A Reform-Based Framework for Observing Teaching

Rebecca Schneider
University of Toledo, Toledo, OH 43606-3390
Tel: 419-530-2504, Fax: 419-530-8459
Email: Rebecca.Schneider@UToledo.edu

Joseph Krajcik & Phyllis Blumenfeld
University of Michigan

Abstract: For research on teaching to succeed in providing meaningful information, a method to examine teaching in complex classroom settings that is also feasible on a larger scale is needed. Our goal was to design a systematic method for observing classroom teaching that was consistent with reform recommendations and adaptable to large scale use. Our work is embedded in an ongoing urban systemic initiative of a large public school district to reform science and mathematics education. Middle school teachers' enactments of a reform-based science unit were videotaped. Student achievement measures included low, medium, and high cognitive level items developed to match science concepts addressed the unit. Analysis identified specific criteria within seven main analysis categories consistent with reform oriented instructional practices were associated with students' achievement scores. Ideas for adapting this framework to large scale use are discussed.

Introduction

Teaching and its measure are critical components of efforts to promote student learning (NCTAF, 1996). Reformers exploring ways to support exemplary teaching and thus advance student learning, depend on suitable methods for gathering information on what is effective. Based on goals for student learning in science, reformers are exploring new ways to help teachers learn how to use inquiry with collaboration supported by use of technology tools to support students in actively constructing deep understanding of important science concepts (NRC, 1996). In our work we are exploring the use of reform-based curriculum materials in conjunction with professional development to promote exemplary teaching in science. This work has led us to examine teachers' classroom practices. In this paper we describe the development of a method to evaluate complex classroom observations that captures the salient features of reform-based teaching and is feasible on a larger scale.

As part of an ongoing systemic initiative of a large urban public school district, we have developed science materials to reflect desired reforms and provide teachers with needed support to learn and enact innovative curriculum. Developers created materials based on the premises of project-based science and were guided by design principles that include: contextualization, alignment with standards, sustained student inquiry, embedded learning technologies, collaboration and discourse, assessment techniques, and scaffolds and supports for teachers (Schneider & Krajcik, 2002; Singer, Marx, Krajcik, & Clay-Chambers, 2000). Professional development opportunities were designed to support reform teaching consistent with the science materials (Fishman & Best, 2000). We were interested in a scalable method of analyzing classroom enactment data to gain meaningful information on which to base revisions of materials and improve support for teachers in learning and enacting new instructional practices.

Observation of classroom teaching is essential to improve our understanding of how to help teachers learn and enact reform-based practices (Anderson & Helms, 2001). Researchers interested in understanding how to support improved teaching consider classroom observation essential to determining the success of their efforts to change teachers' practice (Blumenfeld, Krajcik, Marx, & Soloway, 1994; Palincsar, Magnusson, Marano, Ford, & Brown, 1998; Wood, Cobb, & Yackel, 1991). However, the rich descriptions provided by qualitative methods are time and labor intensive necessitating the observation of only a few teachers. Careful observation and analysis of classroom events including teachers' behaviors and statements is required. Data from a variety of classrooms is needed to develop truly effective programs. We also are interested in the scalability of our reform-based materials and therefore a measure of teaching that is less cumbersome than detailed descriptions of classroom events.

Researchers attempting to identify specific factors that influence student achievement are examining national and state level data. From this work we have some evidence that the quality of teaching is related to student outcomes. For instance, Darling-Hammond (1999) examined state level data on teacher preparation, certification, and experience along with changes in student achievement over several years. She describes teacher professional development as the most important means to improving student achievement scores. However, this approach does not identify what these teachers are doing in the classroom to impact student learning. Likewise, work by Sanders and Horn (1998) indicates a long term affect of individual teachers on student achievement scores. But again this work does not describe what this quality teaching looks like in a classroom. Therefore, these studies cannot point to the features of teacher preparation, knowledge, or experience that are particularly worthwhile. Although quantitative measures are feasible on a large scale they fail to capture the true complexity of what happens in classrooms. This leaves the topic of how to improve teaching and student outcomes open to debate (Cochran-Smith & Fries, 2001).

