% girls) with respect to gender.

Sources of Data

Two primary sources of data were used to address the research questions: videotaped classroom discussions and student learning assessments developed and tested by the research team.

Coding of Video-recorded Classroom Discussions. The primary source of data for analyzing the project's impact on classroom practices of eliciting and developing student thinking was video recording of classroom discussions collected from both treatment and comparison group teachers. We randomly assigned teachers in both the treatment and comparison group to video record three classroom periods of instruction one of two units, including the *Dynamic Planet* unit that is the focus of the current analysis. We asked teachers to video record lessons when adapted components were most likely to be used: at the beginning of an investigation, at the conclusion of an investigation, and during an activity that a teacher decided to implement to address students' persistent difficulties with understanding concepts. We gave all teachers in both groups these instructions; we did not prescribe specific lessons for them to submit. We provided teachers with video cameras and instructions on how to set up cameras to best capture whole class teaching. In addition, the fourth author provided in-class support for video recording to teachers who requested it.

Table 2 shows the number of videos that teachers submitted, broken down by condition, curriculum unit, and type of activity. Overall, there were more videos for the treatment condition than for the comparison condition. They may have had more reason to submit videos to us, because of the materials they received through the study. Across both conditions, teachers submitted more videos for the focal *Dynamic Planet* unit than for the second unit. For this reason, we focused on this unit for our analysis.

Insert Table 2 about here

To analyze the digital video files teachers submitted, we employed a conceptually driven approach to coding teachers' actions in whole class discussions that provide evidence of the extent to which implementation was consistent with the adaptation goals that we highlighted as part of professional development. In terms of fidelity to the goals of the program, we examined videos of whole-class discussions for the extent to which teachers made explicit use of norms to for how students should participate to create classroom culture for academically productive talk and for evidence of talk moves that to promote discussion. The codes we report here reflect talk moves emphasized in our professional development. They are consistent with and represent a subset of the codes applied by O'Connor and Michaels (2011) and defined as "high leverage" talk moves, because they push students to make their reasoning explicit, to build on one another's ideas, and hold one another accountable for building knowledge together. Michaels directly advised the study team in developing the codes to be consistent with her earlier analyses of talk moves.

To facilitate coding, we developed transcripts for each video. We focused the analysis on whole-class discussions and teacher moves for practical reasons: the audio for the teacher in these situations was more readily audible for all files that teachers submitted. In each video, we coded each teacher utterance with one or more talk move codes. We did not focus on turn-by-turn counts of talk moves, in part because the episodes we coded varied widely in length (from roughly 5 to 15 minutes), and so a fair comparison across groups could not easily be made. In addition, defining the boundaries of a turn can be tricky in classroom discourse, because students and teachers interrupt one another, and there is often overlapping talk (Cazden, 1988). Therefore, we counted simply whether a code was present in an episode, for purposes of analysis of our coded data. A limitation of our approach is that by excluding small group interactions and not coding for student turns of talk, we missed being able to characterize teachers' responsiveness to students in some important ways. At the same time, focusing on teacher moves allowed us to analyze how they invoked norms and used talk moves as supports for building a culture of argumentation in their science classrooms.

Table 3 below presents the categories and definitions of codes analyzed in this article. A more complete elaboration of the definitions and examples appear in the Appendix. All examples come from the video records collected as part of the study.

Insert Table 3 about here

We developed this coding scheme through an iterative process: the team developed initial definitions for the codes, then applied them to a subset of videos, and refined the definitions to address questions and incorporate insights generated by our initial coding experiences. Once we were satisfied with code definitions and had identified examples from the videos to illustrate the codes, one of the authors developed the coding guide and distributed videos among four different coders.

To gather reliability evidence for our scheme, we identified seven videos that all coders rated independently at various points during the coding period. On the basis of our coding, we computed two different reliability estimates. The first was an overall average reliability across codes. The average inter-rater reliability for coding, determined by percent agreement, is 89.8 percent with agreements for the individual lessons ranging from 73.7 percent to 95 percent. We also computed a more conservative measure of inter-rater reliability, establishing percent agreement only for the instances when at least one member of the pair of coders had indicated the presence of a discussion move. The inter-rater reliability determined in this fashion is 76.4 percent overall, with agreement for the individual lessons ranging from 43 percent to 90 percent. Disagreements—that is, differences in the codes applied and divergences in the supporting evidence—were resolved through discussion among the research team.

Student Learning Assessments. We developed an assessment to measure student learning for the focal unit in the study. We aimed for it to be completed in a 45-minute class period. The development plan was to identify about 50% more items than required for the pilot study (conducted in 2009-2010), with only the highest performing items retained for the field trial conducted in the subsequent school year that is the focus of this paper's analysis. An attempt was made to produce instruments that were weighted toward constructed response items to elicit students' explanations about Earth science concepts, with multiple-choice items to elicit particular facets of reasoning. A total of 13 constructed response and 15 multiple-choice items were initially identified for the *Dynamic Planet* (DP) unit. To evaluate all of the items, two forms were developed.

The objective of the pilot study was to identify the most useful items to include in the final forms of the assessments for summative purposes. We looked for items that fit well together to measure the targeted construct with sufficient breadth, assessed an appropriate range of abilities, and did not exhibit favoritism or bias. For the constructed response items we additionally looked for items that would elicit different aspects of argumentation (e.g., evaluating claims, weighing alternate explanations of a phenomenon) and for items that raters could score with strong agreement. An example of an item targeting this practice is shown in Figure 2. The scoring rubric awarded more points for a response with an accurate claim that coordinated evidence presented with underlying theory or model, in this case, including reasoning that a volcanic eruption can quickly change the features on Earth's surface.

In fact, a substantial amount of time was invested in developing scoring guides for each constructed response item. The guides aligned levels of student performance (i.e., scores) with expectations expressed in the student model variable definitions and appropriate standards. Then, during facilitated scoring sessions, scoring guides were refined as needed to clarify expectations to the raters, and item prompts were modified to clarify expectations to students in subsequent test administrations. Item response theory (IRT) methods were also employed to produce evidence that items performed as theorized, and as needed for study purposes (e.g., threshold analysis).

Insert Figure 2 about here

To accommodate instruments with a mixture of dichotomous and polytomous scoring we utilized a Rasch-based (Rasch, 1960) partial credit (Masters, 1982) unidimensional model to analyze the items of each instrument. We also utilized a 3-facet model to evaluate the extent to which raters were comparatively lenient or harsh and the degree of variability in item difficulty attributable to rater differences for the constructed response items as an indication of the performance of the scoring guides to produce consistent ratings.

Our pilot data provided us with evidence that the assessments were sensitive to the effects of instruction and could discriminate between low and high levels of proficiency. Instructional sensitivity is an important property of a test intended to measure effects of a treatment; a test that is "insensitive" to instruction is not one that would be able to yield evidence related to the impacts of a treatment that is focused on changing instruction (Polikoff, 2010). In the pilot, the test performed well for assessing pre-to-post change, producing statistically significant pre-post differences in Weighted Likelihood Estimates (WLE; Warm, 1989) of proficiency ($t(1192) = 18.42$, $ES = 1.07$, $p < .01$). The ability of a test to differentiate high and low levels of proficiency is important as well when evaluating a treatment, since a test that does not do so will not likely produce enough variability to estimate whether variation is due to chance or the effects of a treatment. Item discrimination indices provided evidence of consistent differentiation of students into high and low performing groups. Indices for constructed response items ranged from .52 to .73 and for multiple choice items ranged from .21 to .55. The item difficulties on the test assessed students with median or higher abilities better than students with lower than average abilities. The test information functions also indicated that each test assessed students at the mean posttest ability better than students at the mean pretest ability. This is consistent with the tests being more reliable measures of posttest abilities than of pretest abilities. The output of the pilot study was two sets of items for the field test: 7 constructed response and 12 multiple choice items for the DP assessment.

