Title: A Digital Observation Instrument for Quantifying Educator Effectiveness

Authors: Jon Hasenbank & Jennifer Kosiak (UW-La Crosse)

Contact:

Jon Hasenbank (hasenban.jon@uwlax.edu, 608-785-6609)
Mathematics Department
1725 State St.
La Crosse, WI 54601

Abstract

We focus on the evolution of an innovative classroom observation instrument designed to evaluate teachers' instructional emphasis on eight dimensions of procedural understanding in mathematics. The instrument allows the observer to select among a number of categories as instruction unfolds, and the results are compiled to produce a quantitative measure of the time spent in each category. The instrument produces plots of the frequency distribution as well as the chronological-progression from category to category. The instrument will be presented in the context of the professional development projects for which it was developed, and the potential for adapting the instrument to other settings will also be discussed.

Introduction

One impact of the No Child Left Behind Act (2001) has been an increased emphasis on "scientifically based research" in the educational decision-making process; such an emphasis includes a focus on empirical methods of data collection with a reliance on valid and reliable measurements. The gold standard for scientific research is the randomized controlled experiment, but of course educational settings do not lend themselves well to that level of control. Quasi-experimental designs using random assignment to intact groups are more realistic, but then it is critically important to establish the initial equivalence of the groups and to assess differences in implementation between the treatment and the comparison groups. Therefore, one of the challenges facing educational researchers is quantifying instructional change. In this paper we describe the evolution of a classroom observation tool that provides both quantitative and qualitative data about the nature and sequencing of instruction. The evolution of the instrument will be described in the context of the two studies for which it was developed, but we believe it can be easily adapted to more general use as a flexible measure of educator effectiveness.

A pilot study was undertaken in 2006 by the first author and his colleagues at a western land-grant university. The goal was to assess the semester-long impact of an innovative teaching method on students' procedural understanding of college algebra (see Hasenbank, 2006). The study used a quasi-experimental, matched-pairs design with three treatment and comparison sections. The instructional treatment was based on an eight-item *Framework for Procedural Understanding* (Hasenbank, 2006), which treatment instructors were expected to use to deepen the content of their daily lectures[1].

Arguments in favor of teaching for understanding are based on the research from the fields of psychology and mathematics education, which suggest that knowledge that is understood is easier to remember and apply in novel situations (Chappell & Killpatrick, 2003; Star, 2000), and may even make it easier to learn new information (Hiebert & Carpenter, 1992). Carpenter and Lehrer (1999) list five ways in which classrooms can promote learning with understanding. Such classrooms provide students with opportunities to (a) develop appropriate relationships, (b) extend and apply their mathematical knowledge, (c) reflect about their own mathematical experiences, (d) articulate what they know, and (e) make mathematical knowledge their own. The *Framework for Procedural Understanding*, when used to guide instruction and assessment in algebra, captures these opportunities for understanding. (See Table 1).

The *Framework* was developed for use in *Navigating Through Algebra for Grades 9-12* (Burke, Erickson, Lott, & Obert, 2001). Adhering to the multi-modal nature of mathematical understanding, the original framework consisted of six guidelines for teaching and learning mathematics:

1. The student understands the overall goal of the algebraic process and knows how to predict or estimate the outcome.
2. The student understands how to carry out an algebraic process and knows alternative methods and *representations* of the process.
3. The student understands and can *communicate* to others why the process is effective and leads to valid results.
4. The student understands how to evaluate the results of an algebraic process by invoking *connections* with a context or with other mathematics the student knows.
5. The student understands and uses mathematical *reasoning* to assess the relative efficiency and accuracy of an algebraic process compared wit alternative methods that might have been used.
6. The student understands why an algebraic process empowers her or him as a mathematical *problem solver* (Burke, et al., 2001, p. 31, emphasis in original).

Before applying the Framework in a classroom setting, it was useful to re-express these guidelines by way of a series of student-centered questions. These student-centered questions are collectively referred to as the *Framework for Procedural Understanding*.