One approach that merits further development is classroom observation research that links specific curriculum to teachers' instruction (Collopy, 1999; Prawat & et al., 1992; Remillard, 1999). In these studies, analysis of observations is guided by frameworks which are based on recommended curriculum. This approach is facilitated when researchers describe their curriculum in terms of reform guidelines. The reform-based curriculum that is the focus of this study is one such example. A better understanding of how teachers and students interact around specific materials and ideas in classrooms is needed (Ball, 2000). Similarly, measures of student achievement matched to specific curriculum are more likely to capture the impact on student achievement than general measures. (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). In this study we used a reform-based science curriculum unit to develop a framework for observing teaching and linked our observation results to student outcomes also measured by curriculum specific assessments. One outcome of this work is the beginning of a scalable scoring rubric to measure quality of teaching in comparison to curriculum goals.

Methods

This study was conducted in five urban middle schools located in low SES neighborhoods selected to participate in initial stages of the reform effort (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000). Students in these schools were predominantly African American (95% to 100%) with high percentages of students receiving free or reduced lunch (29% to 66%). Scores on local and statewide achievement testing in science were reported as below grade level in four of the five schools. Curriculum material development was considered an essential component of the change effort, particularly to facilitate change within classrooms on a large scale (Singer et al., 2000). To support teachers in science reform, project-based materials were developed to address important science ideas, offer multiple learning opportunities, and provide appropriate instructional supports for students. Teachers were introduced to units at a 2-week summer institute and were supported by monthly workshops (Fishman & Best, 2000). Teachers participating in the reform used materials for an eight-week unit on force and motion, water quality, or simple machines. Teachers were selected for observation based the unit they were enacting and times researchers could visit classrooms. Teaching experience ranged from 6 to 20 years. Prior to the reform effort, each had limited experience with one or more of the following aspects: project-based science, unit specific science content, or the use of technological tools to support inquiry. Although they were not selected as a statistically random sample, their disparate backgrounds made this group representative of middle school science teachers across the district.

Data Analysis

Beginning with four teacher enacting the force and motion unit, five lesson sequences containing experiences with phenomena, investigation, technology use, or artifact development, spanning 3-5 days each were selected for analysis. These lesson sequences were selected because each represented different aspects of inquiry teaching that were to be used to focus descriptions of classroom enactments. These aspects included how teachers a) presented science ideas, b) promoted students' use of inquiry, c) used technology to promote student inquiry and concept development, d) used collaboration to promote student inquiry and concept development, and e) supported and assessed concept development through student artifacts. Detailed descriptions of classroom events were written from the videotape for each lesson sequence and teacher to facilitate initial code development and final coding. Teacher and student behavior and conversation were described in light of the lesson sequence descriptions in the materials. As these descriptions were prepared, we looked for and described: 1) science ideas (content and process ideas presented), 2) contextualization (referring to the driving question or anchor ideas, using real life examples, stating value), 3) linking ideas to previous or future lessons or to other ideas, 4) directions given, 5) emphasis given—such as what ideas or tasks are important, 6) specific strategies such as POE, 7) specific representations such

as motion graphs, 8) scaffolding (modeling, coaching, feedback, or asking for justifications or reasons), and 9) group work (teacher statements on group work, teacher role during group work). We also noted suggested lesson sequences or portions of lesson sequences that were enacted, omitted, or adapted.

The coding scheme used was designed to capture three aspects of enactment—presentation of science ideas, opportunities for student learning, and support to enhance the learning opportunities—each in comparison to what was intended in the materials. Coding schemes used in this analysis were developed through an iterative process of creating codes, coding, modifying and refining codes, and recoding consistent with Miles and Huberman's (1994) recommendations for rigorous and meaningful qualitative data analysis. The independent coding of several enactment episodes by another science education researcher familiar with the curriculum materials assessed reliability of the coding process. Reliability was 88%. After the categories and rating levels were finalized and reliability established, all enactment data were recoded with the final codes. The final coding scheme is presented in Table 1. Entire episodes and the type of activity were considered to assign a rating for each category. A short statement of evidence or justification was written for each assigned rating.

