Measuring Model-Based High School Science Instruction: Development and Application of a Student Survey

Gavin W. Fulmer *

National Institute of Education (Singapore)

1 Nanyang Walk

Singapore 637616

gavin.fulmer@nie.edu.sg

Ling L. Liang

La Salle University

1900 W. Olney Ave.

Philadelphia, PA 19141

Abstract

This study tested a student survey to detect differences in instruction between teachers in a modeling-based science program and comparison group teachers. The Instructional Activities Survey (IAS) measured teachers' frequency of modeling, inquiry, and lecture instruction. Factor analysis and Rasch modeling identified three subscales, Modeling and Reflecting, Communicating and Relating, and Investigative Inquiry. As predicted, treatment group teachers engaged in modeling and inquiry instruction more than comparison teachers, with effect sizes between 0.55 and 1.25. This study demonstrates the utility of student report data in measuring teachers' classroom practices and in evaluating outcomes of a professional development program.

Keywords: measures of instruction; modeling instruction; Rasch modeling; student survey

Measuring Model-Based High School Science Instruction: Development and Application of a Student Survey

1.1 Introduction

Students who experience inquiry-oriented science instruction obtain deeper conceptual understandings of science content and of scientific reasoning (Minner, Levy, & Century, 2010). Model-based science instruction—a type of inquiry science teaching—has demonstrated greater gains in their content knowledge than students in traditional lecture-lab courses at both secondary school and college/university levels (Authors, in press; Brewe, et al., 2010; Clement, 1989, 2010; Hestenes, Wells, & Swackhamer, 1992; Schwarz & White, 2005; Vesenka, Beach, Munoz, Judd, & Key, 2002). This has provided support for implementing a model-based instructional approach, as it indicates that the adoption of model-based curriculum improves student outcomes. It has also led to professional development programs for teachers to use model-based science instruction (e.g., Wells, Hestenes, & Swackhamer, 1995). Yet there remains a need to understand how teachers implement model-based instruction after engaging in such professional development that accounts for students' experience of the instruction. The present study addresses this need, by describing the development and initial application of a student instrument to measure teachers' model-based inquiry instruction.

A model of the relationship between a professional development (PD) program and student learning outcomes requires a mechanism by which the PD affects teachers' instruction and, in turn, how these instructional practices relate to students' learning experiences (Author, 2008; Desimone, 2009). For model-based instruction, the relationships among the PD program, subsequent changes in teachers' instruction, and student learning outcomes are still not clearly articulated. To gauge the effect of participation in a PD program, researchers must understand whether and how teachers implemented the model-based approach and then determine if this differs from the comparison group teachers. This evidence is essential to determine if differences in student outcomes are attributable to the model-based approach or to other differences between teachers' instructional practices.

Furthermore, there is a need for innovative ways to account for students' classroom experiences as one of the measures of instruction. Many studies of PD and of teachers' instruction have used either observational methods or teachers' reports of their instruction (e.g., Adamson *et al*., 2003; Lawrenz, Wood, Kirchhoff, Kim, & Eisenkraft, 2009). Other studies have combined observations with interviews and open-ended surveys of teachers and students (Author, 2008; Venville, Sheffield, Rennie, & Wallace, 2008; Waight & Abd-el-Khalick, 2007). While these methods have many strengths, there is relatively little use of students' ratings (for reviews of exceptions see Aleamoni, 1999; Marsh, 1984), particularly in science education research. This dearth of research contrasts with the importance of students' experiences of the classroom and the potential impact this will have on their learning outcomes.

This study addressed these two issues through the use of a student instrument aligned with the model-based instructional approach. The instrument, the Instructional Activities Survey (IAS), provided direct evidence of students' experiences in the classroom, and was used both with teachers implementing the model-based curriculum and teachers in a comparison group. This study applied this student instrument to understand possible differences between model-based and comparison group teachers.

The following sections present a review of relevant literature, followed by the study's methodology and results, and ending with discussion of the findings and implications for instruction and for future research.

1.2 Conceptual Framework

This study is based in the literature on model-based science instruction (Clement, 1989, 2010; Hestenes, Wells, & Swackhamer, 1992; Schwarz & White, 2005; Wells, Hestenes, & Swackhamer, 1995). Model-based science instruction is a particular case of inquiry-oriented pedagogy. It focuses on the development of students' coherent scientific understandings and their ability to construct and apply scientific models, which is aligned with "teaching science as practice" (NRC, 1996, 2000, 2007). In this study, scientific models are broadly defined as sets of representations, rules, and reasoning structures that allow one to make explanations and predictions (Wells, Hestenes, & Swackhamer, 1995; Schwarz & White, 2005).