[1] These *Framework* items were also used to deepen the daily homework and weekly quizzes.

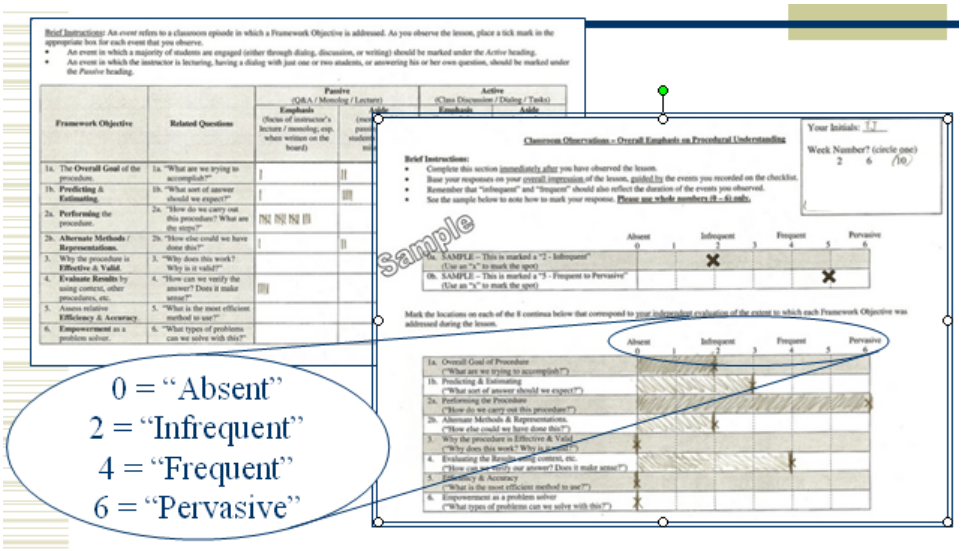
Table 1 - *Framework* Questions

1a. What is the goal of the procedure?
1b. What sort of answer should I expect?
2a. How do I execute the procedure?
2b. What are some other procedures I could use instead?
3. Why is the procedure effective and valid?
4. What connections or contextual features could I use to verify my answer?
5. When is this the "best" procedure to use?
6. What can I use this procedure to do?

The Pilot Instrument

In the pilot study, written assessments with task-specific scoring rubrics were used to provide quantitative measures of students' understanding and skill. To ensure that any gains could be attributed to the intervention itself, we set out to measure the fidelity of the implementation in each section; in part, this included measuring the instructor's emphasis on the *Framework* items in daily lecture. We developed an observation instrument that relied on a system of hand-written tally marks, whereby pairs of observers independently kept a tally of the individual classroom events in each of the *Framework* categories. At the end of the lesson, each observer then reviewed his or her tally and assigned a holistic rating in each of the eight categories. The ratings, which ranged from 0-Absent to 6-Pervasive, were intended to capture the overall emphasis the instructor placed on each of the eight *Framework* categories during the lesson. The holistic scores assigned for pairs of observers were averaged together in the final analysis. Figure 1 shows a sample of the pilot observation form.