Table 1. Categories and rating levels of coding scheme used to analyze classroom enactment data

Accuracy	**Opportunities**	**Instructional Supports**
4 Scientific - all ideas consistent with current scientific ideas	*4 Maximum* - includes ample (number or time) opportunity for student learning	*4 High* - provides many supports for student thinking
3 Sufficient -consistent for main ideas, inaccurate for minor ideas	*3 Sufficient* - includes some opportunities	*3 Medium* - provides some supports for student thinking
2 Semi accurate - inconsistent some main ideas	*2 Insufficient* - includes few opportunities	*2 Low* - provides few supports for student thinking
1 Non scientific - inconsistent for many main ideas	*1 Minimal* - includes almost no opportunities for student learning	*1 None* - provides no supports
Completeness	**Similarity**	**Appropriateness**
4 Thorough - all the appropriate science ideas are addressed	*4 High* - matched to intended lesson	*4 Excellent* - instructional supports always matched to student learning needs
3 Sufficient - all the appropriate main ideas are addressed but some minor ideas are missing	*3 Medium* - closely resembles intended	*3 Sufficient* - supports usually matched to learning needs
2 Incomplete - missing some main ideas	*2 Low* - faintly resembles, major changes	*2 Insufficient* - instructional supports usually not matched to student learning needs
1 Insufficient - missing several main ideas	*1 None* - not consistent with intended	*1 Poor* - supports always not matched to learning needs
	Adaptation	
	4 High - consistent with learning goal and appropriate for students' learning needs	
	3 None - not adapted	
	2 Medium - consistent with learning goal but <u>not</u> appropriate for learning needs	
	1 Low - not consistent with learning goal	

Assigning ratings. The categories of *accuracy* and *completeness* were included to capture information about the science ideas presented by teachers. Both content and process ideas were considered as well as whether the ideas presented were defined as a main or minor idea. The main ideas were defined as those identified in the purpose, objectives, or assessments of the materials for that lesson sequence. Minor ideas were defined as ideas secondary, related, or supporting the main ideas. Teachers presented ideas in a variety of ways. This included teachers' statements, examples, demonstrations, hints, or other types of guidance regarding science ideas. A teacher's response or lack of response to students' actions or statements was also judged as giving students information about science ideas. In this case, a teacher may not have directly stated ideas accurately or inaccurately but, by the type of response they gave, implied that inaccurate student statements were acceptable or vise versa. The rating of accuracy was unrelated to the rating of completeness. Also, unlike any other category, completeness included one rating that could apply in addition to the other ratings. The rating of excessive was used to indicate content related but beyond that intended for students in this unit. A teacher could be incomplete in covering the intended content, yet also excessive by adding other related content.

The categories of opportunities, similarity, and adaptation each refer to the learning opportunities for students. *Opportunities* for student learning included both teacher lead and small-group activities. Take-home activities that were incorporated into class activities were included as opportunities, but work completed entirely at home was not. Opportunities were rated high if the number and time spent was high in relationship to the amount of class time represented in the episode. *Similarity* was rated by considering both that opportunities observed were

intended by the materials, but also that they were in a similar sequence with approximately the same emphasis. For example, if a teacher directed students to make a prediction, but did not allow time for writing the predictions or for sharing some of the predictions in class before the observation phase, similarity would be rated low. *Adaptations* were opportunities provided that were not described in the materials. These activities were judged on whether or not they addressed content specified for the learning sequence and if the activity was likely to help students learn the content. Replacing a discussion of observed phenomena with practice defining terms would be rated as low. The terms may be the ones intended for use but understanding of relationships or application of ideas was the intended learning goal rather than the memorization of definitions. On the other hand, making an investigation more open by allowing students more choices in what to test would be rated high if students appeared to be ready to design an investigation with reduced structure.

The categories of instructional supports and appropriateness each refer to the instructional support for student thinking. *Instructional supports* included wide variety of teacher actions and statements that had the potential to enhance the learning opportunities. These included supports for student thinking as well as supports for organizing and carrying out tasks. Examples included, but were not necessarily limited to: modeling thought processes or actions, coaching, giving hints, using examples, monitoring small-group work, giving reminders, asking for reasons or justification, structuring student work, offering guidance, and giving feedback. Instructional supports were rated high if the number of supports was high. Whether or not the supports appeared that they would help students learn the intended science content was judged in the category of *appropriateness*. Therefore, an episode could be rated high for supports if a teacher gave students many hints, but poor for appropriateness if those hints were likely to lead students in the wrong direction or did not match the type of difficulty students were exhibiting.

