New Instruments for Studying the Impacts of Science Teacher Professional Development

March 2014

Peggy J. Trygstad
Eric R. Banilower
P. Sean Smith
Courtney L. Nelson

Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514

Author Note

This research was supported by National Science Foundation Grant No. 0928177 and Grant No. 03353328. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to: Peggy Trygstad, Horizon Research, Inc., 326 Cloister Court, Chapel Hill, NC 27514-2296. E-mail: ptrygstad@horizon-research.com.

Abstract

The logic model that implicitly drives most professional development (PD) efforts asserts that PD leads to changes in teacher knowledge and beliefs, which leads to improved classroom practice, and ultimately, better student outcomes. However, efforts to study the impacts of PD programs are often hampered by the scarcity of high-quality instruments. This paper describes the development of a set of learning-theory aligned instruments including: coupled teacher and student content assessments that measure conceptual understanding in each of four topics at two different grade levels (upper elementary and middle school); a survey of teacher beliefs about effective science instruction; and a classroom observation protocol. These instruments have been used in a number of research and evaluation projects to study professional development and its impact on teacher content knowledge, beliefs, classroom practices, and/or student achievement.

Introduction

The logic model that implicitly drives most professional development (PD) efforts asserts that PD leads to changes in teacher knowledge and beliefs, which leads to improved classroom practice, and ultimately, better student outcomes (see Figure 1).

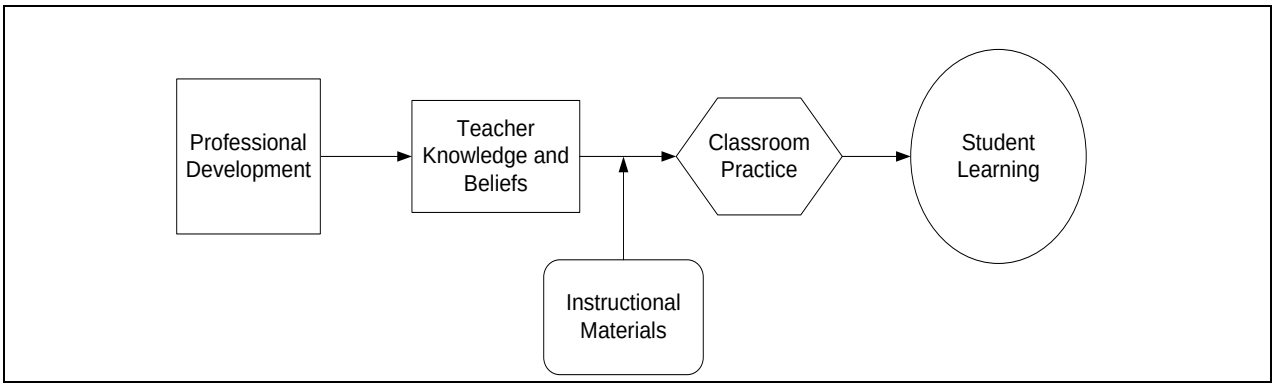

Figure 1. **Theory of Action for Professional Development**

However, efforts to study PD programs are often hampered by the scarcity of high-quality instruments, as evidenced by ongoing federal funding priorities (e.g., National Science Foundation, 2011). Even though a number of instruments have been developed in recent years to help address this need, they cover only a small portion of science content areas. Thus, many researchers face the dilemma of having to piece together instruments in an attempt to assess content that they target, or use established instruments that may only partially align with the topics and skills being addressed. For instance, many researchers rely on data from state assessments that oftentimes cover a much broader range of topics than their project addresses, raising the likelihood that the assessments would not be sensitive to project impacts.

In addition to issues with content alignment, there is a lack of instruments that reflect what has been learned in recent years about effective teaching and learning. For example, research has identified principles and practices of effective science instruction that can inform how teachers support student conceptual development (Bransford, Brown, & Cocking, 1999; National Research Council, 2011). Specifically, there is considerable evidence that instruction is most effective when it elicits students' initial ideas, provides them with opportunities to confront those ideas, helps them formulate new ideas based on evidence, and encourages them to reflect upon how their ideas have evolved (Banilower, Cohen, Pasley, & Weiss, 2010). However, few, if any, existing instruments explicitly reflect these principles.