Trained raters scored the constructed response items in facilitated scoring sessions that included initial consensus building and drift checking to assist in maintaining scoring consistency. The average measure intra-class correlation coefficient (ICCAM) provided an indication of the reliability of the scores obtained from different raters for the same student response. A two-way mixed effects model was used with respondent values varying randomly and rater values fixed. We found excellent agreement among the scorers on the constructed response items, with ICCAM was greater than .90 for all constructed response items on both tests.

The test had a reliability estimate within an acceptable range (*DP:* $\alpha = .79$) (George & Mallery, 2003). All but one item on the test fit a unidimensional one-parameter item response theory (IRT) model well, with both weighted and unweighted mean squares of the items ranging between 0.77 and 1.25. Wright and Linacre (1994) indicate that the range of 0.7 to 1.3 is a conventionally acceptable range for multiple-choice items and 0.6 to 1.4 is acceptable for partial credit (e.g., constructed response) items. The outlying constructed response item on the DP test had a weighted mean square of 1.41 and an unweighted mean square of 1.64 indicating that the responses on that item were less predictable than expected for most students.

Analysis of Data

In the results section, we present three different data analyses:

1. Comparative analysis of teacher use of norms and talk moves in the treatment and comparison classrooms (Research Question 1)
2. Comparative analysis of student learning in the treatment and comparison classrooms (Research Question 2)
3. Correlational analyses of the association between teacher practices and student learning (Research Question 3)

To examine the level of enactment and the perceived value of augmented components by the treatment teachers, we provide descriptive statistics on the observation data for the focal unit of the *Investigating Earth Systems* curriculum. For the analysis, we report descriptive statistics for each of the individual codes, namely whether they were present for a given video. In addition, we report descriptive statistics for the sum of high-leverage talk moves for each of the codes presented in Table 3 above, and we present a teacher-by-teacher analysis, where we selected one video at random to facilitate interpretation of individual variation within and across the treatment and comparison groups.

To examine the impact of the adaptations on teaching and learning, we performed Generalized Estimating Equations (GEE) for the observation and student assessment data. GEEs are an extension of the generalized linear model that relaxes several assumptions of traditional regression models to accommodate correlated (clustering) of data (Zeger & Liang, 1986). That is, GEEs allow us to model data where student scores are correlated with one another within classrooms (because they share a teacher and common learning environment) and where scores are not normally distributed. GEEs are especially appropriate in our study, because they are better able than hierarchical linear models are to provide estimates of impacts when the number of classrooms and students is relatively small (Cheong, Fotiu, & Raudenbush, 2001; Ghisletta & Spini, 2004).

We performed three separate analyses using GEEs. For analyses of teaching practices, we performed separate GEEs for each of the outcomes listed in Table 3. For analyses of student learning, we performed a GEE using the post-test scores on the student learning assessments as the dependent variable and pre-test scores as a covariates in the model. To determine whether implementation of specific tools was associated with student performance, we performed GEEs using the post-test scores on the student learning outcomes as the dependent variable and two teaching variables hypothesized to be linked to scores on the test, explicit use of norms and a summary score of high-leverage moves.

Results

Teachers' Implementation of Strategies in Adapted Materials

We summarized coding data from the *Dynamic Planet (DP)* videos and compared frequencies with which teachers engaged in different practices of eliciting and developing student thinking in whole-class discussions. For the treatment group, included in these practices were coded observations of how well teachers adhered to the pedagogical patterns of the project and their use of spark discussion questions to launch classroom discussions.

In the *DP* unit, teachers were more likely to explicitly invoke the norms provided as part of the adaptation than were comparison teachers and were also more likely to use a wider variety of high-leverage talk moves. Teachers invoked norms in 80 percent of the treatment classroom lessons, but not at all in comparison classroom lessons. In addition, on average, treatment teachers used more high-leverage talk moves to elicit and develop students' thinking (4.6 different types of moves per lesson) than did teachers in the comparison classrooms (1.7 different types of moves per lesson).

Because the number of videos submitted by teacher varied, we examined individual teacher-by-teacher variation, choosing a single video at random for each teacher to analyze. We see a similar pattern of results, looking teacher by teacher. Table 4 shows teachers' use of the four norms. Notably, all five of the treatment teachers invoked at least one of the norms, and four of these teachers engaged multiple norms. Coding analyses document when norms were explicitly reinforced, for example a teacher reminding her students to explain their answers. Analyses also noted when norms were implicit but evident, for instance, when students disagreed respectfully with each other, without being prompted to do so by the teacher (evidence of challenging ideas but not the person). The teacher-by-teacher data on talk moves mirrors the findings on norms

(Table 5). Although similar percentages of teachers used the talk move, “Eliciting student thinking and reasoning,” a low percentage of comparison teachers used any of the other high-leverage talk moves, when compared to the treatment teachers.

Insert Table 4 about here

Insert Table 5 about here

The following excerpts of discussions from treatment and comparison group classrooms illustrate the differences between the two groups that the analysis above highlights. The teachers were selected, because their students had the highest pre-to-post test gains on the student outcome measure for their respective group. The students in the comparison teacher’s classroom gained 7.9 points on average from pretest to posttest, while students in the treatment teacher’s classroom gained 11.6 points on average. Both lessons take place after investigations about rocks.

The activity implemented by the treatment teacher in this excerpt, “Is Water Enough?” is intended to help students understand the mechanisms that cause rock to weather. A key purpose of the activity is to help students apply understandings developed through an investigation of conditions that speed up the reactions that cause iron to rust to a new situation, where students are asked to explain patterns of weathering they see in a photograph. The activity begins when the teacher shows students an image of a rock outcropping in the City of Rocks National Reserve in Idaho. The picture shows a number of large rocks, some jagged and some smooth, in the foreground. In the first part of the activity, students work in small groups to answer two questions about the picture: (1) What evidence of rock weathering can you see in this picture? (2) What evidence tells you that the rocks have weathered at different rates?

What follows the opening part of the activity is a discussion that the teacher orchestrates with students about the factors that shape the rate at which weathering occurs. In the materials we provided her as a treatment group teacher, she has available to her an unusual kind of "key" that is intended to help her listen for different kinds of answers. The key defines a range of correct answers, and it also indicates that students may know that weathering can occur when water can split apart a rock but may not know that temperature and the composition of the rock can both affect the rate of weathering. As the transcript below shows, the teacher both elicits divergent viewpoints while also guiding students toward the range of reasons why the rocks in the City of Rocks National Reserve are weathering at different rates.

In this first excerpt, we see the teacher asking students to restate what others have said, posing questions that elicit student thinking and reasoning, and asking follow up questions of students, using multiple talk moves. In this transcript excerpt, pseudonyms indicate students, and words given emphasis by the speaker are in capital letters. When a speaker interrupts another, we indicate this by an "=" at the end of the turn of the speaker being interrupted and the beginning of the speaker who interrupts. In the third column, we indicated how we coded each talk move. In addition to noting the high-leverage talk moves described in Table 3, we also identify talk moves that are in our coding guide and that were important in shaping the discussion, such as when the teacher restated a student idea or asked students to restate a peer's idea, as well as the teacher's use of low inference questions (in contrast to questions that elicit thinking and reasoning).