Summarizing ratings. Ratings were then summarized across episodes for each lesson sequence. The ratings and the justification statements in each category were compared sequentially for all enactment episodes. Then a judgment was made for a rating of the entire lesson sequence. A justification statement was also written for each lesson sequence rating based on a summary of the individual statements. To guide the summarization process a set of guidelines were developed. When variation was evident, summarizing was done in a way that appropriately reflected the variation in the final rating and justification statement. If the variation was minor, one rating was given but the variation was described in the justification statement. However, when variation was more pronounced, two or more ratings were assigned and the lesson sequence was labeled as varied. The final analysis phase was to examine the coded lesson sequences for patterns across lesson sequences and teachers. Each category was traced across all lesson sequences for each teacher. Justifications for the ratings were also examined for patterns. Data also were examined in the same way for patterns across teachers.

Student Achievement Measures: As part of the larger research effort in which this study was embedded, written assessment instruments were developed to assess student understanding of the curriculum content and science process skills (Krajcik, Marx, Blumenfeld, Soloway, & Fishman, 2000). The assessments were administered to each student participating in the curriculum projects. The assessments consisted of a combination of multiple choice and free response items that were further classified as either curriculum *content knowledge* or *science process* skill items. Content and process items were categorized by one of three cognitive levels required for arriving at a complete answer: *lower* (recalling information; understanding simple and complex information); *middle* (drawing or understanding simple relationships; applying knowledge to new or different situations; shifting between representations such as verbal to graphic; identifying hypotheses, procedures, results, or conclusions); and *higher* (describing or analyzing data from charts and graphs; framing hypotheses; drawing conclusions; defining or isolating variables given in a scenario; applying investigation skills; and using concepts to explain phenomena). The curriculum development teams (including science educators, content specialists, educational psychologists, and classroom teachers) constructed the tests. We analyzed all potential questions according to the scheme described above with teams of three to five raters achieving 95% accuracy in categorizing items. Disagreements were settled by consensus. The use of rubrics for each open-ended question produced over 95% agreement by two to four raters each. Again, disagreements were settled by consensus.

Testing rating criteria: The final step in this study was to develop scoring rubrics from the types of evidence and criteria identified from the justification statements for each rating category. To pilot test these rubrics additional videotaped enactments were examined. Two teachers enacting the force and motion unit in the fall of 2001 and two teachers enacting two other units (water quality and simple machines) were scored using the rubrics. Student achievement was measured with pre-posttests designed for each respective unit.

Findings

In the first analysis, categories and rating levels captured differences by teacher throughout all lesson sequences (Table 2). Ratings also indicated teachers were fairly consistent in their enactments. This finding was backed up by the descriptions of specific observations written in the justification statements. More importantly, this method of describing enactment made possible the identification of two groups of enactments. Two teachers' enactments tended to be a good match for the intended enactment whereas the other two teachers' enactments were less reflective of the intended enactment. Moreover, the distinction between the groups was evident not only in the ratings across analysis categories, but also in the specific aspects of enactments that led to the assigned ratings. In each case, the match of individual teacher's enactment to the respective group was quite reliable. These groups were also distinguished by students' achievement scores. Effect sizes were statistically significant on high and medium cognitive level questions for students in the first group and were not statistically significant on high cognitive level questions for the second group (Table 3). Interestingly, only the category of accuracy was not a unique indicator for either group or for student achievement. Teachers who presented science accurately were in both groups.

Table 2. Enactment ratings by teacher.

	Teachers in first analysis				Teachers in the second analysis			
Analysis Category	Franklin	Wells	Davis	Turner	Wells	Ross	Brooks	Day
Accuracy	3	3	3	2	3	3	3	3
	3	2	3	2	2	2	3	3
Completeness	4	3	1	1	4	2	3	3
	3	3	1	1	3	1	3	2
Opportunities	4	4	2	2	4	3	4	3
	4	4	2	2	4	2	3	2
Similarity	4	4	2	2	4	3	4	3
	4	4	2	1	4	2	3	2
Adaptation	3	2	2	1	3	2	3	2
	3	2	2	1	2	1	3	2
Instructional Supports	3	4	2	1	4	3	3	2
	3	4	2	1	4	2	2	1
Appropriateness	3	4	2	2	4	2	3	2
	3	3	2	1	3	2	2	1
Enactment Group	1	1	2	2	1	2	1	2

Note: Each category is represented by two ratings to represent enactments variation. When ratings did not vary, the category was assigned the same code each time.