The Assessing the Impact of the MSPs: K–8 Science (AIM) project has developed and made available a number of learning theory-aligned instruments for examining elements of the theory of action and the relationships among them. The instruments are:

- Coupled teacher and student content assessments in each of four topics at two different grade levels (upper elementary and middle school) that measure conceptual understanding;
- A survey of teacher beliefs about effective science instruction; and
- A classroom observation protocol.

This paper describes the process used for developing each of these instruments and highlights their key features.

Measuring Teacher and Student Content Knowledge

There is broad agreement that teacher knowledge of disciplinary content directly and positively affects classroom practice and, ultimately, student learning. However, empirical support is thin, largely because of a lack of appropriate measures. Studies rely primarily on proxies of teacher content knowledge, such as certification type (Goldhaber & Brewer, 2000), undergraduate major (Monk, 1994), and courses taken (Druva & Anderson, 1983). However, few studies use direct measures of teacher content knowledge. Furthermore, existing student measures either tend to have weak psychometric properties, or they are very broad (e.g., state-administered assessments), further limiting the likelihood that relationships between teacher knowledge of particular content and student learning will be detected.

AIM has developed tightly coupled assessments of teacher and student content knowledge at both the elementary and middle school levels. These assessments span four content areas:

- Evolution and diversity;
- Force and motion;
- Populations and ecosystems; and
- Properties and states of matter.

The assessments in each content area are closely aligned to the same carefully defined content domain and are tailored for a specific audience with regard to complexity and question contexts.

Procedure

Development of the AIM assessments closely mirrors a development process that has produced teacher and student science assessments with strong evidence of validity and reliability (Smith, 2010). This assessment development process is described below.

Defining the Content Domain

Four topic areas were selected from the *Science Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board, 2008). The *NAEP Framework* was based primarily on the *National Science Education Standards* (National Research Council, 1996) and the *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993), but also reflected developments in science and policy that had taken place since those documents were published. Thus, at the time of instrument development (2009–11), we considered the *NAEP Framework* to be a consensus statement of the most important science

concepts students should understand as a result of K–12 science education. In addition, the concepts were expected to be, and are, reflected in the Next Generation Science Standards (Achieve, Inc., 2013) at the 3–5 and 6–8 grade ranges.

The content in each area for each grade range was unpacked by experts (Ph.D. scientists and science educators) into discrete, assessable statements that represented the science concepts students should learn, as well as the concepts teachers need to know in order to teach the content well. These domain specification documents formed the foundation of the assessments. An example of this unpacking is shown in Figure 2.

4a: Materials have properties

1. The properties of anything are the characteristics used to describe that thing, such as color, weight (mass), size, and so on.
2. Various objects and substances (materials) can be identified and distinguished by their properties.

4b: Samples of materials can be classified by their behavior into solids, liquids, and gases.

1. Solids have a definite shape and volume that cannot be easily changed.
2. Liquids flow to take the shape of their container and have a definite volume that cannot be easily changed.
3. Gases expand to fill any shape and volume container and can easily be compressed into a smaller volume container.

Figure 2. **Sample Elementary Student Ideas for Properties and States of Matter**

Writing Multiple-Choice Items

To enable large-scale research, we set out to create assessments that would be minimally burdensome, both for the test-taker and the researcher. Accordingly, we opted for a multiple-choice format, recognizing both the strengths and limitations of such items. For instance, well-constructed, open-ended items may probe more depth of understanding than multiple-choice items, but they are more burdensome for both the researcher (in terms of scoring costs and training to establish inter-rater reliability) and the test-taker (in terms of time required to complete the assessment).

Research has shown that multiple-choice assessments can be reliable and valid tools for assessing the prevalence of students' alternative conceptions in science (Sadler, 1998). Accordingly, we conducted a thorough search that yielded a list of commonly held alternative conceptions in each content area. These alternative conceptions informed the item-writing process, being incorporated into both question stems and answer choices.

Three types of items for assessing teacher content knowledge, each set in instructional contexts, were written: knowledge of science content (Level 1); assessing teacher content knowledge through the analysis of student thinking (Level 2); and assessing teacher content knowledge through instructional decision-making (Level 3). The instructional contexts make it obvious to test takers that the assessment was written for teachers rather than students. Sample teacher assessment items are shown in Figures 3 and 4. (In all sample items, the correct answers appear in bold text.)