The present study also extends the research on appropriate measurement of instruction and on professional development, addressed in greater detail in the following section. An important distinction that informed this research was between the measurement of teaching and the measurement of learning. As Fenstermacher and Richardson (2005) describe, there is a difference between *good teaching*, which is practice-oriented, and *effective teaching*, which is outcomes-oriented. Good teaching is the set of actions that teachers perform which are perceived to be of high quality by trained and expert observers. To describe good teaching requires measurement of the teacher's actions. On the other hand, effective teaching is detected by observing changes in students' knowledge or skills before and after teaching. To describe effective teaching requires only measurement of students' knowledge or ability before and after instruction. Naturally, there is a relationship between good teaching and effective teaching: good teachers are more likely to be effective (Beeth & Hewson, 1999; Fenstermacher & Richardson, 2005). In particular, if the assessment instrument used to measure students' knowledge or skills (i.e., effective teaching outcomes) aligns well with the curriculum or standards that teachers are expected to implement well (i.e., good teaching practices), then such measures of student outcomes are acceptable proxies for good teaching. However, it is often the case that standardized assessments align poorly with the curriculum or with standards documents (Authors, 2009; Author, 2011; Desimone, 2009), so measures of teachers' classroom actions should be collected alongside measures of students' learning. Furthermore, to improve PD programs for teachers, there is a need to identify what instructional activities make up "effective teaching" in a subject and the more productive ways to support teachers' expertise with those methods through PD. For the above reasons, the present study focuses on the measurement of instruction using the IAS as a measure of effective teaching. Further analyses of the relationship between the IAS and student outcomes is the topic of a related study (Authors; in preparation).

1.3 Assessment of Instruction

To understand the possible impact of teachers' classroom instruction on students or to gauge fidelity of implementation of an instructional approach, teachers' actions must first be measured. Though there is not yet a clear consensus on the best method for measuring teachers' classroom actions, two general approaches are common: self-report data from teachers or observation reports from trained raters.

1.3.1 Classroom observations

Classroom observation protocols—such as the Reformed Teaching Observation Protocol (RTOP; Sawada *et al*., 2002), the Looking for Technology Integration (LoFTI) protocol (SERVE, 2006), and the Classroom Observation Protocol (COP; Banilower, 2005)—require a trained observer to attend the class or watch class videos. During the observation, the rater then either repeatedly records teachers' and students' actions at set time intervals or records overall patterns of teachers' and students' actions across the time interval. Classroom observation methods have been found to be effective in understanding teachers' application of reformed instruction and to relate to students' outcomes (Judson & Lawson, 2007; Park, Jang, Chen, & Jung, 2011; Sawada *et al*., 2002). However, there is much variation in the amount of observation that researchers are able to conduct. In many studies using in-person observation the extent of observation varies considerably, with observations lasting anywhere from single 20-minute sessions to multiple sessions totaling hours of observation per teacher across weeks and months (Koziol & Burns, 1986; Park, Jang, Chen, & Jung, 2011).

Despite its strengths, observation methods have limitations. Collecting too few observations per teacher makes it difficult to attribute correctly whether the observation is a reliable measure of the teachers' typical instruction, whereas collecting and rating repeated observations is costly and time-consuming. Furthermore, there is a potentially large expense to train personnel and to obtain informed consent from all students or their guardians. If using video, this can create large quantities of data that, though rich in potential for analysis, also creates a major difficulty for coding and interpretation.

1.3.2 Teacher self-reports

Regarding teachers' self-reports, concerns over discrepancies between instructors' self-reports on their instruction have been expressed for quite some time, with data demonstrating that teachers give themselves higher marks on general surveys than should be expected (Centra, 1973). However, action-specific teacher self-report instruments are more reliable than general surveys (Porter, 2002) and show high correspondence with trained observers (Koziol & Burns, 1986). In particular, Porter (2002) argued that the Survey of Enacted Curriculum (SEC) instruments did not offer fewer responses on general items because that would allow respondents to distinguish desirable responses. However, because of the large number of items, the SEC instruments typically required between 45 and 90 minutes to complete. Therefore, teacher self-report surveys can be limited either by self-confirmation bias or by the burden on response by the participants.

1.3.3 Student surveys

An alternative to teacher self-report and observer ratings is to collect student reports of classroom instruction. The use of student reports to evaluate instructors is common practice in higher education for both undergraduate and graduate courses, with nearly all members of the Association of American Universities (AAU) engaged in the practice (AAU, 1995). Compared to its use in higher education, the use of student-report data at the K-12 level is much less frequent, with just 5% of U.S. school districts using such methods to study or evaluate teachers (Peterson, 2000).