Figure 1 - The Pilot Observation Instrument

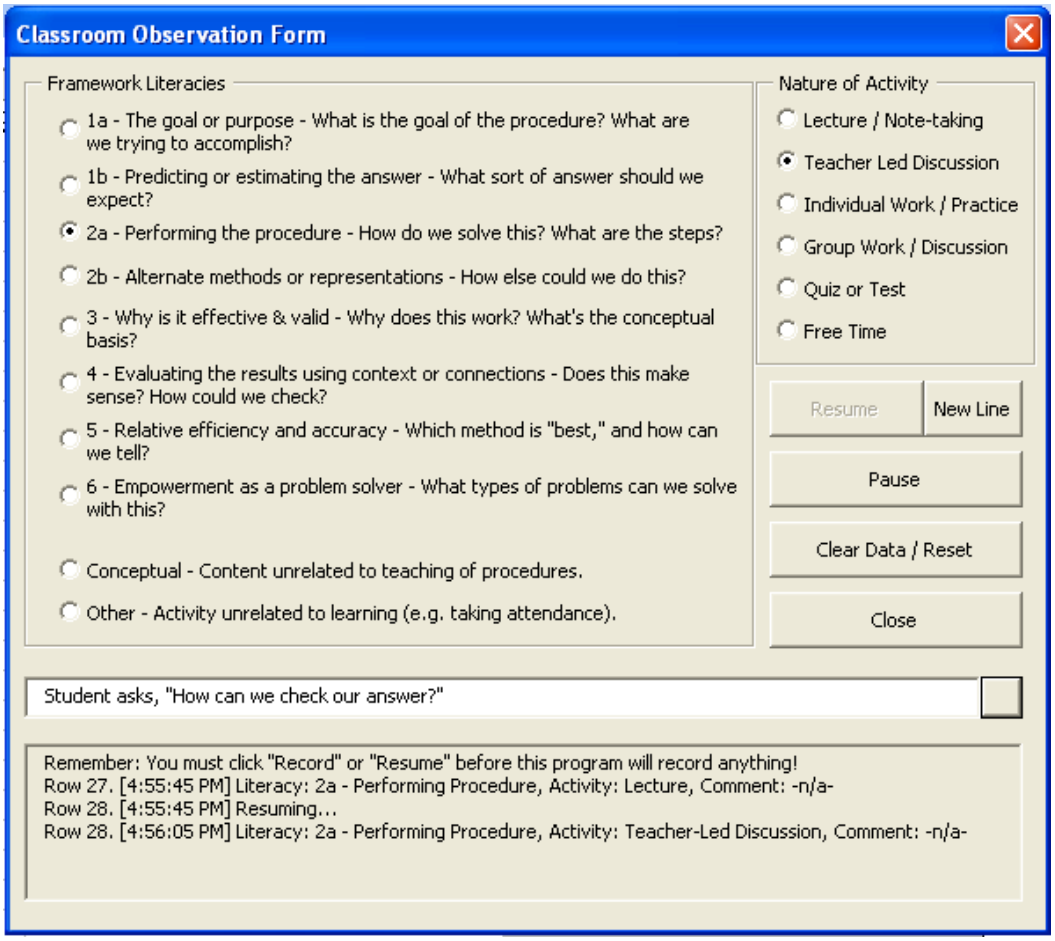


The primary limitation of the pilot instrument was that reliability of individual tally marks was impossible to assess; only the holistic ratings were examined for reliability. Thus, while overall reliability was sufficiently high, there was no way of assessing whether the observers had coded the classroom events consistently. A closely related limitation was that the pilot instrument did not take into consideration the chronological progression of the lesson. This permitted two lessons with widely varying structure to produce identical holistic results. For example, a lesson that sprinkled events related to, say, checking the accuracy of the solution is qualitatively different from one that allocates 10 focused minutes to the subject at the beginning of class, yet the pilot instrument would show the two lessons were similar. We concluded that what was needed was a method for characterizing the chronological progression of the lesson from category to category, much like the method that Schoenfeld (1992) has used in his research on students' problem solving strategies. A digital observation protocol that was chronological in nature would also help record the sequencing of classroom behaviors and precisely measure the time spent in each of the categories in order to systematically reveal patterns in practice.

Designing the Digital Observation Instrument

There are a number of written observation protocols in existence that capture information similar to the digital tool we set out to create, including the Praxis III: Classroom Performance Assessments (ETS, 1993) and the Instructional Skills Observation Instrument (ISOI) based on the Hunter Instructional Theory into Practice model (e.g. Hunter, 1976; Stallings, Robbins, Presbrey, & Scott, 1986). The Praxis III is a comprehensive model for analyzing the teaching of beginning teachers. It includes a written observation protocol that provides an observer space for noting the time, recording comments about the teacher-student interaction, and a selecting a category reflecting the nature of the interaction. Likewise, our digital observation instrument records a timestamp together with a primary and a secondary code for the nature of classroom activity and an optional observer comment. The primary and secondary codes can be easily changed to suit the needs of the observer. For the purposes of the follow-up study described below, we selected the eight dimensions of the *Framework* (plus "conceptual" and "other") as our primary codes. Our secondary categories paralleled the Hunter model of instructional design, which includes four sequenced elements of effective instruction – anticipatory set and focus, instruction, guided practice, and independent practice. The final set of secondary codes we selected for recording the nature of the classroom activity is shown in the screenshot of the user interface in Figure 2.