A teacher asks his students if plants have any competitors in an ecosystem. One student responds:

"Plants do not need to compete with each other, because they make their own food. They're not like animals who have to fight over food."

Based on this statement, which of the following ideas does the student seem to be missing?

A. Plants are producers.
B. Food is not the only resource for which organisms compete.
C. Animals compete with other animals for resources.
D. None. The student has an accurate understanding of competition.

Figure 3. **Example Level 2 Item**

A teacher gives her students the following scenario: *"Three books are sitting on a table. Each has a different mass. If I push each book just as hard for the same amount of time, which book's motion will change the most?"*

Most students agree that all of the books will have the same change in motion because the same force is applied to all of the books. Which of the following would be the best next step to move these students forward in their understanding about the effect of forces on motion?

A. Drop all three books from the same height at the same time and see which book hits the ground first.
B. Push the books across different surfaces that have varying amounts of friction.
C. Show a video that illustrates how the strength of an applied force and the mass of an object affect an object's motion.
D. Have a class discussion about the difference between mass and weight.

Figure 4. **Example Level 3 Item**

Note that in Figure 2, answer choices A and C include scientifically correct statements. However, only choice B addresses the misconception that the student comment suggests; that plants do not change light energy into other forms of energy. Similarly, in Figure 3, each of the answer choices presents an instructional activity that is reasonable to include in a unit on force and motion. However, only one choice pertains to the student comment in the question.

The student assessment items are much more straightforward in that they do not include instructional contexts. A sample elementary grades student assessment item is shown in Figure 5. Note that each of the incorrect answer choices includes a common alternate conception.

The deepest parts of the ocean are dark and very cold. Why are some organisms able to survive even in this environment?

A. Some organisms are strong and fit, so they are able to survive in any environment.
B. Some organisms are able to survive in dark, cold ocean water because they were born in that environment.
C. Different organisms have characteristics that help them survive in different environments.
D. Different organisms can decide to charge their bodies to help them survive in different environments.

Figure 5. **Example Student Assessment Item**

Cognitive Interviews

We next initiated multiple rounds of cognitive interviews (Hamilton, Nussbaum, & Snow, 1997) with the target audience (teachers or students). The interviews revealed whether the teachers/students interpreted the questions as intended and whether they used their knowledge of the targeted content to answer the question. The data collected via cognitive interviews were used in a series of team meetings to collaboratively edit the items.

Piloting the Assessments

The teacher assessments were each completed by 350–450 teachers. Because the assessments are intended to be used to measure change in teacher content knowledge from before professional development (when one expects content knowledge to be relatively low) to after professional development (when content knowledge should be higher), teachers with a broad range of content knowledge were recruited for the pilots to help ensure that the final assessment would be sensitive to change. For the student assessments, AIM recruited teachers to administer the student items to their classes. Between 500 and 600 students, again with a range of knowledge in the targeted area, completed each of the student assessments.

Findings

We conducted both classical and item response theory (IRT) analyses on the pilot data and ultimately used those results to select 20–30 items for each assessment. The IRT reliabilities for the final assessments can be seen in Tables 1 and 2.

Table 1
IRT Reliabilities for Teacher Assessments

	Number of Items	IRT Reliability
Elementary Evolution and Diversity	30	0.88
Elementary Force and Motion	30	0.86
Elementary Populations and Ecosystems	27	0.83
Elementary Properties and States of Matter	30	0.90
Middle School Evolution and Diversity	30	0.85
Middle School Force and Motion	30	0.95
Middle School Populations and Ecosystems	26	0.78
Middle School Properties and States of Matter	30	0.84

Table 2
IRT Reliabilities for Student Assessments

	Number of Items	IRT Reliability
Elementary Diversity of Life[1]	22	0.82
Elementary Force and Motion	25	0.81
Elementary Populations and Ecosystems	25	0.83
Elementary Properties and States of Matter	25	0.77
Middle School Evolution and Diversity	30	0.84
Middle School Force and Motion	30	0.66
Middle School Populations and Ecosystems	26	0.82
Middle School Properties and States of Matter	30	0.79