Kevin	OK. There could be DIFFERENT types of rocks, SOME of which weather faster than OTHERS. For EXAMPLE, SANDSTONE is weaker than GRANITE, so it would WEATHER faster. SOME rocks could be exposed to more SUNLIGHT, therefore more PLANTS can grow there. If PLANTS root in rocks, the PRESSURE from the roots could break apart CERTAIN rocks more than OTHERS.	
Teacher	OK. So, Kevin, BEFORE you add anything ELSE, what FACTORS did you HEAR from this GROUP that could affect the RATE of weathering? What's one, KAREN?	Talk Move: *Asking students to restate someone else's reasoning*
Karen	The amount of SUNLIGHT?	
Teacher	And HOW would amount of SUNLIGHT / be a FACTOR?	Talk Move: *Follow up question*
Karen	Well if like in their EXPERIMENT, the RUST, if the rock has IRON in it and it RAINS a lot there and it gets a lot of SUNLIGHT, then it might, it might OXIDIZE more than if there is a ROCKS where it RAINS a lot but doesn't get a lot of SUNLIGHT.	
Teacher	OK. Now how does sunlight AFFECT how much rusting? Cause we definitely saw… Um, well, WHAT factors did we SEE affecting on to oxidation or the RUST with the STEEL wool? TIA?	Talk Move: *Eliciting thinking and reasoning*
Tia	Well…I guess like if when it's in the SUN and it's WET and like the rock is WET and it has IRON in it, it RUSTS faster?	
Teacher	Was there any=	Talk Move: *Follow up question*
Tia	=I don't know…	

A few minutes later, the teacher tries to bring the class to consider how temperature might affect oxidation. She reminds students of what they varied in their experiment with steel wool – both dampness and warmth. She asks them to consider the pattern of results, namely that warm steel wool rusted faster, and then asks what might explain that pattern. In the excerpt below, we see the teacher both restate the reason one student, Ian, gives for the pattern of results, and engage in a revoicing move, where she opens a speaking slot for Ian to respond to her own rephrasing of his answer:

Teacher	OK. So it seems like a WARM, wet environment is more conducive to RUST being formed than a COLD wet... So more SUN in some of those AREAS might INCREASE the amount of OXIDATION. IAN?	Talk move: *Summarizing*
Ian:	Well the HEAT from the sun could ALSO um LIMIT the rust because it would EVAPORATE a lot of the MOISTURE making it HARDER for the…if it contained IRON to oxidize.	
Teacher	So good POINT. So depending on the situation it MIGHT help or…	Talk move: *Teacher restatement*
Ian	If there is too MUCH it probably would do it. But if was actually just WARM not to the point of too much EVAPORATION…	
Teacher	OK. So maybe DEPENDING on how much the sun INCREASES the temperature in THIS area compared to what it was in another AREA?	Talk move: *Revoicing*
Henry	Yeah.	

This part of the class discussion concludes with an episode in which the teacher presses the students to listen to one of the students, Kevin, who has suggested that sunlight affects the rate of weathering. Our interpretation is that the teacher is aiming to do two things simultaneously: (1) to draw attention to a facet of weathering that students had not yet surfaced but that was in the curricular guidance we provided, and (2) to encourage students to attend carefully to one another's ideas. She presses students to listen, even when some students admit to not hearing their peers' response, as this excerpt illustrates:

Teacher	OK, so who could put in their OWN words what KEVIN was just saying about how the amount of SUN could affect WEATHERING in ANOTHER way. OK, Beth, can you put that in your OWN words?	Talk Move: *Asking students to restate someone else's reasoning*
Beth:	[Inaudible, due to poor video quality]	
Teacher	OK, REALLY important to LISTEN and be thinking about what everyone is SAYING. OK, what did he SAY about why the amount of SUN make a difference? Tamara?	Talk Move: *Asking students to restate someone else's reasoning*
Tamara:	I wasn't listening	
Teacher	OK, Peter?	
Peter:	SUN can make water WARM and if there isn't the same amount of SUNLIGHT or DARKNESS then the temperatures will RISE or the stay the SAME and sometimes the oxidation WON'T stay/ won't go at the same SPEED because of SUNLIGHT.	
Teacher:	OK. So THAT'S a good summary of what we were talking about JUST before, but not with the PLANTS. That relates to the OXIDATION—and THAT makes sense—but he MENTIONED something to do	Talk move: *Teacher Restatement*

	with SUN and PLANTS and how THAT could affect the rate of weathering. Carol?	
Carol:	The SUN, since there's MORE sun, so PLANTS could grow and then the PLANTS of the plants could go up into the rocks and CRACK them?	
Teacher:	OK. Does that sound FAMILIAR?	Talk move: *Low-inference question*
Multiple Students:	Um-hmm.	
Teacher:	Did you see any EVIDENCE of that at Stonesville [a nearby state park the students visited]?	Talk move: *Low-inference question*
Multiple Students:	Yeah	

Though the above episode concludes with what we coded as low-inference questions, it illustrates the ways that the teacher uses the discussion prompts supported by the adaptations to the curriculum materials to connect the investigation students did in class, a field trip to a nearby state park, and the key ideas related to the factors that affect weathering of rock in a given location.

The discussion in the comparison classroom also occurs at the conclusion of the investigation about how rocks are formed. In this excerpt, shown below, the teacher facilitates the discussion, drawing from the guidance provided in the Review and Reflect section of the *Investigating Earth Systems* curriculum. Here, the teacher launches the discussion by posing a low inference question, "What's one type of rock?" Fewer high-leverage talk moves are in evidence in this excerpt (and our analysis suggests, typically in comparison classrooms), and student responses tend to be brief, demonstrating discrete declarative knowledge rather than an articulation of thinking and understanding. The teacher also evaluates students' responses in the course of the discussion, correcting problematic ideas as and when they surface.

Teacher	Talking about TYPES of ROCKS. Tara. What's ONE type of rock?	Talk move: *Low-inference question*
Tara	Igneous?	
Teacher	Igneous. IGNEOUS is a type of rock. What ELSE? What are the…	Talk move: *Low-inference question*
Student	Magma.	
Teacher	Magma isn't a ROCK. Magma COOLS to MAKE a rock, but it's NOT a rock. Lorenzo.	Talk move: *Low-inference question*
Lorenzo	Sedimentary?	
Teacher	SEDIMENTARY. So we've got IGNEOUS. We've got SEDIMENTARY. What ELSE? Victoria.	Talk Moves: *Teacher summarizes student responses, low-inference question*
Victoria	Metamorphic?	
Teacher.	METAMORPHIC. OK. METAMORPHIC. Those are our THREE types. How are they DIFFERENT? How are they DIFFERENT? How are they the SAME? EITHER one. Mary.	Talk Moves: *Teacher summarizes student responses, low-inference question*
Mary	Each one of them is… are formed differently.	
Teacher	They're formed DIFFERENTLY. That's RIGHT. They're formed DIFFERENTLY. So do we know how SEDIMENTARY was formed. WHO can tell?	Talk Move: *Eliciting student thinking and reasoning*

Student	no	
Teacher	me. Carl.	
Carl	When smaller rocks like STICK together [inaudible]	
Teacher	Good. So when SMALL ROCKS stick TOGETHER what's another WORD for SMALL ROCKS. Nick.	Talk move: *Low-inference question*
Nick	Sediment.	

The discussion from the comparison teacher was fast-paced, and many students were successful in coming up with the definitions that the teacher asks them to recall. There is one prompt for students to consider how they would know what a rock type is, though this, too, appears to be a fact students are expected to have memorized rather than a moment that calls for reasoning about the conclusion of their investigation into rocks. In fact, the investigation does not come up at all in this discussion as a resource for students to reason about, in order to develop their understanding of different types of rock. Whereas rock type had explanatory power in the treatment classroom—to help explain why rocks weather at different rates—in this classroom rock types are mainly definitions to be memorized.