Figure 2 - Observation Form User Interface



Using the Instrument: Our Follow-up Study

Our follow-up study in 2007-08 focused on the professional development of middle and high school teachers in Wisconsin[2] for the purpose of developing teachers' abilities to emphasize *Framework*-oriented understanding in their teaching and, ultimately, to improve the depth of students' procedural knowledge of algebra. Nine teachers from six schools were active throughout the project, and an additional five teachers agreed to give the written assessments to their students for comparison purposes. Table 2 demonstrates the close synergy between the *Framework for Procedural Understanding* and the Wisconsin Model Academic Standards for Mathematics. We have also included classroom excerpts from our observations to illustrate how the *Framework* questions might be manifested in a lesson.

[2] The project was funded by the Wisconsin Improving Teacher Quality grant #07-0711, "A Professional Development Partnership for Improving Understanding in Algebra."

Table 2 - Alignment between the Framework and WMAS for Math

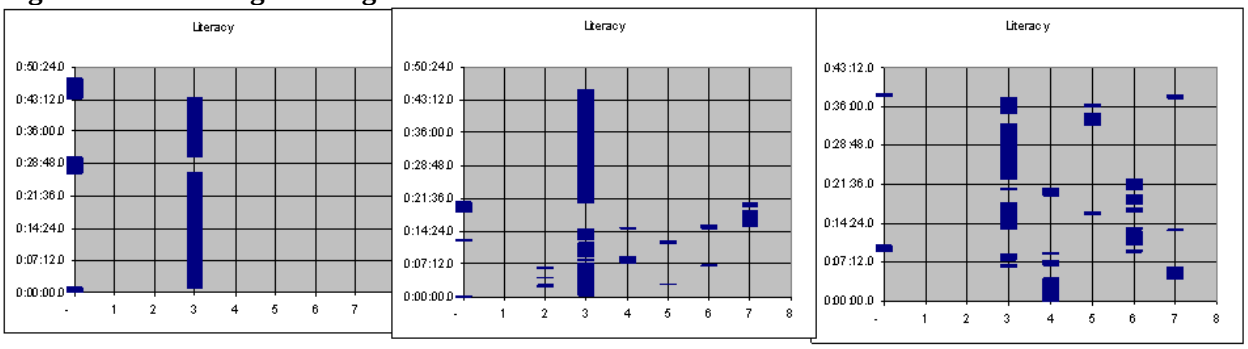
Framework Question	*WMAS Excerpt*	*Classroom Excerpt*
1a. What is the goal of the procedure?	"Students need to know…why [skills and knowledge] are being applied."	Student: "What are we trying to do?"
1b. What sort of answer should I expect?	"Students must be able to communicate their thinking to others."	Student: "What's the answer supposed to look like?" Teacher: "You won't necessarily come up with a number."
2a. How do I execute the procedure?	"Students need to know…how to apply skills and knowledge…"	Teacher: "What is the first step here?"
2b. What are some other procedures I could use instead?	"Learning is easier when students see the connections between various concepts and procedures…"	Teacher: "Graphing takes too long in this one; how can we also get the same information?"
3. Why is the procedure effective and valid?	"Students should be able to provide a reason… why [a] skill works the way it does."	Teacher: "Get into your groups. One person does the work, one person writes the reason."
4. What connections or contextual features could I use to verify my answer?	"Students should habitually check their results and conclusions for their reasonableness; that is, 'does this make sense?'"	Teacher: "Compare with what you did yesterday. Did you get the same answer? We're checking ourselves to see if what we're doing works."
5. When is this the "best" procedure to use?	"Students should be able to provide a reason why they have chosen to apply a particular skill or concept…"	Teacher: "When is it quicker to do intercept method vs. slope intercept form? Which works best for you?"
6. What can I use this procedure to do?	"Important goals for students are… to master specific knowledge necessary for its application to real problems…"	Teacher: "If you had to figure this by hand, you'd be better off simplifying the exponents before plugging in the values."