Measuring Teacher Beliefs about Science Instruction

The importance of teacher attitudes and beliefs about science instruction is evident in the number of attempts to capture different dimensions of the construct. Several well-documented measures exist to measure teacher self-efficacy (e.g., Southerland, Sowell, Kahveci, Granger, & Gaede, 2006; Riggs & Enochs, 1990), teacher attitudes toward science (e.g., Cobern & Loving, 2002; Fraser, 1978), beliefs about science teaching environment (Lumpe, Haney, & Czerniak, 2000), beliefs about the nature of science (e.g., Lederman, Abd-El-Khalick, Bell, Schwartz, & Akerson, 2002; Schwartz, Lederman, & Lederman, 2008), and beliefs about science teaching and learning (e.g., Sampson & Benton, 2006; Luft & Roehrig, 2007). Of course, teacher beliefs are of interest not just in themselves but, more importantly, in relation to science instruction. For example, teacher beliefs and attitudes regarding science as a discipline have been shown to affect lessons on the nature of science (Brickhouse, 1990). Epistemological beliefs influence teacher choices about instructional strategies and the implementation of curricula (Cronin-Jones, 1991). None of these instruments, however, are explicitly aligned with learning theory. To fill this gap, AIM developed a new survey to measure teachers' beliefs about effective science instruction.[2]

Procedure

The process of developing the Teacher Beliefs about Effective Science Teaching (TBEST) Questionnaire closely followed the previously described assessment development sequence.

Defining the Construct

Banilower and colleagues (2010) proposed five "elements" of effective science instruction, based on cognitive science:

1. Motivating the learner;
2. Eliciting the learner's initial ideas about the targeted content;

[1] This assessment addresses ideas that are precursors to evolution concepts but not evolution itself. Thus, "Diversity of Life" was chosen as the title to more accurately represent the content of the assessment. This assessment was recently revised and is currently being piloted.

[2] For additional information about this instrument see Smith, P. S., Smith, A. A., & Banilower, E. R. (in press).

3. Intellectually engaging the learner with phenomena related to the targeted content;
4. Using evidence to make and critique claims about the targeted content; and
5. Making sense of ideas about the targeted content.

These five elements defined the boundaries of the "content domain" for the survey. A group of science education researchers then deconstructed each element into more fine-grained statements. Examples of these statements are shown in Figure 6.

1. Purpose/Motivation

1.1. Learning is enhanced when students can recognize a purpose of what they are doing in a lesson.

1.2. Learning is enhanced when lessons address ideas that students wonder about or are induced to wonder about.

1.3. Learning is enhanced when the teacher/materials points out how what students will learn connects to real-world applications.

1.4. Learning is enhanced when the teacher/materials points out how what students will learn connects to their own lives outside the classroom.

2. Eliciting Students' Prior Knowledge

2.1. Learning is enhanced when students have an opportunity to consider, express, and share their initial ideas about a science concept prior to a sequence of lessons on a concept.

2.2. Teachers need to be aware of their students' initial ideas about a science concept at the beginning of a sequence of lessons on a concept.

Figure 6. **Sample Statements Representing the TBEST Content Domain**

Writing Questionnaire Items

The fully specified content domain was used by researchers to generate questionnaire items. Collaborative item editing meetings provoked spirited discussions among the research team around two themes: practicality and appropriateness. Researchers frequently expressed the concern that if teachers' instruction aligned closely with all elements of effective instruction, teachers would be unable to "cover the curriculum" (a contradiction inherent in national standards documents at the time of development). Additionally, some phenomena do not lend themselves to first-hand investigation because they are inaccessible (for example, convection in Earth's mantle). Because writing items that reflected practical and content-specific constraints proved fruitless, the questionnaire asks respondents to set these constraints aside, focusing on their views of effective science instruction in general. Sample TBEST items are shown in Figure 7.

Practical constraints aside, do you agree that doing what is described in each statement would help most students learn science?

