The Development and Evaluation of a Program for Improving and Assessing the Teaching of Mathematics and Statistics

Dave Brown, Brynja Kohler, & James Cangelosi
Department of Mathematics & Statistics
Utah State University

Introduction and Goals of the Program

On August 23, 2007, the 35-member Department of Mathematics and Statistics faculty at Utah State University voted to undertake a project for assessing and improving how mathematics and statistics courses are taught. Since then, a committee consisting of seven faculty members (a graph theorist, two mathematics educators, a mathematical biologist, a numerical analyst, and two statisticians) has continually worked on the problem of addressing the following goals of the *Assessment and Improvement of Teaching Program* (*AITP*):

A. Develop, field test, refine, and implement a mechanism that provides faculty members with valid formative feedback data that will stimulate them to (i) reflect on specifics of their own teaching practices, (ii) continue successful practices, (iii) experiment with promising innovative practices, and (iv) make needed adjustments in their teaching.

B. Develop, field test, refine, and implement a mechanism that provides summative evaluations of faculty members' teaching practices – a program that will provide university academic administrators with a much more tenable assessments of its faculty's teaching practices than what is now being used university-wide and what is now being used in the vast majorities of post-secondary schools.

C. Assess both the formative and summative mechanisms with respect to the fundamental psychometric principles of measurement *validity* and *usability*.

D. Design and implement this program so that it is congruent with legal standards for "evaluation of personnel" which have been established through a string of litigations occurring over the past 25 years.

Conceptual Framework

The principles upon which this program is now built and continues to be refined are drawn from the research-based literature from four fields of study: (i) *Psychometric* principles for developing and assessing the validity of data-gathering devices, (ii) *qualitative research design* principles for conducting comparative case studies, (iii) *instructional supervision* principles for building collaborative formative feedback teams, and (iv) *mathematics education* principles for focusing both formative feedback and summative evaluation on critical instructional variables. Because of the potential impact of this program on faculty tenure, promotion, and merit compensation issues, established legal precedents influenced program design (e.g., precautions are used so that data gathered from the formative feedback mechanism cannot influence reports resulting from the summative evaluation mechanism).

The major guiding principles gleaned from this literature include the following:

- To help create a non-threatening climate in which faculty members are comfortable making suggestions, sharing ideas, discussing concerns, and critiquing one another's work with regard to their teaching, every aspect of *AITP* focuses only on *instructional practice* rather than the teachers themselves. For example, instead of saying, "Claudia is an engaging instructor," *AITP* participants are trained to make more pointed descriptive comments about what Claudia did rather than who she is (e.g., "While trying to lead her students to discover what the second derivative of a function indicates about the contour of its graph, Claudia paralleled symbolic, numerical, and graphical representations while continually posing questions to her students. This practice helped maintain a high level of student engagement throughout that aspect of the lesson."). Such attention to wording subtleties in *AITP* documents and interchanges among participants is a critical factor in creating and maintaining the collaborative environment necessary to the program's success.

- Evaluations of the quality of one's teaching need to be data based. But quality of instruction is a function of a chaotic mix of very complicated variables. So for this project, "data" is interpreted broadly to include not only strings of numbers but also records of empirical observations that are not easily quantified. But even for seemingly informally-gathered data, an argument needs to be made that those data are the results of *valid measurements*. For purposes of clarity, consider the following definitions of psychometric terms used in this paper:

    - *Measurement*: A measurement is a process by which data or information are gathered via empirical observations and recorded or remembered.
    - *Measurement Relevance*: The relevance of a measurement used in an evaluation of instructional practice depends on how well the data or information it generates pertains to the question addressed by that evaluation.
    - *Measurement Reliability*: A measurement is reliable to the degree that it can be depended upon to yield to yield consistent, non-contradictory data or information.
    - *Measurement Validity*: A measurement is valid to the degree that it is both relevant and reliable.
    - *Measurement Usability*: A measurement is usable to the degree that it is inexpensive, brief, easy to administer and score, and does not interfere with other activities.
    - *Evaluation*: An evaluation is a judgment about the quality, value, effectiveness, or impact of something (e.g., teaching).
    - *Formative Evaluation of Teaching*: An evaluation of teaching is formative if its *sole* purpose is to provide information that is useful for decisions about how to teach.
    - *Summative Evaluation of Teaching* : An evaluation of teaching is summative if it is a

judgment of instructional effectiveness that is used for purposes other than helping teachers decide how to teach. Unlike formative evaluations, summative evaluations may influence administrative decisions about that teacher's retention, salary, or promotion.