As in the pilot study, observers in the professional development program visited classrooms in pairs (when possible) to independently code each lesson's emphasis on each of the *Framework*-items and the nature of instructional activities. The observation instrument was installed on the observers' laptop computers, and the software recorded categories specified by the observer as the lesson progressed. The observers also typed brief qualitative descriptions of the classroom events, which the software recorded together with the category and timestamp. In this way, the instrument captured the researchers' perception of the lesson progression; later, we were able to compare observers' independent ratings of individual classroom events. A total of 22 lessons were observed. Of those, 14 were algebra lessons, half of which were observed simultaneously by both researchers.

The observation instrument automatically generates two visual displays of the lesson progression. The first is a simple bar graph showing the frequency distribution of the relative amount of time dedicated

to each of the prescribed primary (or secondary) categories. The second is a graph of the chronological progression of the lesson; Figure 3 illustrates this latter type. It shows the category to category (horizontal axis) progression over time (vertical axis) for three separate lessons as taught by three different teachers. Qualitatively, we can see clear differences between the lessons. From the perspective of our 2007-08 professional development program, we interpret the results of Figure 3 as evidence that the two teachers on the right have implemented much more balanced lessons (in terms of the *Framework* categories) than the teacher on the left.

Figure 3 - Chronological Progression of Three Different Lessons

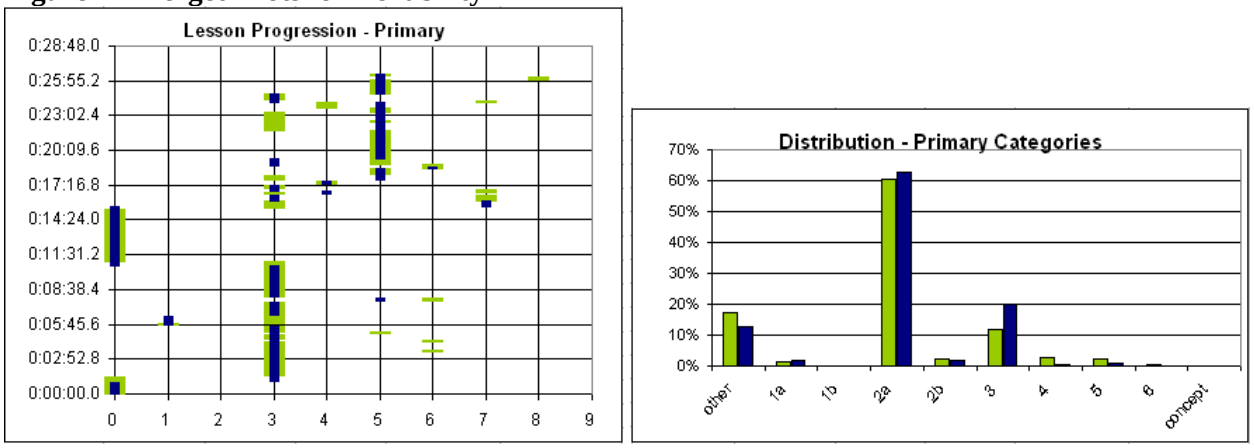


Instrument Reliability:

Inter-rater reliability was assessed by merging the data from the independent observations generated by the researchers. The method used to examine reliability involves triangulating a number of visual displays and numerical calculations. First, the software discretizes the data into fixed time intervals (typically, a resolution of 5-seconds is appropriate). The chronological progression plots are then merged and side-by-side frequency distributions are generated (

Figure 4). These plots give an initial visual estimate of the consistency between the two observations.