	Strongly Disagree	Moderately Disagree	Slightly Disagree	Slightly Agree	Moderately Agree	Strongly Agree
a. Teachers should provide students with opportunities to connect the science they learn in the classroom to what they experience outside of the classroom.	1	2	3	4	5	6
b. At the beginning of instruction on a science concept, students should be provided with definitions for new scientific vocabulary that will be used.	1	2	3	4	5	6
c. Hands-on activities and/or laboratory activities should be used primarily to reinforce a science concept that the students have already learned.	1	2	3	4	5	6

Figure 7. **Sample TBEST Items**

Cognitive Interviews

We conducted cognitive interviews (Desimone & Le Floch, 2004) with middle grades science teachers nationally to ensure that the questionnaire items were being interpreted as intended. The data collected via cognitive interviews were used to make edits to the items.

Piloting the Instrument

Researchers composed over 100 items intended to conceptually align with the five elements of effective science instruction. Approximately, 950 middle grades science teachers responded to the first pilot of the items, which was conducted online. A number of important and related findings emerged from the data. First, the four-point agreement response-option formats did not generate sufficient variation in teacher responses. (Several had no variation in responses and were eliminated from the survey.) Second, the data suggested that some respondents did not answer the questions thoughtfully. For instance, some individuals gave the same response to adjacent items that had opposite meanings. Our hypothesis was that the lack of thoughtfulness was due to the length of the questionnaire.

Based on the results, we chose the importance response-option scale and 23 items for the second phase of piloting, also conducted online. The items were chosen based on coverage of the content domain and variation in responses. Middle grades science teachers were recruited for participation, and an exploratory factor analyses (EFA) was conducted on the resulting sample of just under 250 respondents. The EFA was run using an oblique rotation, which allowed any underlying factors to correlate. The analysis suggested five factors, which, based on the items, were labeled: (1) the importance of situating learning; (2) the importance of using evidence in sense making; (3) the importance of connecting new learning and prior learning; (4) the importance of using activities to confirm concepts that have already been taught (which we refer to as confirmatory instruction); and (5) the importance of hands-on instruction. However, some of the factors were highly correlated (e.g., the correlation between situating learning and

confirmatory instruction was –0.57), causing concern about whether the factors were indeed distinct dimensions.

In order to assess the robustness of the five-factor structure, a third pilot was conducted. At this point, we addressed a disconcerting feature of the survey. Although the importance response-option format produced sufficient variation in responses, it seemed a force fit for many of the statements, requiring respondents to mentally alter the item or the response options to create alignment. Rather than continue with this response-option format, we returned to the agreement format but expanded it to six points, rewording the items to make them appropriate for the response options. The result was much better alignment between the items and the response options.

Approximately, 250 middle grades science teachers responded to the new version of the questionnaire. Using the five-factor solution suggested by the EFA, a confirmatory factor analysis (CFA) using Mplus version 5.2 was applied. However, the CFA results did not support the five-factor solution, and follow-up analyses suggested a three-factor solution was more appropriate. The three factors were conceptually coherent and were labeled: (1) Learning-theory-aligned science instruction; (2) Confirmatory science instruction; and (3) All hands-on all the time.

Next, we investigated the psychometric soundness of the survey's underlying structure across administration modes (paper versus online) and grade levels (K–12). In the first of these studies, just over 600 teachers were randomly assigned to receive either an online or paper version of the instrument. The previous pilots had been exclusively online; however, we anticipated that other researchers might prefer a paper-and-pencil version. Therefore, it seemed important to establish that similar results would be obtained regardless of administration mode. We decided to conduct an EFA on data from the paper version followed by a CFA on data from the web version. The same three-factor solution fit for both modes of administration, and there were no statistically significant differences in factor composite means, suggesting that the instrument produces similar scores regardless of whether it is administered on paper or online.

We were also interested in the robustness across grade levels, anticipating that researchers might want to use the TBEST in studies of elementary, middle, or high school science teaching. A final study was designed in which we administered the TBEST to a total of 900 elementary, middle, and high school teachers. To test whether the factor structure was the same across grade levels, a multiple-group CFA procedure was followed, again using Mplus version 5.2. This procedure involves conducting an initial CFA for each grade range separately, followed by a multiple-group CFA.

Findings

The analyses provide support for the same three-factor model for each grade range. The factors were not highly correlated with each other, suggesting distinct constructs. (See Table 3.) Furthermore, the reliabilities (Cronbach's alpha) of the composites for each grade range are above 0.70. (See Table 4.) These findings were consistent across all grade ranges.