Student Learning Effects

Our analysis of student learning effects indicated that there were significant differences between treatment and comparison classrooms (see Figure 3). Students in the treatment classrooms (M_t = 13.0, SD_t = 6.6) performed higher on the 19-item test compared to the students in the comparison group classrooms (M_c = 9.1, SD_c = 4.6). At the same time, the students in the treatment teacher classrooms had a higher pre-test scores (M_t = 6.5, SD_t = 3.8) on the assessment compared to the comparison group teacher classrooms (M_c = 5.5, SD_c = 3.4). Based on the GEE analysis, we found that the students in the treatment teacher classrooms performed about three points higher on the assessment after controlling for the initial differences in the pre-test scores between the two groups and student ethnicity (p = 0.006).

Insert Figure 3 about here

Correlational Analysis Relating Implementation and Student Learning Outcomes

We performed additional analysis with the video observation data to determine whether implementation of particular strategies was associated with student learning gains. In a set of regression analyses, we analyzed whether use of norms in a lesson was associated with higher test scores and whether the number of high-leverage talk moves used in a lesson was related to learning outcomes. We found that the increase in explicit use of norms in the classrooms was significantly associated with higher student assessment post-test scores (β = .90, p = 0.03). Students in the classrooms with teachers who explicitly invoked norms scored about 4 points higher on the post-test after controlling for the initial differences in their pre-test scores. The association between gains and the use of high-leverage talk moves was positive, but it was not statistically significantly associated with student performance after controlling for the pre-test performance (β = .21, p = 0.09).

Discussion and Conclusions

In the study, we found evidence that a program with multiple, integrated components for helping teachers establish norms, pose well-designed questions, and orchestrate discussions in which students were pressed to give reasons for their ideas could improve teacher practice and increase student learning. The inquiry-based Earth science curriculum provided a good foundation for these practices, but teachers were struggling to help their students make connections between the investigations and the core ideas. The facet-based questions, targeting opportunities to elicit student ideas at the beginning and end of investigations, coupled with supports to promote productive academic talk helped to surface and develop student reasoning. As illustrated in the transcript excerpts, when teachers became practiced with both the norms and the talk moves, they flexibility interwove these moves in ways that elicited student thinking and reasoning and made connections between core ideas and investigations.

Importantly, we demonstrated the value of adapting existing materials, when adaptations are accompanied with professional development. From our perspective, the professional development was an integral component of the adaptation, not separate from it. It formed one component of what others have called a "curricular activity system" (Roschelle, Knudsen, & Hegedus, 2010), that is, a coherent set of instructional materials for teachers, student learning and assessment activities, and professional development plans organized around a specific model of learning. Of course, it might be useful to investigate systematically the contributions of different components to the potential teaching and learning gains, but our study design did not permit us to do so. A district leader might want to know which components to emphasize or how much professional development is necessary. While we cannot answer such questions, our correlational analysis finding that the use of norms by teachers related to outcomes does, however, suggests one potentially important component future research should investigate.

It was puzzling to us that teachers' use of talk moves was not associated with greater gains among the Contingent Pedagogies teachers. We would have expected on the basis of prior research (e.g., research reported in Resnick, et al. 2010) to find a strong relationship between the use of talk moves and learning gains. Our sample transcripts, though, illustrate a more global quality of the most effective Contingent Pedagogies classrooms that other scholars have termed "responsiveness" (Coffey, Hammer, Levin, & Grant, 2011; Elby et al., 2014; Pierson, 2008) and that may account for differences in learning gains. A teacher is responsive when there is evidence of careful attention to students' ideas and questions; the teacher may share their own thinking, but it is in the service of making sense of students' ideas (Pierson, 2008). Coffey and colleagues (2011) call this careful attention to students' ideas and reasoning "the heart of the matter for formative assessment" (p. 1112) and caution that strategies (e.g., talk moves) should be in the service of this broader attention. It is possible that less successful Contingent Pedagogies teachers used talk moves, but not in the service of attending to students' thinking and reasoning.

We were similarly surprised by the power of relatively simple reminders of norms in partly explaining our learning gains. On the one hand, we designed the norms poster and introduced it to teachers because teachers thought it would help to create a positive culture of academically productive talk in their classrooms. Moreover, past research in other fields led us to believe that social norms centered on disciplinary forms of reasoning could support students' learning (e.g., Yackel & Cobb, 1996). We thought that our design was relatively weak, however, until we saw how some teachers took up the poster and integrated mentions of norms seamlessly into their classrooms to get students to participate, to defend their ideas, and revise their thinking. The norms proved to be a persistent reminder of the purposes of talk in a science classroom. Future research should consider how such relatively simple additions to practice might support not only students' engagement in talk but also help the teacher become more responsive to student ideas in ways that can facilitate student learning.

Limitations of the Study

One limitation of our study was the amount of missing data from classroom observations. We did not collect data from all of the teachers for the focal unit in this analysis, and teachers did not provide us with all the data we requested. Therefore, analyses of the teaching data, including correlations with outcomes, should be treated as useful for generating hypotheses rather than as conclusive of the value of our approach for supporting changes to teaching.

A second limitation of our study was that even though the project was a district-level initiative, it was limited in scope to a single grade level and did not include all teachers in that grade level. The current study might best be viewed as an important early test of and approach that would need to be replicated at a larger scale, before it could be judged adequate to producing reliable improvements across a wide range of settings (see Bryk, Gomez, Grunow, & LeMahieu, 2015). In addition, further studies would be needed that fully explore the conditions that are required for adaptations to be successful and sustainable. As we noted in our introduction, we began with some conjectures about those conditions, including having strong district leadership and a partner with expertise in research on student learning to support the design and testing of curriculum adaptations.

Still, demonstrating that adaptations focused on eliciting and making use of student thinking can yield improvements to student learning at scale has heretofore remained elusive. Earlier related efforts found wide variation in implementation and limited impacts on student learning (Furtak et al., 2008; Yin et al., 2008). By contrast, our study found significant—if modest—effects on student learning. We view findings as particularly significant, given the limitations of most current curricula and the need to strengthen specifically components related to sensemaking about phenomena students have investigated and opportunities to construct and use models to support predictions and explanations.

Implications

This study is important for demonstrating how relatively small, but purposeful curriculum adaptations, in this case targeting the elicitation of and responsiveness to student ideas, can make a difference in student learning outcomes. Adaptions can be described as purposeful because they were grounded in research, informed by practitioners' problems of practice, and iteratively refined with input from teachers. While the scope of the adaptations was fairly focused on developing productive elicitation and discussion techniques that were situated within a specific disciplinary context, the notion of purposeful reflected the perspectives of key stakeholders involved in implementing the adaptations.

It is important to note that while our study took place just before the latest wave of reforms in science education in the United States, the adaptations we tested involved significant revisions to an existing inquiry-oriented curriculum to strengthen connections between students' understanding of disciplinary core ideas and student-led investigations already included in the materials. As such, it embodies one key element of the *Framework for K-12 Science Education* (National Research Council, 2012), namely the integration of disciplinary core ideas with science practices. The facet clusters and corresponding questions inform how core ideas can emerge in classroom conversations, helping teachers to anticipate the problematic ideas as well. Importantly, these questions coupled with talk moves can engage students in applying these core ideas through the practices of argumentation, constructing explanations and developing models. While these curricular resources would require revision to fully align to the NGSS, the approaches applied in this study would likely be generalizable to other curricular and disciplinary contexts aiming to meet the NGSS.