Figure 4 - Merged Plots for Reliability



Simultaneously, the software generates a table reflecting the pair-wise associations of the primary codes assigned by the two observers (

Figure 5). If the observers were perfectly consistent, all entries would appear on the main diagonal of the table. Summing the entries along the main diagonal gives the percentage of discretized intervals on which the researchers' codes agreed (72% in the example shown). Examining the off-diagonal elements allows us to identify the most common areas of disagreement (e.g., code 5 with code 3 occurred 7% of the time in our example), which can be cross-referenced with the chronological progression plot and the observers' comments for clarification.

Figure 5 - Code-Pairs

	-1	0	1	2	3	4	5	6	7	8	9	Row:
-1	-	-	-	-	-	-	-	-	-	-	-	0%
0	-	20%	-	-	2%	-	-	-	-	-	-	22%
1	-	-	1%	-	3%	-	-	-	-	-	-	3%
2	-	-	-	-	-	-	-	-	-	-	-	0%
3	-	2%	-	-	31%	-	3%	2%	3%	-	-	41%
4	-	-	-	-	1%	1%	-	-	1%	-	-	3%
5	-	-	-	-	7%	2%	18%	1%	-	1%	-	29%
6	-	-	-	-	-	-	-	1%	-	-	-	1%
7	-	-	-	-	2%	-	-	-	-	-	-	2%
8	-	-	-	-	-	-	-	-	-	-	-	0%
9	-	-	-	-	-	-	-	-	-	-	-	0%
Col:	0%	22%	1%	0%	45%	3%	20%	4%	4%	1%	0%	

Finally, we use the code-pairs table to compute Cohen's *kappa*, defined below, which represents the excess of the observer agreement over that expected purely by chance (Agresti, 1990).

$$kappa = \frac{(observer\ agreement) - (chance\ agreement)}{1 - (chance\ agreement)}$$

Benchmarks for interpreting *kappa* are presented in Table (Landis & Koch, 1977). The presence of several *kappa* values in the *fair agreement* range (0.2-0.4) reflect observations for which agreement rates that were not ideal. Our investigations revealed that one cause for the disagreement between codes is a lack of strict independence between the primary categories we selected for our observation. In particular, there were instances during our observations where we felt that it would be appropriate to select multiple codes simultaneously. For instance, a discussion about checking the results of a calculation (*Framework item 4)* by seeking alternate representations of the problem (*Item 2b*) might be difficult to code consistently. These instances reflect the complexity of classroom discourse, and they highlight the importance of using a set of carefully crafted and well-defined non-overlapping categories for the observations.

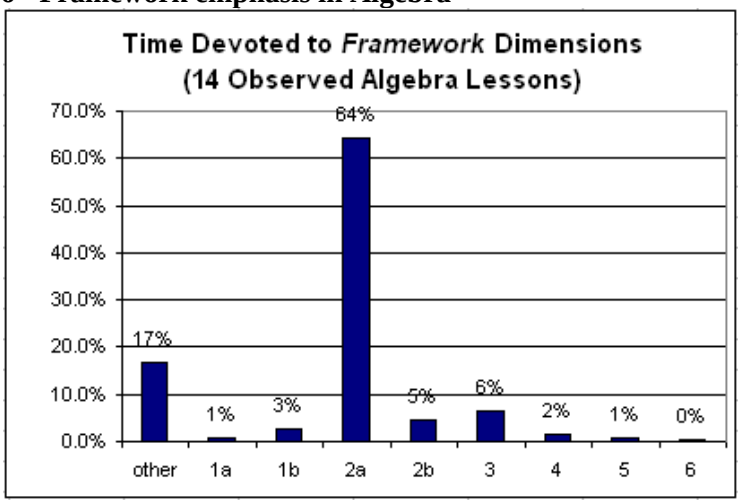
Table 3 - Agreement rate (kappa) for Classroom Observations