Table 3
Correlations Among Factors[†]

	Learning-Theory-Aligned Science Instruction	Confirmatory Science Instruction	All Hands-on All the Time
Learning-Theory-Aligned Science Instruction	1.00		
Confirmatory Science Instruction	-0.18	1.00	
All Hands-on All the Time	-0.07	0.45	1.00

[†] Factor correlations were similar across grade ranges

Table 4
Cronbach's Alpha Reliability Coefficients by Grade Range Taught

	Grade Range			
	Overall	Elementary	Middle	High
Learning-Theory-Aligned Science Instruction (11 items)	0.713	0.766	0.739	0.761
Confirmatory Science Instruction (7 items)	0.771	0.758	0.775	0.784
All Hands-on All the Time (3 items)	0.758	0.794	0.747	0.732

To summarize, the resulting questionnaire contains 21 items using a six-point agreement response scale. The items fall into three factors: (1) Learning-theory-aligned science instruction; (2) Confirmatory science instruction; and (3) All hands-on all the time. Statistical findings support the psychometric structure of the survey across different modes of administration and across teachers of various grade ranges.

Gauging Student Opportunity to Learn Science Ideas

Assessments and questionnaires capture what teachers know and believe about teaching, but perhaps the best way to understand their classroom practice is through observation. There are many observation protocols available to the field; some are content neutral, such as the Framework for Teaching (Danielson, 2011) and the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008), and others are specific to science, such as the Reformed Teaching Observation Protocol (Sawada et al., 2002), the Science Teacher Inquiry Rubric (Beerer & Bodzin, 2003), and the Local Systemic Change Classroom Observation Protocol (Horizon Research, Inc., 2005). The protocols vary in their intent; however, none were explicitly designed with the aforementioned elements of effective science instruction in mind. The AIM Classroom Observation Protocol (COP) was developed to address this need. The protocol provides a structure for examining classroom practices in order to gauge student opportunity to learn targeted science ideas.

Procedure

The AIM COP is not intended to advocate a particular set of instructional strategies, but rather focuses on students' opportunities for conceptual change regardless of pedagogy. Observers are asked to rate five components of instruction, related to the elements of effective science instruction:

1. Appropriateness of science content;
2. Opportunities to surface prior knowledge;
3. Engaging with examples/phenomena;
4. Using evidence to draw conclusions/make claims about the examples/phenomena; and
5. Sense-making of the targeted ideas.

In each of these sections, observers first rate the extent to which several key features were present in instruction. Additionally, observers are asked to consider in their rating what proportion of students were engaged in the instruction related to each feature. The key features considered for opportunities to surface prior knowledge, as well as the rating scale, can be found in Figure 8.

	Not at all			**To a great extent**
Deliberate opportunities provided to surface students' prior knowledge:				
a. were structured/implemented so that students would be aware of their own prior knowledge.	1	2	3	4
b. surfaced students' reasons for how they were thinking.	1	2	3	4
c. had students record aspects of their prior knowledge.	1	2	3	4
d. had students make public aspects of their prior knowledge.	1	2	3	4
e. allowed students' ideas to be surfaced without judgment.	1	2	3	4

***Figure 8*. Key Features of Opportunities to Surface Prior Knowledge**

After rating the presence of the key features, observers rate the extent to which these features of instruction aligned with the targeted science idea. Finally, observers combine this information to make a holistic rating of the extent to which the opportunities for students in that domain were likely to be sufficient for their learning of the targeted idea. Observers are asked to support each of their ratings with evidence from their observations.

Two scenarios that show different ways students' prior knowledge may be surfaced can be found in Figures 9 and 10. In both scenarios, the instruction is meant to address the same targeted idea—that in a contact push/pull interaction, the force ceases to exist as soon as contact between the interacting objects is lost.

A teacher asks her students to answer the following question:

Imagine a soccer player taking a shot on goal. She runs up and kicks the ball which flies toward the goal, where the goalkeeper catches it.

Which of the choices below is closest to when you think the force of the kick stopped acting on the ball?

a) Before the ball lost contact with the foot
b) At the moment the ball lost contact with the foot.
c) After the ball lost contact with the foot, but before it got to the goalkeeper.
d) When the goalkeeper stopped the ball moving.