- *Formative Evaluation of Student Learning*: An evaluation of student learning is formative if its *sole* purpose is to guide students' learning activities.
- *Summative Evaluation of Student Learning* : An evaluation of student learning is summative if it is used to report how well students achieved learning goals.

- A clear distinction needs to be made between *measurements* and *evaluations*. Unfortunately, many of the so-called "measurement" or "observation" forms commonly used in evaluations of instruction include prompts for "observers" (e.g., department heads) to respond with their judgments rather than their observations. One such mislabeled form uses a Likert scale in which the "observer" is prompted to record a level of agreement with the statement, "The teacher demonstrates knowledge of mathematics." *AITP*'s guiding principles do not suggest that judgments about such qualitative characteristics shouldn't be part of the evaluative process. But they should not be considered empirically observable; they should be recognized as opinions informed by measurement results. Consider the following example drawn from an *AITP* experience:

Example 1:

One of the aspects of *AITP* is for each teacher who volunteers to participate in the program to have a *Summative Evaluation Team* (*SET*) consisting of her/himself and two evaluators who are also Mathematics and Statistics Department faculty members. Kathryn and John serve on Belinda's *SET*. After observing a string of Belinda's classes for a number theory course, Kathryn writes the following to be incorporated in a *SET* report that the team will eventually submit to the Department Head:

> "In my judgment, Belinda displayed careful attention to precise mathematical language. This judgment is based on multiple in-class observations as well as examination of study notes Belinda posted online for her students. Here is an example of one of several observations I made that is evidence of how she models attention to precise mathematical language:
>
> While examining relationships involving perfect numbers, Belinda made reference to the following definition she had formulated earlier in the class meeting:
>
> $p \in$ {perfect numbers} $\Leftrightarrow p = \sum_{i}^{k} d_i$ where ($|\{d_1, d_2, d_3, ..., d_k\}| = k$ and $\{d_1, d_2, d_3, ..., d_k\}$ = {proper divisors of p})
>
> Michonn (a student), while referring to '$|\{d_1, d_2, d_3, ..., d_k\}| = k$ ', asked, 'What's the point of that? I don't understand what that means.' Belinda replied with a question, 'How many elements are there in this set?' as she writes on the board,

‘$\{a_1, a_2, a_3, ..., a_n\}$.’ Michonn and others say, ‘n.’ Belinda asks for a show of hands to indicate agreement with ‘n’ as the answer. About 20 of the 30 students raise their hands. Belinda says, “Matt didn’t raise his hand. Why not, Matt?’ ‘I don’t know; I don’t understand,’ is the reply. Belinda replies, “I’m with Matt on this one. I don’t know either. I agree with Michonn that there could be as many as n numbers there but – okay, Vanessa, you look like you have something to say.” Vanessa says, ‘It just dawned on me that we didn’t specify anywhere that those numbers had to be distinct. Maybe $d_5 = d_{17}$ which would mean that $|\{d_1, d_2, d_3, ..., d_k\}| < k$. So I see now why you put that part in the definition.’ The conversation continued along these lines for another three minutes.”

In this example, Kathryn judges that Belinda displayed precise use of mathematical language and identifies an empirical basis for that judgment. She distinguishes her evaluations from her measurements. Of course it isn’t practical for Kathryn to write down every observation that influenced her many judgments about Belinda’s teaching. What she did do was to support each evaluation with examples of supporting measurement results. Kathryn chose the particular classroom episode because it also provides an example of observations she made that support other evaluations she made of Belinda’s teaching (i.e., “Belinda used formative feedback from students to make adjustments in her in-class activities with students.” And, “Belinda’s style of interacting with students invites them into conversations that deepen understanding.”).