Although the *Framework's* emphasis on how student understanding develops over time was not reflected in our adaptation approach, our findings have potential relevance to efforts to adapt curriculum materials that are currently available in ways that reflect the hypothetical learning progressions outlined in the *Framework* and in the NGSS (see especially Appendix E, "Progressions within the NGSS"). Learning progressions are testable, empirically supported hypotheses about how student understanding develops toward specific disciplinary goals (Duschl et al., 2011; National Research Council, 2007b; Smith, Wiser, Anderson, & Krajcik, 2006). A number of researchers have used learning progressions to guide curriculum development (Fortus & Krajcik, 2012; Parker, de Los Santos, & Anderson, 2015; Wiser, Smith, & Doubler, 2012). Future research might investigate how teachers might adapt existing materials to reflect hypothetical learning progressions. At present, one of the authors is engaged in just such a collaborative effort, building on research that emphasizes the need for coherent storylines to guide science teaching (Reiser, 2014; Roth et al., 2009).

All of these efforts are more likely to be successful, we argue, when conducted in partnerships between researchers and practitioners (Penuel, 2015; Penuel & DeBarger, in press). We attribute the success of implementation and outcomes in part to the close ties we maintained with district leaders to ensure the fit of the treatment within district constraints and to the involvement of teachers in helping make materials more usable and accessible to teachers. Today's districts are not only limited in resources; they provide myriad sources of guidance to teachers that must be coordinated to ensure implementation effectiveness (Bryk et al., 2015; Hopkins & Spillane, in press). Future research on curriculum adaptation may need to contend with ever expanding sources of guidance to teachers about what and how to teach, and not just from standards. This, no doubt, complicates the task but underscores the need for continued research on the dynamics and outcomes of purposeful curriculum adaptation as a strategy for improving teaching and learning.

References

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

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

Bell, P. (2002). Using argument map representations to make thinking visible for individuals and groups. In T. Koschmann, R. Hall, & N. Miyake (Eds.), *CSCL 2: Carrying Forward the Conversation* (pp. 449-485). Mahwah, NJ: Lawrence Erlbaum Associates.

Berman, P., & McLaughlin, M. W. (1975). Federal programs supporting educational change, Volume 4: The findings in review. Santa Monica, CA: RAND.

Bianchini, J. A., & Kelly, G. J. (2003). Challenges of standards-based reform: The example of California's science content standards and textbook adoption process. *Science Education, 87*(3), 378-389.

Blumenfeld, P., Fishman, B. J., Krajcik, J., 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.

Borko, H., & Klingner, J. K. (2013). Supporting teachers in schools to improve their instructional practice. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds). *Design-based implementation research. National Society for the Study of Education Yearbook, 112*(2), 298-319.

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for practices of science education. *Science Education, 92*(3), 473-498.

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice*. Cambridge, MA: MIT Press.

Brown, A. L., & Campione, J. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289-325). Hillsdale, NJ: Erlbaum.

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

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard University Press.

Calabrese Barton, A., & Tan, E. (2009). Funds of knowledge, discourses and hybrid space. *Journal of Research in Science Teaching, 46*(1), 50-73.

Cazden, C. (1988). *Classroom discourse*. Portsmouth, NH: Heinemann.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*(2), 199-218.

Cheong, Y. F., Fotiu, R. P., & Raudenbush, S. W. (2001). Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics, 26*, 411-429.

Cofer, C. (1952). Problems in the implementation of curriculum. In Study Group Supported by the Carnegie Corporation of New York and the Grant Foundation (Ed.), *Improving undergraduate instruction in psychology* (pp. 46-52). Ithaca, NY: Cornell University.

Coffey, J., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*(10), 1109-1136.

Conley, D. T. (2003). *Who governs our schools? Changing roles and responsibilities*. New York, NY: Teachers College Press.

Cromley, J. G., Newcombe, N. S., Wills, T. W., Wills, M., Karakus, M., & Batchelor, M. (2011). *Teaching middle school students to reason with visual representations in science*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1880-1990* (2nd ed.). New York: Teachers College Press.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312. Retrieved from http://spot.colorado.edu/~otero/Driver_Norms.pdf

Duschl, R. A., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and synthesis. *Studies in Science Education, 47*(2), 123-182.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*(1), 39-72.

Eisenhart, M., Finkel, E., & Marion, S. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal, 33*, 261-295.

Elby, A., Richards, J., Walkoe, J., Gupta, A., Russ, R. S., Luna, M., . . . van Es, E. (2014). Differing notions of responsive teaching across mathematics and science: Does the discipline matter? In J. L. Polman, E. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. R. Lee, & L. D'Amico (Eds.), *Proceedings of the 11th*

International Conference of the Learning Sciences (Vol. 2, pp. 1406-1415). Boulder, CO: International Society of the Learning Sciences.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933.

Fortus, D., & Krajcik, J. (2012). Curriculum coherence and learning progressions. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 783-798). Dordrecht, the Netherlands: Springer.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessment and its relation to student learning. *Applied Measurement in Education*, *21*, 360-389.

Furtak, E. M., Thompson, J., Braaten, M., & Windschitl, M. (2012). Learning progressions to support ambitious teaching practices. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 405-434). Rotterdam, the Netherlands: Sense Publishers.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4 ed.). Boston, MA: Allyn & Bacon.

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics, 29*(4), 421-437.

Harris, C. J., Phillips, R. S., & Penuel, W. R. (2012). Examining teachers' instructional moves aimed at developing students' ideas and questions in learner-centered science classrooms.

Journal of Science Teacher Education, 23(7), 768-788. Retrieved from http://dx.doi.org/10.1007/s10972-011-9237-0

Hopkins, M., & Spillane, J. P. (in press). Conceptualizing relations between instructional guidance infrastructure (IGI) and teachers' beliefs about mathematics instruction: Regulative, normative, and cultural-cognitive considerations. *Journal of Educational Change*.

Jackson, K., & Cobb, P. (2013). Coordinating professional development across contexts and role group. In M. Evans (Ed.), *Teacher education and pedagogy: Theory, policy and practice* (pp. 80-99). New York, NY: Cambridge University Press.

Kastens, K. A., & Rivet, A. (2010). Using analogical mapping to assess the affordances of scale models used in Earth and environmental science education. In C. Holscher, T. E. Shipley, M. O. Blelardinelli, J. A. Bateman, & N. S. Newcombe (Eds.), *Spatial Cognition VII* (pp. 112-124). Dordrecht, the Netherlands: Springer.

Kastens, K. A., Rivet, A., Lyons, C., & Miller, A. R. (2011, October). *Bridging the gap between tabletop models and the Earth system: Classroom-based research on how teachers teach and how students learn from physical models.* Paper presented at the Geological Society of America Annual Meeting, Minneapolis, MN.

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.

Lagemann, E. (2002). *An elusive science: The troubling history of education research*. Chicago, IL: University of Chicago Press.

Levidow, B. B., Hunt, E., & McKee, C. (1991). The DIAGNOSER: A HyperCard tool for building theoretically based tutorials. *Behavior Research Methods, Instruments & Computers, 23*(2), 249-252.

Magnusson, S. J., Palincsar, A. S., & Templin, M. (2004). Community, culture, and conversation in inquiry-based science instruction. In L. B. Flick & N. G. Lederman (Eds.), *Scientific inquiry and nature of science* (pp. 131-155). Dordrecht, The Netherlands: Kluwer.

Marx, R. W., Freeman, J. G., & Krajcik, J. S. (1998). Professional development of science teachers. In B. J. Fraser & K. B. Tobin (Eds.), *International Handbook of Science Education* (Vol. 2, pp. 667-680). Dordrecht, the Netherlands: Kluwer.