The teacher instructs the students to record their answers in their notebooks and then move to one of four different locations in the classroom depending on which answer they chose. The students in each of the four groups discuss their reasoning for their response, and one student is selected to share each group's ideas with the whole class.

Figure 9. **Scenario A for Surfacing Prior Knowledge**

A teacher asks her students the following question:

What are some examples of forces that you saw on your way to school this morning?

The teacher instructs the students to record their answers in their notebooks and share their ideas with a partner. She then calls on several students to share their ideas with the whole class.

Figure 10. **Scenario B for Surfacing Prior Knowledge**

The instruction in both scenarios provides opportunities for most students to discuss their prior knowledge, record aspects of their prior knowledge, and make public aspects of their prior knowledge (i.e., share their ideas with others). However, only the instruction in Scenario A is closely aligned to the targeted idea and surfaces students' reasons for how they are thinking. Therefore Scenario A would receive a rating of a 4 on the protocol while Scenario B would be rated lower.

Piloting the Instrument

AIM piloted the use of this classroom observation protocol during the 2011–12 school year. In this pilot, researchers observed the science instruction of 28 teachers during their units on force and motion. Researchers took field notes while in the classroom and wrote up lesson summaries afterwards. After instruction on a targeted idea was complete, a researcher used all lesson summaries related to a targeted idea to complete an observation protocol for that idea. Early in the study, researchers collaborated on completing observation protocols and were given feedback by the lead researchers. Later, researchers worked independently so that inter-rater reliability (IRR) could be assessed.

Findings

Using the five ratings (one for each section of the protocol), IRR was examined using percent agreement and the intraclass correlation coefficient (ICC). Overall, researchers agreed exactly on 77 percent of their ratings, and the ICC was 0.86. The measures are both above the minimum standard described in the literature (Graham, Milanowski, & Miller, 2012), indicating sufficient IRR among researchers.

Conclusion

The instruments described in this paper were created to provide the field with tools to study the complex relationships among professional development, teacher knowledge and beliefs, and student learning. The AIM project has been using these instruments for a number of studies, such as examining professional development to identify key features associated with teacher learning. AIM has also used these instruments to examine the impacts of a professional development model explicitly tied to learning theory on teacher knowledge, beliefs, classroom practices, and student learning.

In addition, the instruments have been used in a number of evaluation and research projects conducted by other researchers. For example, the 16 content assessments have been used by a number of NSF- and state-funded Math Science Partnership projects to look at the impacts of their PD on teacher and student learning. Typically, these projects use a pre/post, or pre/post/delayed-post design to look at changes in assessment scores over time. The TBEST has been used to study impacts on teachers' beliefs resulting from an eight-day summer workshop focused on kit-based instruction. Participant scores on the learning-theory-aligned beliefs composite increased significantly and their scores on the confirmatory instruction composite decreased significantly. These findings suggest that the experience changed teachers' beliefs in positive ways (i.e., more consistent with what is known from cognitive science).

All of these instruments are being made available to the field at no cost, as tools for conducting research about professional development and its impact on teacher content knowledge, beliefs, classroom practices, and/or student achievement.[3] Further, we anticipate these instruments to have utility beyond studies of in-service PD programs. For example, the beliefs questionnaire has been used by researchers at one university to study how pre-service teacher beliefs about science instruction change over the course of their preparation program. In addition, these researchers used the classroom observation protocol as a basis for practicum observations and post-observation conferences with pre-service teachers.

[3] Instructions for accessing these instruments can found at: http://www.horizon-research.com/aim/instruments/

References

Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS. (2013). *Next generation science standards*. Retrieved August 8, 2013 from http://www.nextgenscience.org/next-generation-science-standards.

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

Banilower, E., Cohen, K., Pasley, J., & Weiss, I. (2010). *Effective science instruction: What does research tell us?* (2nd ed.). Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Beerer, K. M. & Bodzin, A. M. (2003). Promoting inquiry-based science instruction: The validation of the science teacher inquiry rubric (STIR). *The Journal of Elementary Science Education, 15*(2), 39–49.