- Teaching is an extremely complex art involving inextricably-interrelated variables. The improvement-of-teaching aspects of *AITP* require an advanced organizer for analyzing some proper subset of those variables. Furthermore, the summative evaluation aspects require *relevant measurements* as previously defined. Measurements can hardly pertain to the questions addressed by the evaluation unless those questions have been specified. And such questions are specified by enumerating well-defined variables affecting teaching practice. From its examination of the research-based literature in mathematics education, the *AITP* Development Committee generated the following advanced organizer for categorizing teaching variables:

A. *Appropriateness of learning objectives* (Are the objectives consistent with curriculum guidelines (e.g., course description, the published syllabus, what students are conventionally expected to gain from the course, and research-based learning principles (e.g., conceptual learning w/r mathematical content should often precede skill-building w/r that content))? Are the objectives sufficiently, but not overly, ambitious?)

B. *Appropriateness of lesson designs* (Considering the (i) aptitudes and prior achievements of the students, (ii) the time, resources, and technologies available to the teacher, and (iii) the targeted learning objectives, what pedagogical principles and teaching strategies are applicable (e.g., direct instruction is appropriate for leading students to develop algorithmic skill, whereas, inquiry instruction is appropriate for leading them to discover relationships)? Are such principles and strategies incorporated in the design of lessons?)

C. *Appropriateness of the conduct of the lessons* (Are the learning activities actually conducted in accordance with the pedagogical principles upon which the lessons were designed?)

D. *Validity and usability of measurements that impact summative evaluation of student achievement of learning objectives* (To what degree is the evaluation of student achievement based on measurements that are valid and usable?)

E. *Appropriateness of teaching activities w/r side-effects* (*i.e., effects not directly associated with specific learning objectives*) (Does the manner in which the teacher interacts with students and conducts lessons provide students with healthy, positive experiences that stimulate their desire to learn mathematics or statistics and lead them to feel that they are contributing members of a learning community?)

- Formative feedback data gathered during the improvement-of-teaching aspects of *AITP* should be blocked from influencing evaluations made as part of the summative evaluation aspects. "*Data curtain*" is a legal expression commonly used in litigations involving faculty-related job actions. The literature related to instructional supervision is replete with arguments for maintaining a *data curtain* that hides formative data from those making summative evaluations. Instructional supervisors are responsible for helping teachers be more effective with their students. They can hardly meet this responsibility without making accurate formative evaluations of teachers' instructional practices. However, if teachers suspect that instructional supervisors' formative evaluations may influence administrative decisions regarding their retention, salaries, or promotions, the trusting, collegial relationships necessary for effective instructional supervision is threatened. Consequently, formative evaluations of instruction can hardly serve their purpose unless they are completely divorced from summative evaluations of instruction. Although this data-curtain principle has been consistently upheld by a string of judicial proceedings since the mid 1980s, it is almost universally violated by school, college, and university administrators. As reflected in the Program Description section of this paper, *AITP* 's design is consistent with the data-curtain principle.

- Teaching is virtually always improved when small collaborative teams of teachers provide one another with formative feedback on very focused aspects of their instructional practices. Focused in-depth analyses of instructional practices emphasizing descriptions of what occurs during a limited time frame (e.g., level of student engagement in planned learning activities during a unit on solving optimization problems in a beginning calculus course) is far more effective than the more common practice of drop-in observations by supervisors who rate the teacher on poorly-focused traits (e.g., "knowledge of content" or "friendliness").

- For *AITP* to take root and make a positive impact on the teaching across the Department, it should begin as an experiment with a few small teams of volunteers who report on their experiences in a very descriptive, non-judgmental fashion to the faculty as a whole. No faculty member should be required to participate. Development of the program is never complete; it is forever considered a work in progress.