Massell, D., Kirst, M. W., & Hoppe, M. (1997). Persistence and change: Standards-based reform in nine states. Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McWilliams, H., McAuliffe, C., Lockwood, J., Larsen, J., Rosenberg, M., & Hoover, M. (2006). Earth Science by Design: A handbook for professional developers. Cambridge, MA: TERC.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in the Philosophy of Education*, *27*, 283-297.

Michaels, S., & O'Connor, M. C. (2011). Talk science primer. Cambridge, MA: TERC.

Minstrell, J., & Kraus, P. (2005). Guided inquiry in the science classroom. In National Research Council (Ed.), *How students learn: History, mathematics, and science in the classroom* (pp. 475-514). Washington, DC: National Academies Press.

Morozov, A., Herrenkohl, L., Shutt, K., Thummaphan, P., Vye, N., Abbott, R. D., & Scalone, G. (2014). Emotional engagement in agentive science environments. In J. L. Polman, E. Kyza, K. O'Neill, & I. Tabak (Eds.), *Proceedings of the 11th International Conference of the Learning Sciences* (pp. 1152-1156). Boulder, CO: International Society of the Learning Sciences.

National Governors Association, Council of Chief State School Officers, Achieve, Council of Great City Schools, & Education, N. A. o. S. B. o. (2013). Publishers' Criteria for the Common Core State Standards. Washington, DC: Authors.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

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

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

National Research Council. (2005). How students learn: History, mathematics, and science in the classroom. Washington, DC: National Academies Press.

National Research Council. (2007a). Ready, set, SCIENCE: Putting research to work in K-8 science classrooms. Washington, DC: National Academies Press.

National Research Council. (2007b). Taking science to school: Learning and teaching science in grades K-8. Washington, DC: National Academies Press.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Research Council.

National Research Council. (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

National Research Council. (2015). Guide to implementing the Next Generation Science Standards. Washington, DC: National Academies Press.

O'Connor, M. C., & Michaels, S. (2011). *Problematizing dialogic and authoritative discourse, their coding in classroom transcripts, and realization in the classroom.* Paper presented at the ISCAR Congress, Rome, Italy.

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), 33-84.

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.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching, 47*(4), 422-453.

Oliveira, A. W., Akerson, V., & Oldfield, M. (2012). Environmental argumentation as sociocultural activity. *Journal of Research in Science Teaching, 49*(7), 869-897.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching, 40*(7), 692-720.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Parker, J. M., de Los Santos, E. X., & Anderson, C. W. (2015). Learning progressions and climate change. *The American Biology Teacher, 77*(4), 232-238.

Pea, R. D., & Collins, A. (2008). Learning how to do science education: Four waves of reform. In Y. Kali, M. C. Linn, & J. E. Roseman (Eds.), *Designing coherent science education* (Vol. 3-12). New York: Teachers College Press.

Penuel, W. R. (2015). Learning progressions as evolving tools in joint enterprises for educational improvement *Measurement: Interdisciplinary Research & Perspectives, 13*(2), 123-127. doi:10.1080/15366367.2015.1055145

Penuel, W. R., & DeBarger, A. H. (in press). A research-practice partnership to improve formative assessment in science. In A. J. Daly & K. S. Finnigan (Eds.), *Thinking systemically: Improving districts under pressure*. Washington, DC: AERA.

Penuel, W. R., & Gallagher, L. P. (2009). Preparing teachers to design instruction for deep understanding in middle school Earth science. *The Journal of the Learning Sciences, 18*(4), 461-508.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in Earth science: A comparison of three professional development programs. *American Educational Research Journal, 48*(4), 996-1025.

Penuel, W. R., McWilliams, H., McAuliffe, C., Benbow, A., Mably, C., & Hayden, M. M. (2009). Teaching for understanding in Earth science: Comparing impacts on planning and instruction in three professional development designs for middle school science teachers. *Journal of Science Teacher Education, 20*(5), 415-436.

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning, 2*(1), 51-74.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3-14.

Radinsky, J., Oliva, S., & Alamar, K. (2010). Camila, the earth, and the sun: Constructing an idea as a shared intellectual property. *Journal of Research in Science Teaching, 47*(6), 619-642.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, UK: Nielsen & Lydiche.

Reiser, B. J. (2014, April). *Designing coherent storylines aligned with NGSS for the K-12 classroom.* Paper presented at the National Science Education Leadership Association Meeting, Boston, MA.

Resnick, L. B., Michaels, S., & O'Connor, M. C. (2010). How (well structured) talk builds the mind. In D. Preiss & R. J. Sternberg (Eds.), *Innovations in educational psychology: Perspectives on learning, teaching, and human development* (pp. 163-194). New York, NY: Springer.

Rivet, A., & Kastens, K. A. (2012). Developing a construct-based assessment to examine analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching*, *49*(6), 713-743.

Roschelle, J., Knudsen, J., & Hegedus, S. J. (2010). From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning. In M. J.

Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 233-262). New York: Springer.

Roth, K. J., Chen, C., Lemmens, M., Garnier, H. E., Wickler, N. I. Z., Atkins, L. J., . . . Zembal-Saul, C. (2009). *Coherence and science content storylines in science teaching: Evidence of neglect? Evidence of effect?* Paper presented at the NARST Annual Conference, Garden Grove, CA.

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., . . . Gallimore, R. (2006). Teaching science in five countries: Results from the TIMSS 1999 video study. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88,* 345-372.

Schunn, C. D., Merlino, F. J., Cromley, J. G., Massey, C., Newcombe, N. S., & Nokes, T. J. (2010). *Translational science of cognitive science in middle school science curricula*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Schunn, C. D., Richey, J. E., & Alfieri, L. (2011). *Contrasting cases can facilitate hands-on middle school science learning at scale*. Paper presented at the 23rd Annual Convention of the Association for Psychological Science, Washington, DC.

Shayer, M. (1978). The analysis of science curricula for Piagetian level of demand. *Studies in Science Education, 5*(1), 115-130.

Shutt, K., Phillips, R. S., Vye, N., Van Horne, K., & Bransford, J. D. (2010, April). *Developing science inquiry skills with challenge-based, student-directed learning.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Shutt, K., Vye, N., & Bransford, J. D. (2011). *The role of agency and authenticity in argumentation during science inquiry*. Paper presented at the NARST Annual Meeting, Orlando, FL.

Singer, J. E., Krajcik, J., Marx, R. W., & Clay-Chambers, J. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist, 35*(3), 165-179.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective, 4*(1&2), 1-98.

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.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching, 39*(9), 889-910.

Thompson, J., Braaten, M., & Windschitl, M. (2009, June). *Learning progressions as vision tools for advancing teachers' pedagogical performance.* Paper presented at the Learning Progressions in Science Conference, Iowa City, IA.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Tzou, C. T., & Bell, P. (2010). *Micros and Me:* Leveraging home and community practices in formal science instruction. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Proceedings of the 9th International Conference of the Learning Sciences* (pp. 1135-1143). Chicago, IL: International Society of the Learning Sciences.

van Zee, E. H., Iwasyk, M., Kurose, A., Simpson, D., & Wild, J. (2001). Student and teacher questioning during conversations about science. *Journal of Research in Science Teaching, 38*(2), 159-190.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427-450.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, *92*(6), 941-967.

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 359-404). Rotterdam, the Netherlands: Sense Publishers.

Wright, B. D., & Linacre, M. J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wrightstone, J. W., Parke, M., & Bressler, M. (1944). Introducing child study practices in selected schools in New York City. *The Journal of Educational Research, 37*(7), 512-520.

Yackel, E., & Cobb, P. (1996). Socio-mathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education, 27*(4), 458-477.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., . . . Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335-359.