Bransford, J., Brown, A. L., & Cocking, R. R. (1999). *How people learn: brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brickhouse, N. W. (1990). Teachers' beliefs about the nature of science and their relationship to classroom practice. *Journal of Teacher Education, 41*(3), 53–62.

Cobern, W. W. & Loving, C. C. (2002). Investigation of preservice elementary teachers' thinking about science. *Journal of Research in Science Teaching, 39*(10), 1016–1031.

Cronin-Jones, L. L. (1991). Science teacher beliefs and their influence on curriculum implementation: Two case studies. *Journal of Research in Science Teaching, 28*(3), 235–50.

Danielson, C. (2011). *The framework for teaching evaluation instrument, 2011 Edition*. Retrieved March 26, 2014 from: http://www.danielsongroup.org/article.aspx?page=FfTEvaluationInstrument

Desimone, L. M. & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1–22.

Druva, C. A. & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching, 20*(5), 467–79.

Forbes, C.T., Biggers, M., & Zangori, L. (2013). Investigating Essential Characteristics of Scientific Practices in Elementary Science Learning Environments: The Practices of Science Observation Protocol (P-SOP). *School Science and Mathematics,* 113(4), 180-190.

Fraser, B. J. (1978). Development of a test of science-related attitudes. *Science Education*, 62(4), 509–515.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Washington, DC: Center for Educator Compensation Reform.

Goldhaber, D. D. & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129–45.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181–200.

Horizon Research, Inc. (2005). 2005–06 Local Systemic Change Classroom Observation Protocol. Retrieved March 26, 2014 from: http://www.horizon-research.com/LSC/manual/0506/tab6/cop0506.pdf

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, *39*(6), 497-521.

Luft, J. A., Roehrig, G. H. (2007). Capturing science teachers' epistemological beliefs: the development of the teacher beliefs interview. *Electronic Journal of Science Education, 11(2)*, 38–62. Retrieved August, 13, 2009 from http://ejse.southwestern.edu/volumes/v11n2/v11n2_list.html

Lumpe, A. T., Haney, J. J., & Czerniak, C. M. (2000). Assessing teachers' beliefs about their science teaching context. *Journal of Research in Science Teaching*, *37*(3), 275–92.

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, *13*(2), 125–45.

National Assessment Governing Board. (2008). Science Framework for the 2009 National Assessment of Educational Progress. Washington, DC: U.S. Department of Education.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Science Foundation, (2011). Discovery Research K-12 (Solicitation 11-5888). Retrieved August 9, 2013, from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=500047.

Pianta, R., La Paro, K., & Hamre, B. K. (2008). *Classroom assessment scoring system*. Baltimore: Paul H. Brookes.

Riggs, I. M. & Enochs, L. G. (1990). *Toward the development of an elementary teacher's science teaching efficacy belief instrument*. Retrieved August 14, 2009, from: http://www.eric.ed ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED308068.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265–296.

Sampson, V. & Benton, A. (2006). Development and validation of the beliefs about reformed science teaching and learning (BARSTL) questionnaire. *Annual Conference of the Association of Science Teacher Education (ASTE). Portland, Oregon. Retrieved February* (Vol. 16, p. 2008).

Sawada, D., Pilburn, M., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*(6), 245–253.

Schwartz, R. S., Lederman, N. G., & Lederman, J. S. (March, 2008). *An instrument to assess views of scientific inquiry: the VOSI questionnaire.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Baltimore, MD

Smith, P. S. (2010, March). *New tools for investigating the relationship between teacher content knowledge and student learning.* Paper presented at the National Association for Research in Science Teaching, 2010 Annual International Conference, Philadelphia, PA.

Smith, P. S., Smith, A. A., & Banilower, E. R. (in press). Situating beliefs in the theory of planned behavior: The development of the teacher beliefs about effective science instruction questionnaire. In C. M. Czerniak, R. Evans, J. Luft, & C. Pea (Eds.), *The role of science teachers' beliefs in international classrooms: From teacher actions to student learning.*

Southerland, S., Sowell, S., Kahveci, M., Granger, D. E., & Gaede, O. (April, 2006). *Working to measure the impact of professional development activities: developing an instrument to quantify pedagogical discontentment.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.