Program Description

AITP as it is currently constituted is managed by an oversight committee. Anyone who teaches for the Department and would like to take advantage of AITP's services meets with the Oversight Committee for the purpose of establishing two teams, an *Instructional Improvement Team* (*IIT*) and a *Summative Evaluation Team* (*SET*). The *IIT* consists of the teacher her/himself and two faculty members; the *SET* also includes the teacher as well as two faculty members who are not also members of the *IIT*. The following example is presented to clarify how *IIT*, *SET* pairs operate:

Example 2:

Laura requests *AITP* services because she wants to improve her instructional practices and build a case for *excellence in teaching* for her promotion and tenure file. She recruits Jessica and Jorge for the *IIT*. In consultation with Laura, the Oversight Committee selects Belinda and Les from its pool of faculty volunteers to join Laura on the *SET*. An Oversight Committee member reviews *AITP* procedures with the *IIT*, addressing such issues as how to secure the data curtain. Laura is responsible for setting *IIT* 's agenda (e.g., by identifying specific aspects of her teaching on which to focus). One role of *IIT* is to help Laura prepare for the subsequent evaluation of her teaching by *SET*.

During a 90 minute meeting, Laura's *IIT* makes the following decisions:

1. The effort is to focus on a three-week unit of a Calculus I course in which Laura attempts to lead students to accomplish the following:

   A. Comprehend plausibility arguments for the validity of theorems underlying basic algorithms for computing derivatives of polynomial, exponential, and trigonometric functions (Reference Sections 3.1 (Derivatives of Polynomial and Exponential Functions), 3.2 (The Product and Quotient Rules), 3.4 (Derivatives of Trigonometric Functions), 3.5 (The Chain Rule), 3.6 (Implicit Differentiation), and 3.7 (Derivatives of Logarithmic Functions) of J. Stewart's *Calculus: Concepts and Contexts* (*3rd ed.*, pp. 182–247)).

   B. Develop computational fluency with the basic algorithms listed in the Sections 3.1–3.7 of the Stewart text .

C. Solve a wide variety of word problems, each requiring one to decide which combinations (if any) of the algorithms listed in the Sections 3.1–3.7 of the Stewart text should be employed.

2. During a class period in which Laura finishes up work from the prior unit, Jessica and Jorge would make an ecological observation (i.e., they would observe for the purpose of soaking up the culture of the classroom, beginning to understand Laura's modus operandi without focusing on any particular aspects of her teaching).

3. Shortly after the ecological observation, the team will meet to clarify Laura's plans for the upcoming unit and to determine which aspects of Laura's teaching will be the focus of the effort.

4. Jessica and Jorge will make focused observations in as many class sessions as they reasonably can. After each class period, the team meets to discuss what occurred during the session and to plan for subsequent sessions.

5. Jessica and Jorge will work with Laura to design the test to be administered near the end of the unit; later they will discuss students' responses to the test prompts.

The five stages of the plan are implemented. The initial ecological observation facilitate Laura's explanations of the rationale for the objectives, her unique teaching style, plans for leading students to achieve the objectives, plans for dealing with some of the special challenges associated with this particular group of students, plans for monitoring students' progress, and plans for making a summative evaluation of how well objectives are achieved near the end of the unit. The team decides to focus on six instructional variables:

1. Engagement levels of students – particularly for six of the 32 students whose body language suggest boredom or hardly ever volunteer to speak in class

2. Quality of formative feedback Laura picks up from students during class so that she better monitors what they understand and what they misunderstand

3. The clarity of her explanations and directions for students

4. Choice of illustrative examples and problems used during class and on assignments

5. Use of in-class illustrations, especially computer-based graphics

6. Validity of the unit test

To provide a flavor for how the *IIT*'s efforts played out for these six variables, here is a small sample of the suggestions Laura decided to use and how they worked for her:

1. *Instead of being confined to the front of the room while delivering explanations, frequently move among students observing what they are writing, reading their body language, and using proximity to cue them to be on-task.* It took awhile for Laura to feel comfortable speaking on the move away from her comfort zone near the board. But she discovered that students' attention was productively cued by her movements between (i) the front of the room where she writes on the board or displays a computer-generated graphic and (ii) locations among the students. For example, the students were soon conditioned to take notes while she wrote on the board and then to discuss or attend to her explanation as everyone (including Laura) focused on what they had all written. Laura was amazed at how well she was able to pick up on students' perceptions, conceptions, and misconceptions of the mathematical topics by being among them as she taught. The focus of attention shifted from Laura herself to the mathematical topics at hand.