Collecting data from student responses is typically faster and more cost-efficient than classroom observations (it can be done by the teacher, using a survey proctor, or online), the students may be less likely to over-report desirable actions than their teachers would, and the large volume of data from students' responses allows the exploration of relationships in the data. Previous literature has demonstrated much potential value for student surveys to gauge student dispositions, educational attainment, and instructional practices (e.g., Aleamoni, 1999; Marsh, 1984; Peterson, Wahlquist, and Bone, 2000). However, student surveys have frequently been phrased in very general terms (e.g., "I understand how to do assignments") rather than focusing on specific instructional practices. This issue is still present in an ongoing, large-scale study of methods for measuring instruction, the Measures of Effective Teaching (MET) project (Bill and Melinda Gates Foundation, 2010b), which includes general items about teachers' caring for students, such as "My teacher is nice to me when I ask questions" (Bill and Melinda Gates Foundation, 2010a, p. 12). While undoubtedly important aspects of instructional practice and the relationships between teachers and students, such general items do not measure the extent to which teachers implement model-based or other inquiry-oriented science instructional practices.

The IAS used in the present study builds on this literature. It focuses intentionally on the teachers' instruction practices as measures of effective teaching (Fenstermacher & Richardson, 2005). It is a student survey instrument, which provides proximal data on students' perceptions (Aleamoni, 1999; Marsh, 1984). Its questions are action-specific rather than general (Porter, 2002), which provides information on instructional practices particular to the science classroom that other methods do not (*cf*. Bill and Melinda Gates Foundation, 2010a). However, recognizing the potential burden on respondents (Porter, 2002), the survey was focused on a specific set of classroom activities which the study's professional development was intended to impact. More information on the instrument development is included in the Methodology section. In the next section, a review is provided of the professional development literature that informed the present study.

1.4 Professional Development

The National Science Education Standards (National Research Council [NRC], 1996) include four standards that stress that PD programs should: (1) promote teachers' science content knowledge through inquiry experiences, (2) help teachers' integrate content knowledge with pedagogical knowledge, (3) move teachers toward lifelong learning, and (4) be coherent and context-sensitive. Further research has explored the qualities of PD that can achieve these goals.

PD is effective when it is explicit in the intended content and pedagogy (e.g., Akerson, Abd-el-Khalick, & Lederman, 2000; Garet *et al.*, 2001), when it focuses on practical usage (e.g., Desimone, Porter, Garet, Yoon, & Birman, 2002; van Driel, Beijaard, & Verloop, 2001), and when professional learning communities are focused on a common goal (Johnson, Duvernoy, McGill, & Will, 1996; Lieberman, 2000). Further, PD should include reflection as a transformation of practice (e.g., Radford, 1998; Schön, 1983; Zeichner & Liston, 2006) and should be coherent, in that it should relate to the local influences on teachers' practice (Duran & Duran, 2005; Garet, *et al*., 2001). Projects that implement high quality PD programs for teachers are also able to demonstrate significant differences of the treatment on students' performance over time (e.g., Johnson, Kahle, & Fargo, 2007).

As reported in the previous research, teachers often struggle with their teaching when engaging their students in a model-centered inquiry environment (Schwarz, & Gwekwerere, 2006; Wells, Hestenes, & Swackhamer, 1995). Many teachers themselves have never learned science through inquiry or model-based instruction and, therefore, do not understand the use of models and modeling processes in science (Van Driel, & Verloop, 1999; Schwarz, & Gwekwerere, 2006). For PD on modeling instruction to be fruitful, teacher participants first need to be engaged in a cooperative, inquiry-oriented learning environment and experience the model-based curriculum materials as learners. Additionally, the teachers need to participate in discussions to share ideas with the members of a professional community and in reflections on their pedagogy as teachers. These features were incorporated into the PD implemented for this study; further detail on the PD is presented in the methodology section.

2.1 Methodology

This study examines students' responses to a survey about their teachers' instruction to identify common factors in such instruction and to explore differences in the instruction between the comparison and modeling groups. In doing so, it contributes to the study of teaching practices and the impact of professional development on instruction, by advancing alternative methods for measuring teacher actions. The following sections describe the research sample and setting, the professional development (PD), the instrument development, and the analyses.

2.2 Research Sample and Setting

The present study was part of a comparative case study involving two high schools in the northeast region of the United States to study the impact of modeling instruction PD on teaching practices and student learning. The two schools were selected based on closely matched student demographics and similar scores on statewide standardized reading and mathematics tests. The students in both schools were predominantly Whites (91-95%) from middle income households as defined by the state ($37,501 to $57,000). Student participants completed the IAS about their teachers' instructional practices toward the end of the respective introductory physics course (offered every semester following a block schedule), so that their responses would be indicative of their overall experiences in class. Completed data was available for 228 students from the treatment and comparison schools. The modeling instruction classes contained 72 students—49 in ninth-grade and 23 in twelfth-grade—taught by three instructors in respective sections, with about 24-28 students per class section. The comparison group classrooms contained 156 students—46 in eleventh-grade and 110 in twelfth-grade—taught by two instructors, with about 20-24 students per class section.