Range (interpretation)	*List of observed* kappa
<0.0 (worse than chance)	-
0.0-0.2 (slight agreement)	-
0.2-0.4 (fair)	.22, .24
0.4-0.6 (moderate)	.42, .45, .48, .48
0.6-0.8 (substantial)	.60, .61, .70, .74, .75, .79
0.8-1.0 (almost perfect)	-

Results – The Nature of Algebra Instruction:

Our observations of the teachers participating in our professional development program can be viewed as a snapshot into the nature of algebra instruction in Wisconsin. These teachers were trained to cultivate classroom norms that promote understanding using the *Framework* as a guide. Therefore, our classroom observations demonstrate what can be interpreted as an upper bound on the emphasis that Wisconsin teachers place on understanding in their algebra classes. Individual analysis of teacher efforts will be conducted as part of our program evaluation; here, we present a cumulative frequency distribution of the relative emphasis these teachers placed on understanding along the dimensions of the *Framework* (see Figure 6). We see that these teachers devoted 64% of class time to Item 2a – Performing the Procedure. This is skills-based instruction, and is an important component of teaching about algebra. Summing across all other categories, we see that 18% of class time was spent on other categories related to teaching for understanding. By triangulating these results with the results of our student assessments, we hope to gain a clearer picture of the impact that this emphasis had on students' knowledge and skill.

Figure 6 - Framework emphasis in Algebra



Note: The "other" category includes all activities that did not fit under the umbrella of "teaching about procedures." This may include some unstructured time but may also include time spent teaching about concepts as well as time spent on routine classroom activities such as taking attendance and handing back papers. We warn that "other" is not synonymous with "unstructured."

Implications

There are significant implications for other professional development educators and classroom researchers. The digital classroom observation instrument is useful for evaluating a wide variety of professional development programs aimed at changing teacher practice. By adjusting the categories displayed on the user-interface form, the instrument can be adapted to these and other purposes. For example, the digital instrument can be modified to code classroom behavior categories consistent with the Praxis III: Classroom Performance Assessment or the four instructional skills components in the ISOI. Another professional development program aimed at incorporating formative assessments into teachers daily lessons might benefit from a version of the instrument that includes categories such as "Think-pair-share," "Individual reflection," "Brief interview," "Class board work," etc. We are eager to share this instrument with others in the hopes that it might assist them with program evaluations throughout the region.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Burke, M., Erickson, D., Lott, J. W., & Obert, M. (2001). *Navigating Through Algebra in Grades 9-12*. Reston, VA: National Council of Teachers of Mathematics.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and Learning Mathematics with Understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19-32). Mahwah, N.J.: Lawrence Erlbaum Associates.

Chappell, K. K., & Killpatrick, K. (2003). Effects of Concept-Based Instruction on Students' Conceptual Understanding and Procedural Knowledge of Calculus. *Primus, 13*(1), 17-37.

Educational Testing Service (1993). Formative Studies of Praxis III: Classroom Performance Assessment. Princeton NJ: Author.

Hasenbank, J. F. (2006). *The Effects Of A Framework For Procedural Understanding On College Algebra Students' Procedural Skill And Understanding.* Unpublished Doctoral Dissertation, Montana State University, Bozeman.

Hiebert, J., & Carpenter, T. P. (1992). Learning and Teaching with Understanding. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 65-97). New York: Macmillan.

Hunter, M. (1976). Teacher competency: Problem, theory, and practice. *Theory into Practice* 15(2), 162-171.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

No Child Left Behind Act. (2001).

Schoenfeld, A. H. (1992). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 334-370): Macmillan.

Stallings, J., Robbins, P., Presbry, L., & Scott, J. (1986). Effects of instruction based on the Madeline Hunter Model on students' achievement: Findings from a follow-through project. *The Elementary School Journal 86*(5), 571-587.

Star, J. R. (2000). *On the Relationship Between Knowing and Doing in Procedural Learning.* Paper presented at the Fourth International Conference of the Learning Sciences.