Zangori, L., Forbes, C. T., & Biggers, M. (2013). Fostering student sense making in elementary science learning environments: Elementary teachers' use of science curriculum materials to promote explanation construction. *Journal of Research in Science Teaching, 50*(8), 989-1017.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*(1), 121-130.

Zubrzycki, J. (2014). Denver Public Schools "back to the drawing board" in search for Common Core-aligned curriculum. *Chalkbeat Colorado*. Retrieved [12/24/15] from http://co.chalkbeat.org/2014/12/17/denver-public-schools-back-to-the-drawing-board-in-search-for-common-core-aligned-curriculum/

Table 1.

Causes of Earthquakes, Volcanoes and Mountain Building Facet Cluster

Goal Facets

00. Movements of plates cause earthquakes, volcanoes and mountains. Earth's surface layer (lithosphere or crust) is composed of tectonic plates that move in various directions with respect to each other. There are three types of plate boundaries: divergent boundaries (e.g., spreading ridges), convergent boundaries (e.g., subduction zones) and places where two continents slide past each other (called transform boundaries). Earthquakes can occur at any one of these plate boundaries.
    - 01. When two plates move away from each other along what are called divergent boundaries, the movement produces mid-ocean mountains and rift valleys and possibly volcanoes due to rising and cooling magma
    - 02. When two plates collide along what are called convergent boundaries, the movement produces mountains, trenches, or volcanic arcs. The type of crust (oceanic or continental) that collides determines the characteristic features of these areas (because oceanic plates are more dense than continental plates).
    - 03. When two plates slide past each other along transform boundaries, fault lines and fault scarps can occur.
    - 04. When weakened areas within plates move over hot spots, volcanoes can form.

Problematic Facets

20. Volcanoes are caused by pressure building up inside earth and being relieved by breaking through the crust.

30. Only plate collisions create landforms.
    - 31. Mountains and volcanoes only form when plates collide or rub together.
    - 32. Nothing happens if plates don't collide with each other.

40. The student thinks that volcanoes, mountains and earthquakes happen on plates, but may not recognize that they occur where there is movement at the plate boundaries.

50. The student thinks that all volcanoes and mountains form in the same way.
    - 51. Mountains form when magma is not strong enough to make a volcano.
    - 52. The student thinks that magma is only found under volcanoes.
    - 53. Volcanoes are mountains with holes.

60. The student relates the creation of volcanoes to heat and fire, both inside and outside the crust.
    - 61. Volcanoes and mid-ocean ridges are caused by the inner core.
    - 62. Fire and heat create volcanoes and ridges.

70A. The student thinks that ditches, sink holes, empty spaces form at divergent boundaries.

70B. The student thinks that earthquakes cause volcanoes and mountains to form.

80. The student thinks that landforms on Earth's crust are only due to weathering and erosion.

90. The student thinks that weather causes earthquakes to occur and volcanoes to form.

Note. Each facet has a two-digit number. The X0's indicate more general statements of student ideas. These may be followed by more specific examples, which are coded X1 through X9. Problematic facets begin with the numbers 2X through 9X. Problematic facets are roughly ranked from least to most problematic, with higher facet numbers (e.g., 7X, 8X, 9X) representing the most problematic ideas.

Table 2.

Videos Submitted by Unit and Activity Type

		UNIT	
Condition		Second Unit	Dynamic Planet
Treatment	Beginning of Investigation	9	8
	End of Investigation	8	7
	Supplementary Activity	7	5
Comparison	Beginning of Investigation	1	3
	End of Investigation	4	7
	Supplementary Activity	0	5
	TOTAL	29	35

Table 3.

Description of High-Leverage Talk Moves and Examples in Video Analysis

High-Leverage Talk Moves	*Code Definition*	*Example from Data*
Elicit Thinking and Reasoning	The teacher poses questions that (are designed to) elicit explanations that provide a (typically causal) account for phenomena. This move is a press for reasoning.	T: [Student name], Why would somebody think that water wore that rock down the way it is?
Follow-Up Question	The teacher asks questions that (are designed to) probe for the thinking that underlies a student's reasoning. The conversational exchanges in which these moves occur consist of a sequence of contiguous turns between the teacher and the same student.	Student reads out the group's explanation to account for the differential weathering of rocks in the City of Rocks Park. T (addressing class): What factors did you hear from this group that could affect the rate of weathering? S1: The amount of sunlight? **T: And how would the amount of sunlight be a factor?** S1: It's like (inaudible) with the rust, if the rock has iron in it, it rains a lot there, it gets a lot of sunlight, then it might oxidize more than if there's a rock where it rains a lot, but it doesn't get a lot of sunlight.
Revoicing	The teacher repeats or rephrases some or all of what the student has said and then asks the student to verify or whether or not the teacher's rephrasing reflects the student's thinking.	S: When the water is going this way and that, it's not the strongest of currents, and it might become like a few (inaudible), I don't think it is strong enough to really take away any of that, a large factor of the sediments T: **So you're saying that it is typically gentle?** S: Yeah
Adding On	The teacher asks, "Who can add on?" to invite participation from anyone to join in and respond to or build on someone else's idea.	T: Why are you saying that sedimentary would weather faster than the igneous? S: Because it is made of sediments and igneous, it cools from lava and everything, well, magma, not lava. **T: [Student name], What would you add?**

		Student: I think we were kind of thinking about our experiment, because in our experiment, sedimentary rocks weathered a lot more than igneous rocks.
Weighing Perspectives	The teacher asks a question that asks students to consider an idea that is part of the discussion, put forward their perspective on the topic, note their agreement or disagreement, and explain why.	The teacher invites the class to explain why the riverbeds are not always deep enough to contain all of the water that is in the river during spring season. A student explains that sediment leaves the riverbed stays at a constant height because sediment is both deposited and withdrawn by the movement of the water. **T: [Student Name 1] has presented a position where there is sediment wearing away, leaving, but also sediment arriving, thus making the rivers, the riverbeds, at a specific height. Do you agree, disagree, and why? [Student name 2]?**
Other Ideas	The teacher invites students to contribute other ideas to the discussion. It does not ask respondents to relate their comments to the current idea. This move is therefore different from the "Adding on" move.	T: So, I'm thinking about this one, all rock is changing over time. **Is there anything else you can bring to the discussion about ways that rocks change, in addition to what's already been said? Anything you remember from the reading last night?**
Summarize	The teacher synthesizes or consolidates the key ideas and understandings that the class has arrived at during the lesson or activity.	Three students have offered explanations to account for why the river bed is not deep enough T: Excellent. **So we have a third one, which is a combination of the sediment deposit-withdrawal. So, we have [Student Name 1], with sediment/ withdrawal, [Student Name 2] with a sponge soaking it up, and [Student Name 3] combination of sediment-withdrawal and water's probably not changing too much come spring season, so, so the depths of the river banks is about the same, combination.**

Table 4.

Use of Norms for Orchestrating Academically Productive Talk by Treatment (T) and Comparison (C) Teachers

Teacher	Everyone Participates	Support Claims with Evidence	Challenge Ideas but Respect the Person	Revise and Rethink Often
T1	Yes - Implicit	Yes - Explicit	Yes - Implicit	No
T2	No	Yes - Explicit	No	Yes
T3	Yes - Explicit	Yes - Explicit	Yes - Explicit	Yes - Explicit
T4	No	Yes - Implicit	No	No
T5	No	Yes - Explicit	Yes - Explicit	No
T %	40%	100%	60%	40%
C1	No	No	No	No
C2	No	No	No	No
C3	No	No	No	No
C4	No	No	No	No
C5	Yes -Explicit	No	Yes - Explicit	No
C %	20%	0%	20%	0%

Table 5.