Table 3. Student performance on pre- and post-tests for each teacher.

Force and Motion 8th grade	Pre-test M (SD)	Post-test M (SD)	Effect Size[a]	Pre-test M (SD)	Post-test M (SD)	Effect Size[a]
	Enactment Group One			**Enactment Group Two**		
Fall 1998	Ms Franklin (N = 29)			Ms Turner (N = 25)		
High level (18 points)	1.66 (1.08)	3.97 (2.23)	2.14***	0.88 (1.05)	0.88 (1.01)	0.00
Medium level (19 pts)	6.34 (1.45)	10.03 (2.23)	2.54***	5.04 (1.90)	6.00 (2.40)	0.51*
Low level (16 points)	8.03 (2.28)	9.59 (3.21)	0.68*	4.48 (2.22)	6.00 (2.87)	0.68*
Overall (53 points)	16.03 (3.45)	23.59 (6.16)	2.19***	10.40 (3.31)	12.88 (5.10)	0.75**
Fall 1999	Ms Wells (N = 56)			Mr. Davis (N = 25)		
High level (4 points)	0.63 (1.59)	1.25 (1.96)	1.06***	0.44 (0.65)	0.72 (1.02)	0.43
Medium level (9 points)	3.79 (1.39)	4.41 (1.69)	0.45*	3.60 (1.32)	3.72 (1.51)	0.09
Low level (8 points)	2.73 (1.27)	3.63 (1.36)	0.70***	2.40 (1.29)	4.40 (1.85)	1.55***
Overall (21 points)	7.14 (2.11)	9.29 (3.04)	1.01***	6.44 (1.87)	8.84 (3.16)	1.28***

[a]Effect Size: effect size was calculated by difference between the means divided by standard deviation of pre-test.
*$p < .05$. **$p < .01$. ***$p < .001$.

Creating rubrics: The identification of eight categories, rating levels, and types of evidence made it possible to construct scoring rubrics for each category (see Figure 1). Further, the specific examples used to justify the assigned ratings describe the characteristics of two to six types of evidence that are consistent with high or low ratings for each category. For example, the types of evidence for *appropriateness* of instructional supports that guided observations and ratings include: 1) questions and prompts, 2) hints and reminders, 3) student ideas, and 4) feedback. The criteria for judging the evidence is also important. For example, when questions are used to guide students to consider important content ideas this evidence contributes to a high rating. Conversely, when questions are used to elicit definitions this evidence contributes to low ratings in for instructional strategies. By rating each type of evidence an overall rating for the category is possible and justified.

***Appropriateness* of Instructional Supports**

Types of evidence	***Poor*** supports always not matched to student learning needs	***Insufficient*** supports usually not matched to student learning needs	***Sufficient*** supports usually matched to student learning needs	***Excellent*** supports always matched to student learning needs
Questions and prompts	*Answered or explained by the teacher* *Guide students to definitions*			*Guide students to focus on appropriate ideas*
Hints and reminders	*Address task completion*			*Address ideas with which students may have trouble*
Students ideas	*Not requested*			*Requested* *Connected to previously stated students' ideas*
Feedback	*Identifies mistakes or wrong answers*			*Directs students to appropriate ideas*
Student questions and difficulties	*Not addressed*			*Addressed*
Overall rating				

Figure 1. Sample rubric based on criteria identified in descriptive ratings.

When these rubrics were used to score additional videotape enactments, again the pattern of ratings was unique for each teacher (Table 2). It was necessary, however, to examine a variety of lesson types including teacher presentation of ideas, whole class and small group work, and investigations in order to arrive at valid ratings in each category. Also additional subcategories were necessary such as the process of modeling in addition to investigations. The types of evidence listed on the rubrics were the types of evidence seen in enactments and did facilitate scoring. Student achievement scores again were aligned with enactment ratings (Table 4). Accuracy, although not indicative of overall enactment group, may be related to student achievement on low cognitive level items.

Table 4. Student performance on pre- and post-tests for each teacher.