The model-based physics program was developed in previous NSF-funded projects based at the Arizona State University (Wells, Hestenes, & Swackhamer, 1995). Built on the learning cycle approach designed by Robert Karplus for the Science Curriculum Improvement Study (SCIS; Karplus, 1977), and the Modeling Theory of Physics Instruction (Wells, Hestenes, & Swackhamer, 1995), the curriculum or course content is organized around a small set of basic models, while instruction is organized into *modeling cycles* which move students systematically through all phases of model development, evaluation, and application in concrete situations—thus developing skills and insight in the procedural aspects of scientific knowledge.

2.3 Professional Development

All three physics teachers in the treatment school participated in a three-week-long, intensive summer institute on modeling-instruction prior to their implementation of the inquiry-based and model-centered program mandated by the school district. During the summer institute, teacher leaders facilitated and modeled the instructional practices by engaging teachers as learners of physics and of physics pedagogy. The teacher participants were introduced to the model-based pedagogy as a systematic approach to the design of curriculum and instruction through: 1) examining implications of educational research in physics learning and teaching; 2) rotating between roles of student and instructor as they practicing instructional strategies that engage and guide learners in cooperative inquiry, developing and applying models, evaluating evidence, and conducting discourse; 3) exploring ways to integrate computer technology and electronic resources in physics teaching; and 4) collaborating on rethinking and redesigning the high school physics course and curriculum materials for enhanced learning. The teacher participants were also required to take a force concept inventory and other evaluation instruments to become aware of likely student misconceptions or naïve non-scientific understanding, then given opportunities to discuss their ideas with colleagues and frequently reflect upon their experience in their journals. In addition, in line with the literature on the development of professional learning communities (PLCs; e.g., Vescio, Rossa, & Adams, 2008; Webster-Wright , 2009), all PD participants in the study were connected through a nation-wide modelers' list-serve for ongoing communication and knowledge sharing immediately after the summer institute.

The teachers in the comparison group did not receive PD on modeling instruction during the study. It was expected that students in the comparison classes would experience more traditional lecture and confirmative laboratory instruction, and the students in the modeling classes would be engaged in guided inquiry through model development, evaluation, and application.

2.4 Instrument

The Instructional Activities Survey (IAS) was developed based on the key features of the model-centered approach (Wells, Hestenes, & Swackhamer, 1995), the *Fundamental Abilities of Inquiry* (grades K-12) section of the *National Science Education Standards* (NRC, 1996), and instructional survey items released from the Trends in International Math and Science Study (TIMSS). The IAS was designed to distinguish between the model-based instructional practices promoted by the PD workshop and lecture-based instruction. Both model-based and lecture-based instructional practices are included in IAS items, because it was not assumed that teachers who participated in the PD would necessarily implement all aspects of model-based instruction. Additionally, it was not assumed that all comparison group teachers would necessarily use lecture-based instruction. That is, this study avoided the presumption that instruction in comparison classrooms would necessarily be "commonplace" (Wilson, Taylor, Kowalski, & Carlson, 2010, p. 282) by including primarily teacher-led discussions, presentations, demonstrations, or performing verification laboratories.

The IAS asks students to rate how often they completed various actions in class. The instrument contained 24 items that included inquiry-oriented, model-based instruction such as "Develop conceptual models using scientific evidence," and more traditional lecture-lab type of instruction such as, "Listen to the teacher's lecture-style presentations." All items used a four-point Likert-type scale (1= Never or almost never; 2= Sometimes; 3= About half of the lessons; and 4= Most of the lessons). The complete instrument and relevant information were provided in a different article (Authors, in press).

2.5 Analyses

The IAS data were prepared using factor analysis and Rasch modeling. Factor analyses were conducted using the *fact_anal* package in the R statistical environment (Ihaka & Gentleman, 1996); all Rasch model estimation was conducted using the WINSTEPS software package (Linacre, 2007). An exploratory factor analysis (Lawley & Maxwell, 1962; van Prooijen & van der Kloot, 2001) indicated that three factors were appropriate for the items, based on examination of the scree plot (*cf*. Floyd & Widaman, 1995) and the proportion of variance explained: 41% of the variance was explained by the three factors, with a fourth factor accounting for only an additional 2% of variance. While 41% explained variance is low, in the current study the three-factor solution is consistent with the scree plot and allows a parsimonious measurement model for the newly-developed instrument. Each factor was then analyzed separately using a polytomous Rasch model (Andrich, 1978) in WINSTEPS. Rasch measurement analysis provides benefits over classical item analysis in that it simultaneously calculates measures for items and persons, and estimates the reliability for both item and person measures. Rasch modeling is the basis for many standardized tests, such as the Programme for International Student Assessment (PISA; OECD, 2009) and many US state-wide tests such as

Ohio (*cf*. Ohio Department of Education, 2011) and Texas (*cf*. Texas Education Agency, 2005). Items were not reverse-coded before the Rasch analyses; the estimation of Rasch measures allows items to have negative measures, so reverse coding is not necessary.