2. *To collect formative feedback on what students are understanding and misunderstanding, use pointed questions directed to individual students rather than vague lip-service questions such as, "Does everyone understand?"* For example, Laura writes the following on the board, "We can use the fact that $\frac{d}{dx}e^{x} = e^{x}$ to evaluate $\frac{d}{dt}e^{\sqrt{t}}$. In this case, we have *x* is $\sqrt{t}$, thus, $\frac{d}{dt}e^{\sqrt{t}} = e^{\sqrt{t}}$. " She moves away from the board noticing some students frowning, mumbling to themselves, or checking with their calculators while most appear content with what she wrote. Laura: "Doesn't *e* to the *x* make our lives easy, Tom?" Tom: "Yea, I like these kind of problems." Henri: "But that's not what I got on my calculator." The discussion continues with Laura leading students to comprehend the distinction between $\frac{dy}{dt}$ and $\frac{dy}{d\sqrt{t}}$; students conclude they should have applied the chain rule. Such episodes provide Laura with rich feedback about students' progress.

3. *No change is needed in the way explanation and directions are delivered.* However, Laura noticed an improvement in students' comprehension of her explanations and directions after she implemented the first, second, and fifth suggestions from this list.

4. *Instead of illustrating techniques with five or six relatively simple problems, work through only a couple of relatively-complicated problems that have more of a real-life flavor to them. In-class problems should typically be harder to solve than those presented on tests*. Laura did not detect that this practice improved students' test scores. However, she did notice two benefits: (i) Before diving into a problem, she and her students developed the habit of stepping back for the purpose of comprehending what the problem was all about. (ii) Students complained less about the difficulty of tests than they had before.

5. *Depend less on the white board and more on the tablet laptop for in-class illustrations.* By taking this suggestion, Laura could display clear graphics and move about the room more efficiently than before. She is now able to store what she otherwise would have written on the board for subsequent use and to send electronic copies to students.

6. *Prior to constructing the unit test, decide on the relative importance of various topics* (e.g., 20% of the test prompts will involve implicit differentiation) and skills (33% of the test points will reflect algorithmic skills). This practice led Laura to be more systematic in how she designed not only tests but also curricula. Thinking more about test design stimulated her to read about the art and science of assessing student achievement. She now routinely computes reliability and item efficiency indices on test scores.

Laura's attention shifts to work with her *SET*. Two templates are used for *SET* reports, one for the teacher her/himself that Laura uses to teach Belinda and Les about her plans for the unit that will be the basis for their evaluation of her teaching. In Section 1, "General Information," Laura inputs information about the course and unit that will be the focus of the evaluation (e.g., course syllabus, title of the unit, and schedule of class meetings that pertain to the unit). The prompts for Section 2, "Instructional Unit Description," are as follows:

2.1. Describe your activities related to planning this unit.

2.2 Describe your purpose in conducting this unit by listing a string of learning objectives for the students to achieve during the unit. Also explain your rationale for targeting these objectives.

2.3 Describe your plan of instructional activities for leading students to achieve the objectives. Also provide your rationale for the planned instructional activities.

2.4. Describe your plan for making formative evaluations of students' progress as they engage in learning activities, and your plan for making summative evaluations of students's achievement of the learning objectives.

2.5. Describe any relevant concerns, motives, thoughts, and plans regarding side effects resulting from students' activities during this unit, or any special considerations important for evaluating this unit.