	Pre-test M (SD)	Post-test M (SD)	Effect Size[a]	Pre-test M (SD)	Post-test M (SD)	Effect Size[a]
	Enactment Group One			**Enactment Group Two**		
	Force and Motion, 8th grade					
Fall 2001	Wells (N = 49)			Ross (N = 43)		
High level (6 points)	0.98 (0.99)	2.12 (1.06))	1.15***	1.57 (0.93)	1.71 (0.96)	0.15
Medium level (9 points)	3.76 (1.22)	5.65 (1.34))	1.55***	4.48 (1.11)	6.00 (1.46)	1.37***
Low level (9 points)	3.50 (1.07)	3.65 (0.97)	0.14	3.64 (0.94)	3.98 (1.01)	0.36
Overall (24 points)	8.06 (1.99)	11.29 (3.08)	1.62***	9.72 (2.44)	11.86 (2.79)	0.88***
	Water Quality, 7th grade			Simple Machines, 6th grade		
Spring 2000/ Fall 1999	Brooks (N = 63)			Day (N = 87)		
High level (6/16 points)	0.32 (0.91)	1.25 (0.88)	1.02***	0.79 (1.10)	1.52 (1.41)	0.66*
Medium level (9/10 pts)	4.00 (1.48)	5.39 (1.65)	0.94***	4.21 (1.80)	5.14 (1.76)	0.52*
Low level (9/7 points)	4.32 (1.97)	6.29 (1.87)	1.00***	2.22 (1.15)	3.76 (1.31)	1.32***
Overall (24/33 points)	8.95 (2.67)	12.89 (3.73)	1.48***	9.84 (3.82)	14.50 (5.02)	1.22***

[a]Effect Size: effect size was calculated by difference between the means divided by standard deviation of pre-test.
*$p < .05$. **$p < .01$. ***$p < .001$.

Discussion

Measures used to research teaching in reform need to be reflective of the reform goals. The seven analysis categories are consistent with reform recommendations because they were developed from a reform-based curriculum framework. Moreover, these categories are not specific to the unit used to develop them. In this study, the types of evidence listed were applied to enactments of two other units. The categories should be adaptable to other reform-oriented science programs. Any quality program will be concerned with how content is presented, that students have opportunities to learn and that teachers give students guidance and support. In fact, consistent with reform ideas, many of the types of evidence and criteria identified in this study focus on respect for student ideas and an emphasis on accurate and important science ideas. The ratings were able to separate teachers enactments into two groups that are correspond to two groups indicated by student achievement scores. This suggests the categories and rating levels are capturing something important about teachers' practices that lead to student learning.

The link from specific aspects of classroom teaching to student learning is an important one. Whereas others have shown that teachers effect student learning they have not identified specific instructional practices that lead to improve student outcomes (Darling-Hammond, 1999). In addition, it is an important finding that measuring specific teacher behaviors is not sufficient to determine quality of teaching (Ball & Cohen, 1999; Lampert, 1998). It is not enough to know whether teachers are asking questions. It is also important to consider teachers' goals in asking these questions. If we can learn what teachers can do to help students learn we also can learn what types of support teachers need to learn and enact these practices.

The rubrics are based on enactment data and are formatted to facilitate scoring enactments directly. This should reduce the need to collect, prepare, and analyze videotape or detail descriptions of classroom events. However, these rubrics have only limited field-testing. Although the categories and types of evidence have proven to be useful and informative, other types of evidence may emerge from further observations of reform-based enactments. For example, examining a water quality unit highlighted the need to include modeling as well as investigation in all categories. Also categories and types of evidence directly related to what students are doing or saying may be necessary to more accurately predict student achievement. In addition, we do not know if it is possible to score enactments in real time in the class although it was possible to score videotaped enactments. Videotape facilitates pausing and rewinding, useful features when scoring multiple dimensions of complex events. Although much simpler than careful qualitative analysis, these rubrics remain complex. It is likely improvements can be made in these rubrics based on more extensive use in classrooms or with enactment videotape.