Use of High Leverage Talk Moves by Treatment (T) and Comparison (C) Teachers

Teacher	Elicit Thinking and Reasoning	Follow-Up Question	Revoicing	Adding On	Weighing Perspectives	Other Ideas	Summarize
T1	Yes*	Yes	No	No	Yes	Yes	No
T2	Yes	No	Yes	No	No	No	No
T3	Yes	Yes	No	No	Yes	Yes	No
T4	Yes	Yes	Yes	Yes	Yes	Yes	Yes
T5	Yes	Yes	Yes	Yes	No	No	Yes
T %	100%	80%	60%	40%	60%	60%	40%
C1	Yes	No	No	No	No	No	No
C2	Yes	No	Yes	No	Yes	No	No
C3	Yes	Yes	No	No	No	No	No
C4	No	No	No	No	No	No	No
C5	No	No	No	No	No	No	No
C %	60%	20%	20%	0%	20%	0%	0%

*Yes means the teacher used this move at least once during the lesson.

Elicitation Question

Pose "Clicker" Question

What can happen at or near plate boundaries? Be prepared to explain how these processes might result from plate collision.

a. Magma can form.
b. Volcanic activity can occur.
c. Mountains can form.
d. The ocean floor can spread.
e. Earthquakes can happen.

Spark Discussion

How might [a, b, c, d, e] be the result of plates colliding? [All answers are correct and correspond in some way to Goal Facets.]

Reflect-and-Revise Question

Pose "Clicker" Question

In the middle of the Atlantic Ocean floor, and on other ocean floors, there is a long ridge of volcanic rock. What might be going on in and on Earth's surface to cause this? Be prepared to explain your answer.

a. Magma rises and then cools where the two plates diverge.
b. Earthquakes cause a ridge to rise along the bottom of the ocean.
c. Heat and fire in the crust cause the ridge to form.
d. A ridge rises as the plates slide past each other along a transform boundary.

Spark Discussion

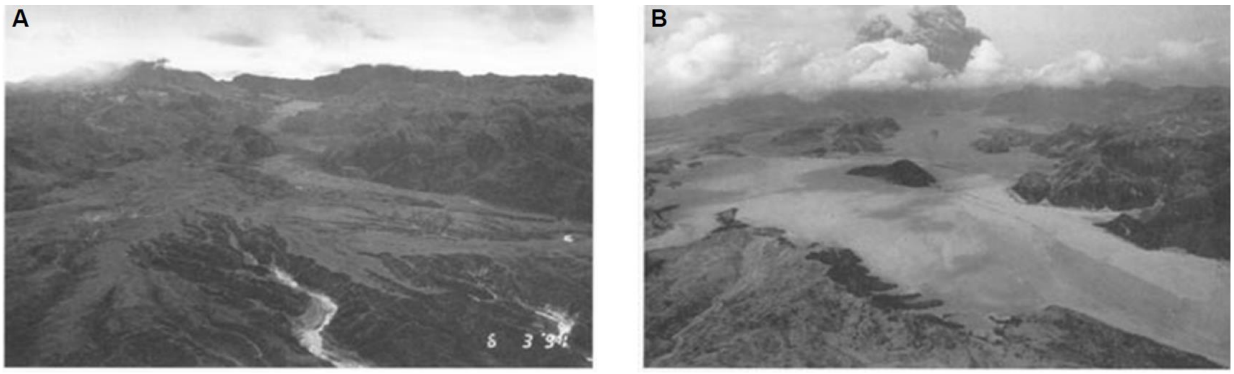
Why would [a, b, c, d] result in a long ridge of volcanic rock? [Listen for Goal and Problematic Facets in student reasoning.]

- Response a: Correct Response. Students may know that magma can rise where two plates diverge in the middle of oceans, creating a ridge of volcanic rock as the magma cools. Volcanic rock can also be found at other kinds of plate boundaries.
- Response b. Students may know that earthquakes can occur at plate boundaries, including in the ocean. However, earthquakes do not create ridges of volcanic rock.
- Response c: Students may know that magma has a very high temperature and can form in the crust. However, the heat comes from deep within Earth's interior and does not create ridges of volcanic rock.
- Response d: Students may know that when plates collide along a transform boundary, the plates slide past each other. However, the middle of the Atlantic is a divergent plate boundary.

Figure 1.

Elicitation Question and Reflect-and-Revise Question

DP2-14. These two pictures show Mount Pinatubo, a large active volcano in the Philippines. Sam thinks the pictures were taken hundreds of years apart since Earth's surface looks so different in each picture. Ronaldo thinks the pictures could have been taken just days apart. Who is right? Explain your answer.



(Source for Images: https://en.wikipedia.org/wiki/File:River_valley_filled_in_by_pyroclastic_flows,_Mt._Pinatubo.jpg)

Figure 2.

Examples of Items on Dynamic Planet (DP) Assessment

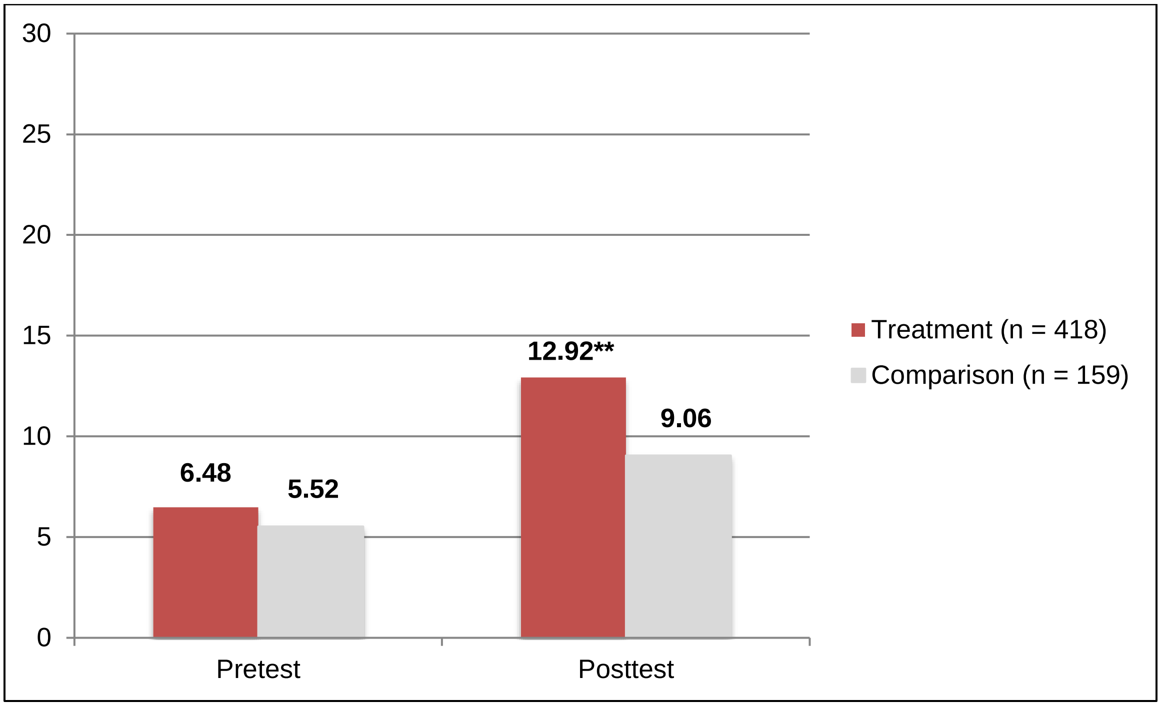


Figure 3.

Treatment and Comparison Group Student Scores