During the Rasch modeling stage, items were dropped from the model that showed poor fit statistics (with z-scores of magnitude greater than 2; Bond & Fox, 2001). Each scale was estimated such that the student measures, in logit units, would have mean of 0 and standard deviation of 1. However, because the scales use separate items, the logit measures for each scale are not necessarily equal size. Table 1 presents the factor loadings and Rasch measures for the items retained in the final Rasch model for each of the three subscales.

INSERT TABLE 1 ABOUT HERE

The Rasch model estimation yielded measures of the IAS subscales for each student. Based on review of the item text within each subscale, the subscales were renamed *Modeling and Reflecting* (MR), *Communicating and Relating* (CR), and *Investigative Inquiry* (II). Table 2 presents sample items for each subscale. Item-reliability coefficients for the three subscales are: 0.98 (MR); 0.96 (CR); and 0.92 (II). Person-reliability coefficients for the three subscales are: 0.82 (MR); 0.29 (CR); and 0.62 (II). A separate study (Authors, in press) describes the relationship of IAS measures with other measures of instruction such as RTOP.

INSERT TABLE 2 ABOUT HERE

After preparation, the students' Rasch IAS measures were analyzed using linear mixed-effects regression. These regression models were calculated using the *lme4* package in the R statistical environment (Ihaka & Gentleman, 1996). The predictor variables were entered as dummy-codes (0 or 1), so that results would be identical to analysis of variance (ANOVA). Since students in the same class will experience similar instruction, their IAS subscale measures are not necessarily independent. To control for the hierarchical nature of the data, the regression models used classroom as a nesting term with both Intercept and Treatment as random-coefficient variables. This nesting provides more accurate estimates of standard error (Raudenbush & Bryk, 2002, p. 116). While it would also be possible to include teacher as a nesting variable, there was inadequate power at this level to support this analysis. Additionally, it was anticipated that teachers may use differing instructional practices in different classrooms, depending on the students comprising the class and other factors beyond the control of the study. The data exhibited intraclass correlations (ICC; Raudenbush & Bryk, 2002, p. 36) of 0.324 (MR), 0.067 (CR), and 0.201 (II) for the three subscales, indicating that between 6% and 32% of the variance in the data was attributable between classes (rather than between individuals).

3.1 Results

Analyses indicated that there were significant treatment effects for all three IAS factors (Table 3). The coefficients, standard errors, and t-statistics are calculated using linear mixed-effects regression with nesting within classroom. Therefore, the standard errors are more conservative than would be obtained using traditional, student-level regression. The modeling-group students' IAS Rasch measures were significantly different from the comparison group students' ratings for all three factors.

INSERT TABLE 3 ABOUT HERE

As Table 4 shows, the treatment group had much higher IAS Rasch measures than the comparison group, with moderate to high effect sizes (ES; Cohen, 1988) for all subscales. The higher effect size for the MR (*Modeling and Reflecting*) subscale (ES = 1.25) and II (*Investigative Inquiry*) subscale (ES = 0.98) correspond with the intervention professional development's focus on modeling and inquiry instruction. The CR (*Communicating and*

Relating) subscale showed relatively lower differences between treatment and comparison teachers (ES = 0.55).

INSERT TABLE 4 ABOUT HERE

4.1 Discussion and Conclusions

The present study examined the use of student survey responses in evaluating the effect of a professional development on teachers' instruction. The results indicate that the students of teachers from a model-based professional development (PD) program showed higher incidence of model-based and inquiry-oriented instruction. The effect size estimates reveal that the modeling-group students experienced much greater use of modeling and reflecting (MR) instructional strategies than did comparison group students (effect size of 1.25) as well as investigative inquiry (II) instruction (effect size of 0.98). This is consistent with the PD program's emphasis, which implemented modeling-based teaching as a core example of inquiry-oriented pedagogy.

The results showed a lower effect size found for the communicating and relating (CR) scale (effect size of 0.55) for the difference between the modeling and comparison-group students' experiences. The professional development did address communication with students, such as through Socratic dialogue. However, the lower value reflects that the communication aspect may not have been as well-developed in the PD. This indicates that long-term support and mentoring on communication strategies may be needed to demonstrate fully the effects of this aspect of the PD.

This study's findings demonstrate that students' survey responses may identify differences in their experiences that relate to their teachers' participation in a PD program. The end goal of a teacher preparation or professional development program naturally must include student performance (Levine, 2006). The study and refinement of teacher professional development efforts must also include understanding how a PD program affects teachers' instruction and, subsequently, how this relates to changes in what students know or can do. This reflects the continued attention to unpacking the differences in teachers' instruction that result from teacher PD, and the subsequent effects on student learning outcomes (Author, 2008; Desimone, 2009).