The process used to identify categories, rating levels, and specific types of evidence that could be used to characterize teaching was time and labor intensive. However, when these have been identified future evaluations will be much simpler. Further studies with more teachers enacting reforms would increase the reliability of these recommendations. Through this work, an observation framework that is appropriate for larger scale studies could be created. These categories should be presented in a format easily adapted to various classrooms and curriculum. This will make the much needed large-scale studies of teacher enactments feasible. We developed these rubrics to evaluate the efficacy of a curriculum centered reform effort but they can be adapted to use in other reform-oriented curriculum research questions. For example, Davis (2002) is using student teachers' unit plans to answer questions about how novices learn to teach. Others are using reform-based materials to promote student learning (Prawat & et al., 1992; Songer, Lee, & Kam, 2002). An evaluation scheme like the one presented here would be helpful to gauge how closely enactment reflects an intended curriculum plan without looking for strict implementation.

This importance of this work lies in its ability to provide a tool to facilitate research on teaching. One area of weakness is the lack of studies that bridge the gap between teacher preparation, classroom teaching, and student outcomes on a large scale. We know that teachers need to learn about teaching in the context of the classroom but we do not know how to efficiently support their learning (Putnam & Borko, 2000). Although we used this observation framework to inform the design of materials and support for teachers in reform, this framework will be valuable in many areas of research on teaching. A method to evaluate teaching that is meaningful and usable on a large scale is needed to inform teacher education and professional development research.

Endnote

(1) More information about this work including the curriculum materials used in this study, can be obtained from our project's web site at this address: http://hi-ce.org/teacherworkroom/middleschool/physics/index.html

References

Anderson, R. D., & Helms, J. V. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching, 38*(1), 3-16.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education, 51*(3), 241-247.

Ball, D. L., & Cohen, D. K. (1999). *Teacher learning and instructional capacity: Interaction and intervention*. Paper presented at the American Educational Research Association Annual Meeting, Montreal.

Blumenfeld, P. C., Fishman, B. J., Krajcik, J. S., Marx, R. W., & Soloway, E. (2000). Creating usable innovations in systemic reform. *Educational Psychologist, 35*(3), 149-164.

Blumenfeld, P. C., Krajcik, J. S., Marx, R. W., & Soloway, E. (1994). Lessons learned: A collaborative model for helping teachers learn project-based instruction. *Elementary School Journal, 94*, 539-551.

Cochran-Smith, M., & Fries, M. K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30*(8), 3-15.

Collopy, R. (1999). *Teachers use of and learning from curriculum materials*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence*: Center for the study of teaching and policy, University of Washington.

Davis, E. A. (2002). Scaffolding prospective elementary teachers in critiquing and refining instructional materials for science. In P. Bell, R. Stevens & T. Satwicz (Eds.), *Keeping Learning Complex: The Proceedings of the Fifth International Conference of the Learning Sciences (ICLS)* (pp. 71-78). Mahwah, NJ: Erlbaum.

Fishman, B., & Best, S. (2000). *Professional development in systemic reform: Using worksessions to foster change among teachers with diverse needs*. Paper presented at NARST, New Orleans, LA.

Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Inquiry based science supported by technology: Achievement among urban middle school students*. Paper presented at NARST, New Orleans.

Lampert, M. (1998). Studying teaching as a thinking practice. In J. Greeno & S. G. Goldman (Eds.), *Thinking practices* (pp. 53-78). Hillsdale, NJ: Lawrence Erlbaum.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks: Sage Publications.

National Commission on Teaching and America's Future. (1996). *What Matter's Most: Teaching for America's Future*. New York: National Commission of Teaching and America's Future.

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

Palincsar, A. S., Magnusson, S. J., Marano, N., Ford, D., & Brown, N. (1998). Designing a community of practice: Principles and practices of the GIsML community. *Teaching and Teacher Education, 14*(1), 5-19.

Prawat, R. S., & et al. (1992). Teaching mathematics for understanding. *Elementary School Journal, 93*(2), 145-152.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry, 29*(3), 315-342.

Ruiz-Primo, M. A., Shavelson, R., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform. *Journal of Research in Science Teaching, 39*(5), 369-393.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Schneider, R., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. *Journal of Science Teacher Education, 13*(2), 167-217.

Singer, J., Marx, R. W., Krajcik, J. S., & Clay-Chambers, J. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist, 35*(3), 164-178.

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

Wood, T., Cobb, P., & Yackel, E. (1991). Change in teaching mathematics: A case study. *American Educational Research Journal, 28*(3), 587-616.

Acknowledgement

This study was funded in part by the National Science Foundation as part of the Center for Learning Technologies in Education grant 0830 310 A605. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.