The document expands to four pages with Laura's responses which she explains to Belinda and Les in a two-hour meeting. The *SET* focuses on a unit involving specific applications of differentiation to solve related-rates problems and describe the behaviors of various functions. Belinda and Les observe every class session during the two-weeks of that unit and the team meets after each observation, not to communicate opinions or provide feedback about the class session (as an *IIT* would do) but to clarify what happened. For example, Les asks, "Why did you call on Vessela instead of one of the students with their hands up?" Laura explains her rationale but neither Les nor Belinda indicate whether or not they agree with the decision. After the unit Les and Belinda respond to the following prompts of Section 2, "Evaluator's Summative Judgment of Teacher Performance Relative to the Unit," of their template:

> For each category below, provide your summative judgments regarding the teaching performance relative to the unit, as well as your record of empirical bases or research literature bases (i.e., evidence) for your summative judgments.
>
> 2.1. Appropriateness of learning objectives
>
> 2.2. Appropriateness for the teacher's plan for leading students to achieve the objectives
>
> 2.3. How well the teacher executed the plan for leading students to achieve the objectives
>
> 2.4. Appropriateness of the teacher's plan for making formative and summative evaluations of students' progress and achievement of objectives
>
> 2.5. Side effects on students resulting form their experiences with the teacher during the time period of the unit

The evaluators are prompted to write summarizing comments in a third section.

Laura reviews both Belinda's and Les' reports which are finalized at their culminating meeting. The two *SET* reports are forwarded to the Department Head and incorporated in Laura's tenure and promotion file.

<u>*AITP*'s Experiences to Date</u>

To date the process has been employed by five teachers with 10 people serving as *IIT* or *SET* members. All of them indicate that the process has had a far more positive effect on not only the teaching that was analyzed but also on the teaching of those who provided suggestions as *IIT* members and those who wrote summative evaluation reports in their roles on *SET*s. Belinda's comment in the journal she keeps as part of the qualitative research component of *AITP* is indicative of the expressions of most participants: "I was amazed at what I learned just from my first ecological observation in John's linear regression course. Not only did I gain understanding about time series analysis and pick up some pedagogical ideas to use in my own teaching, but I also recognized nine students in John's class who are also in my number theory class. Without even planning to do so, I found myself associating number theory content to times series analysis

content from John's class. My students really appreciated the connections."
All of the *AITP* participants have chosen to extend their involvement as both teachers and evaluators. Faculty members who have not yet participated express support for the program but most worry that it will be too time consuming, detracting from time to do research. However, because *IIT*s and *SET*s operate for only a few weeks at a time, because the impact on the teaching of all involved, and because the *SET* reports provide data based, meaningful evaluation of teaching that may help make the case for promotion or tenure, those who have *AITP* experience believe the program is a very efficient use of faculty time.

References

American Psychological Association. (1999). *Standards for educational and psychological testing*, 6th ed. Washington, Author.

Brock, A., Cangelosi, J., Carlson, T., Hansen, M., Hunt, A., Manning, T., Meldurm, J. Merril, A., Moody, S., & Walker, R. (2006). *Measurement analysis tool: User's guide for the measurement analysis coding form*. The Professional Development Assessment Project. Logan: Utah State University.

Cangelosi, J. S. (1991). *Evaluating classroom instruction*. New York: Longman.

Cangelosi, J. S. (2000). *Assessment strategies for monitoring student learning*. New York: Addison-Wesley.

Cangelosi, J. S. (2003). *Teaching mathematics in secondary and middle school: An interactive approach*. (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Hansen, M. Manning, T., Merrill, A., Brock, A., Moody, S., Hart, L., Carlson, T., Hunt, A, & Meldrum, J. (2005). *Classroom dynamics: User's guide for the video-coding observation form*; Professional Development Assessment Project. Logan: Utah State University.

Joint Commission on Standards for Evaluating Educators. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. Newbury Park: Sage. Joint Committee on Testing Practices. (2004). *Code of Fair Testing Practices in Education*. Washington, Author.

Puchner, L., & Taylor, A. (2006). Lesson study, collaboration, and teacher efficacy: Stories from two school-based math lesson study groups. *Teaching and Teacher Education: An International Journal of Research Studies*, 22(7), 922–934.

Sedlin, P. (1999). *Changing Practices in Evaluating Teaching.* Bolton MA: Anker.