This study piloted an instrument, the Instructional Activities Survey (IAS). The IAS combined released survey items from TIMSS and new items that reflect the model-centered approach (Wells, Hestenes, & Swackhamer, 1995) and the *Fundamental Abilities of Inquiry* (grades K-12) section of the *National Science Education Standards* (NRC, 1996). The IAS was intended to capture both model-based and lecture-based instruction. The instrument development and study design did not assume that teachers who participated in the professional development would necessarily exhibit more model-based instruction or that comparison teachers would only use lecture-based instruction. The exploratory factor analysis and Rasch modeling procedures used would allow these practices to be treated independently if borne out by the data. However, the analysis demonstrated that the same factor had positive loadings for aspects of the model-based instruction and negative loadings for some aspects of lecture-based instruction (see Table 1). This suggests that, in the present sample, these practices were negatively associated rather than orthogonal: students who reported experiencing more modeling instruction also reported less lecture instruction, in general.

This study's contributions to the literature are twofold. First, it focuses in particular on measurement of science instruction from the student perspective. By contrast, other studies that use student surveys have examined teachers' dispositions (Aleamoni, 1999; AAU, 1995) or have used general items that do not reflect science content or potential differences between inquiry-oriented and lecture-based science instruction (Bill and Melinda Gates Foundation, 2010a). The IAS uses students' responses about their classroom experiences related to either model- or lecture-based instruction. The present study fits within current efforts to incorporate a broader suite of measurement methods to the assessment of instruction in addition to student outcome measures, such as teacher self-reports, classroom observations, and student reports. As described previously, student-report data are the most proximal to the students' experience of the class; so, this line of work is promising for the development of measures of teachers' practices using student report data.

Secondly, this study contributes to the literature through its examination of latent factors in students' experiences of instructional practices using factor analysis and Rasch modeling. Unlike traditional survey item analysis methods, the present study explored the item fit and construct dimensionality for students' experience of instruction. Therefore, the present study demonstrates an approach to student survey analysis that builds on latent factor approaches that will allow more objective measurement of students' classroom perceptions.

4.2 Implications

Previous research has indicated that the adoption of model-based curriculum improves student outcomes. The present study attempted to understand the mechanisms underlying this improvement by identifying meaningful constructs associated with classroom activities in model-centered instructions. It has two implications for professional development implementation and future research. First, it provides initial evidence that the model-based professional development program used in the study was associated with greater incidence of model-based classroom instruction. To promote more effective teaching, future teacher PD providers may want to spend more time on how to cultivate teachers' expertise in *Modeling and Reflecting*, *Communicating and Relating*, and *Investigative Inquiry*. This result is suggestive, but there are limitations. Data on teachers' instructional practices prior to the workshop were not available for this analysis. Future research should collect data on participating and comparison teachers' classroom instructional practices prior to the PD, which would allow for stronger causal links between the professional development and changes in teachers' instructional practices, and with a larger sample of teachers and schools.

Second, despite the potential usefulness of the IAS, there is also significant room for improvement. As shown in Table 1, few items were retained for Subscale 2 (Communicating and Reflecting). As summarized in the Methodology section, the IAS had strong item reliability (ranging from 0.92 to 0.98), but person reliability measures ranged widely, from 0.29 for CR to 0.82 for MR. The low person reliability for CR is attributable to the low number of items that were retained in the Rasch model due to low person-item fit indices. The reliability of the factors and the number of items retained in the Rasch measure calculation suggest that additional work is required—refining the items and developing additional items to bolster each factor—before the IAS instrument will be appropriate for broader use.

5.1 References

Adamson, S. L., Banks, D., Burtch, M., Cox III, F., Judson, E., Turley, J. B., Benford, R., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, *40* (10), 939-957.

Aleamoni, L.M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153-166.

Akerson, V. L., Abd-El-Khalick, F., & Lederman, N. G. (2000). Influence of a reflective explicit activity-based approach on elementary teachers' conceptions of Nature of Science. *Journal of Research in Science Teaching*, *37*, 295-317.

American Association for the Advancement of Science (AAAS). (1989). *Science for all Americans*. New York: Oxford University Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-73.

Association of American Universities. (1995). *Survey of Undergraduate Education Activities*. Washington, DC: Author.

Author. (2008). Book.

Authors. (2009). Science Education.

Author. (2011). Journal of Educational and Behavioral Statistics.

Authors. (in press). Journal of Science Education and Technology.

Banilower, E. R. (2005). *A study of the predictive validity of the LSC Classroom Observation Protocol*. Chapel Hill, NC: Horizon Research, Inc.

Beeth, M. E., & Hewson, P. W. (1999). Learning goals in an exemplary science teacher's practice: Cognitive and social factors in teaching for conceptual change. *Science Education*, *83*, 738-760.

Bill & Melinda Gates Foundation. (2010a). *Learning About Teaching – Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Author. [Accessed November 1, 2010 from http://www.metproject.org/reading.]

Bill & Melinda Gates Foundation. (2010b). *Working with teachers to develop fair and reliable measures of effective teaching*: *Framing paper of the Measures of Effective Teaching project*. Seattle, WA: Author. [Accessed November 1, 2010 from http://www.metproject.org/reading.]

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd Ed.). New York, NY: Routledge.

Carnegie Corporation of New York (2009). *The Opportunity Equation: Transforming Mathematics and Science Education for Citizenship and the Global Economy*. New York: Author.

Chambers, J. G., Lam, I., & Mahitivanichcha, K. (2008). *Examining context and challenges in measuring investment in professional development: A case study of six school districts in the Southwest Region*. Washington, DC: US Department of Education Institute of Education Sciences.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Darling-Hammond, L. & Baratz-Snowden, J. (2005). *A good teacher in every classroom: Preparing the highly qualified teachers our children deserve*. New York, NY: John Wiley & Sons.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, (3), 181–199.

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

Duran, E., & Duran, L. B. (2005). Project ASTER: A model staff development program and its impact on early childhood teachers' self-efficacy. *Journal of Elementary Science Education*, *17* (2), 1-12.

Fenstermacher, G. D. & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107* (1), 186-213.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299. doi:10.1037/1040-3590.7.3.286

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

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5* (3), 299-314.

Johnson, D.K., Duvernoy, R., McGill, P., & Will, J.F. (1996). Educating teachers together: Teachers as learners, talkers, and collaborators. *Theory into Practice*, *35* (3), 173-178.

Johnson, C. C., Kahle, J. B., & Fargo, J. D. (2007). A study of the effect of sustained, whole-school professional development on student achievement in science. *Journal of Research in Science Teaching*, *44* (6), 775-786.

Judson, E., & Lawson, A. E. (2007). What is the role of constructivist teachers within faculty communication networks? *Journal of Research in Science Teaching*, *44* (3), 490-505. DOI: 10.1002/tea.20117

Koziol Jr., S. M., & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research*, *79* (4), 205-209.

Lawley, D. N. & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society*. Series D (The Statistician), *12* (3), 209-229.

Lawrenz, F., Wood, N. B., Kirchhoff, A., Kim, N. K., & Eisenkraft, A. (2009). Variables affecting physics achievement. *Journal of Research in Science Teaching*, *44* (6), 775-786.

Levine, A. (2006). *Educating School Teachers*. New York: Education Schools Project.

Lieberman, A. (2000). Networks as learning communities: Shaping the future of teacher development. *Teacher Education*, *51*, 221-227.

Linacre, J.M. (2007). WINSTEPS (Version 3.61.2) [Computer Software]. Chicago: Winsteps.com.

Loucks-Horsley, S., & Matsumoto, C. (1999). Research on professional development for teachers of mathematics and science: The state of the scene. *School Science and Mathematics*, *99* (5), 258.

MacIsaac, D., Sawada, D., & Falconer, K. (2001). Using the Reformed Teaching Observation Protocol (RTOP) as a catalyst for self-reflective change in secondary science teaching. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76* (5), 707-754.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47* (4), 474-496.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Washington, DC: Author. [Accessed online December 29, 2009, from: http://standards.nctm.org/document/chapter3/index.htm]

National Research Council (NRC). (1996). *National science education standards*. Washington D.C.: National Academy Press.

Neale, D. C., Smith, D. C., & Johnson, V. G. (1990). Implementing conceptual change teaching in primary science. *Elementary School Journal*, *91* (2), 109-132.

OECD. (2009). The Rasch Model. In OECD (author), *PISA Data Analysis Manual: SPSS* (2nd Ed.). Paris: OECD Publishing. doi: 10.1787/9789264056275-6-en

Ohio Department of Education (2011). OHIO ACHIEVEMENT ASSESSMENTS MAY 2011 ADMINISTRATION: STATISTICAL SUMMARY. Columbus, OH: Author. [Retrieved 1 December 2011 from http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=107732/]

Park, S., Jang, J.-Y., Chen, Y.-C., Jung, J. (2011). Is pedagogical content knowledge (PCK) necessary for reformed science teaching? Evidence from an empirical study. *Research in Science Education*, *41* (2), 245-260. DOI: 10.1007/s11165-009-9163-8.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14* (2), 135-153.

Peterson, K.D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. (2nd ed.). Thousand Oaks, CA: Corwin Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Sawada, D., Piburn, 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.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction*, *23*, 165-205.

Schwarz, C.V., & Gwekwerere, Y. N. (2007). Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education*, *91,* 158-186.

SERVE. (2006). CAPE Evaluation Framework: Looking For Technology Integration (LoFTI). SERVE Center-UNC Greensboro. [Retrieved August, 2009, from http://www.serve.org/Evaluation/Capacity/EvalFramework/resources/LoFTI.php]

Texas Education Agency. (2005). *Technical Digest for the Academic Year 2004-2005*. A Collaborative Effort of the Texas Education Agency, Pearson Educational Measurement, Harcourt Educational Measurement, and Beck Evaluation and Testing Associates, Inc.

Austin, TX: Author. [Retrieved 1 December 2011 from http://www.tea.state.tx.us/student.assessment/]

Van Driel, J. H. and Verloop, N. (1999) Teachers' knowledge of models and modeling in science. *International Journal of Science Education*, 21, 1141–1153.

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

van Prooijen, J.-W., & van der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement*, *61*, 777-792.

Venville, G., Sheffield, R., Rennie, L. J., & Wallace, J. (2008). The writing on the wall: Classroom context, curriculum implementation, and student learning in integrated, community-based science projects. *Journal of Research in Science Teaching*, *45* (8), 857-880.

Vescio, V., Rossa, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education*, *24* (1), 80-91.

Vesenka, J., Beach, P., Munoz, G., Judd, F., and Key, R. (2002). A comparison between traditional and "modeling" approaches to undergraduate physics instruction at two universities with implications for improving physics teacher preparation. *Journal of Physics Teacher Education Online*, *1* (1), 3-7. [Retrieved March 2, 2007, from http://phy.ilstu.edu:16080/jpteo/issues/june2002.html]

Waight, N., & Abd-El-Khalick, F. (2007). The impact of technology on the enactment of "inquiry" in a technology enthusiast's sixth grade science classroom. *Journal of Research in Science Teaching*, *44* (1), 154-182.

Webster-Wright, A. (2009). Reframing professional development through understanding authentic professional learning. *Review of Educational Research*, *79* (2), 702–739. DOI: 10.3102/0034654308330970

Wells, M., Hestenes, D., and Swackhamer, G. (1995). A modeling method for high school physics instruction. *American Journal of Physics*, *63*, 606-619.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, *47* (3), 276-301.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14* (2), 97-116.

Yerrick, R., Parke, H., & Nugent, J. (1997). Struggling to promote deeply rooted change: The "filtering effect" of teachers' beliefs on understanding transformational views of teaching science. *Science Education*, *81* (2), 137-159.

Zeichner, K., & Liston, D. P. (2006). *Teaching student teachers to reflect*. In D. Hartley & M. Whitehead (Eds.) Teacher education: Professionalism, social justice and teacher education (Volume IV; 5-34). New York: Routledge.

Acknowledgements

Portions of this work were supported by a grant from the National Science Foundation (NSF; award number DUE 03-14806) and by an Independent Research and Development (IR/D) project to the first author. Any opinions expressed are those of the authors, and do not necessarily reflect the views or policies of the NSF.

Tables

Table 1. Factor loadings and Rasch scale measures for IAS items.

Item	Factor Loading	Rasch Measure	Model SE
		Subscale 1	
g	0.51	-0.30	0.10
h	0.58	0.45	0.10
i	0.60	0.61	0.10
m	0.71	0.89	0.10
o	0.50	-1.50	0.11
q	0.52	-0.40	0.10
r	0.73	0.30	0.10
x	0.67	-0.04	0.10
		Subscale 2	
p	0.47	0.02	0.11
v	0.99	-0.73	0.12
w	0.53	0.71	0.10
		Subscale 3	
a	-0.31	-0.83	0.10
c	0.43	0.81	0.09
d	0.91	1.04	0.09
e	0.66	1.00	0.09
f	0.33	0.25	0.09
s	0.46	-2.27	0.15

Note. For item texts, please see Appendix A. For subscale 1, item reliability is 0.98 and person reliability is 0.82. For subscale 2, item reliability is 0.96 and person reliability is 0.29. For subscale 3, item reliability is 0.99 and person reliability is 0.62.

Table 2. Sample items from the three IAS subscales

Modeling and Reflecting [MR]	• Recognize and analyze alternative explanations by weighing evidence and examining reasons. • Develop conceptual models using scientific evidence. • Reflect on our own thinking and learning.
Communicating and Relating [CR]	• Work together in small groups to discuss our ideas. • Relate what we are learning in science to our daily lives.
Investigative Inquiry [II]	• Ask scientifically oriented questions. • Formulate our own hypotheses or predictions to be tested in an experiment or investigation. • Listen to the teacher's lecture-style presentations. (negatively loaded).

Table 3. Combined results from univariate analyses of treatment effects on the IAS Rasch subscale scores

Variable	Source	Coefficient	SE	t-value	p
MR	Intercept	-0.293	0.204	-1.438	>.1
	Treatment	1.740	0.256	6.809	<.001
CR	Intercept	1.225	0.125	9.800	<.001
	Treatment	0.816	0.229	3.564	<.001
II	Intercept	0.477	0.097	4.940	<.001
	Treatment	1.181	0.172	6.874	<.001

Note: For estimating p-values in all significance tests, df=213; Control group N=156, Modeling group N=72.

Table 4. Means and standard deviations by treatment group:

Measure	Comparison		Modeling		Effect Size
	M	SD	M	SD	d
MR	-0.264	1.396	1.447	1.301	1.25
CR	1.227	1.525	2.040	1.387	0.55
II	0.477	1.181	1.658	1.258	0.98

Note: Control group N=156, Modeling group N=72.