Essays on Teaching Excellence

Toward the Best in the Academy

Volume 16, Number 5, 2004-05

A publication of The Professional & Organizational Development Network in Higher Education (www.podnetwork.org).

Validity, Research, and Reality: Student Ratings of Instruction at the Crossroads

Jennifer Franklin, University of Arizona

Many faculty use student ratings of instruction to get feedback to assess their teaching practices and course designs. Over the last decade, however, faculty have become increasingly aware of new ways to understand and practice teaching. As a result, the teaching that student ratings seek to evaluate itself has become a moving and changing target. While acknowledging that there are other useful ways to gather feedback, I am writing to raise this question: *Since the research that supports the use of student ratings of instruction was conducted during a time when most courses were given using conventional, face-to-face teaching methods, how can we use ratings to get feedback when we adopt new teaching methods and/or technologies unexamined by ratings resea*rch?

What has changed, and why should it matter?

One striking change is an emerging shift in roles and responsibilities from the teacher as presenter of content to a facilitator of learning, where focus moves away from what the teacher does with course content to what the learner needs and can do. Certainly, the lecture is alive and well, but such slogans as "Sage on the stage versus guide on the side" and the emergence of a learner-centered education movement signal a sea change. New methods often include an emphasis on collaborative and cooperative instructional strategies in which student-to-student interaction works as an arena within which students construct meaning and develop skills; strategies in which students must discover for themselves the content they would have been given to memorize in past years; a stronger emphasis on problem solving; and strategies taking students beyond the acquisition of concepts to analysis of the structure of an underlying knowledge domain.

Add to the mix the growing use of computer-based instructional technologies that offer whole new ways to communicate instructional content and mediate communication with and among students in our classes. For example, we have web-based systems such as Blackboard, WebCT, Weblogs (blogs), as well as instant messaging and conferencing systems (e.g., WebMeeting and Breeze) and mobile computing using PDAs. These tools will offer ways to teach that we could not have imagined before their advent.

What should we ask students that will provide unambiguous indicators of how well these strategies and technologies work? First, looking at established collections of questions, some universities are updating their forms to include collaborative instructional methods (see University of Washington's IAS system (http://www.washington.edu/oea/iasforms.htm), but there are few published item collections such as The Flashlight Student Inventory (http://www.tltgroup.org/programs/flashcsi.html) that aim to assess salient aspects of new teaching modalities. The overall quality of collections so far appears uneven. Insufficient formal investigation of item or instrument validity or reliability has been conducted. As a result, the ratings research literature will offer little direct guidance. Faculty should be prepared to further assess the validity and reliability of adopted or adapted items.

A problem with ratings research: "shelf life"

The body of research that guided the development of ratings questionnaires and the types of items in current use was largely conducted during a forty-five-year period (roughly 1955-2000) in settings where teaching practices (lecture, seminar, lab, and discussion) that most faculty would recognize from their own experience as students ruled. Thus, arguments for the validity and reliability of rating data and methods for constructing items are predicated on data mostly collected before current innovations took hold.

That same research revealed stable dimensions of teaching in conventional face-to-face courses, e.g., presentation, rapport, feedback, interaction, workload and difficulty (see, for example, Marsh & Dunkin, 1977). Do such dimensions translate to fully distance internet courses or courses that blend face-to-face meetings and online activities? Are there new, important dimensions of teaching effectiveness that we have not yet recognized? How do things like web design factor in? Unlike passive textbooks, interactive websites actually have "behavior."

Hundreds of respectable studies have examined the association of teacher, student, and course variables, e.g., gender, age, personality, grade point average, class size, academic discipline, and course level, with ratings and found significant associations among some, but in other cases did not show associations where popular opinion held they existed. Could new and unanticipated sources of systematic variation or bias in students' responses

appear when ratings are used in new settings? Many of us launching into constructivist pedagogies have heard students waxing nostalgic for the days when they were told "what to learn". Do students' expectations and orientations to one kind of teaching methodology dispose them to prefer certain methods regardless of how effective the instruction? From my own teaching practice I know not all students adapt equally well to fully distance courses.

I speculate that much of this work will generalize to new settings, but there is no way to prove that assertion until the research is updated to reflect changes in teaching practices and philosophies (and implied values). It will take time for scholarship in the field to catch up. Studies of these issues are on the rise, but the quality of the studies is highly variable. Meanwhile, we should assume that instructional innovation can pose real threats to both the validity and the reliability of diagnostic ratings items predicated on research conducted in conventional courses and proceed with due caution.

What faculty can do until the researcher arrives

So, should faculty embarking on instructional innovation avoid the use of ratings altogether until another twenty years of research accumulates? Of course not. Well-crafted ratings items remain an efficient way to get crucial timely and anonymous feedback for improving teaching. However, the emphasis in that sentence is on "well-crafted." Good diagnostic items are informed by careful analysis of how a teaching method is supposed to work and are constructed with respect for established practices for developing survey type items.

A starting point for faculty who need to develop their own diagnostic items can be found in the work of Murray (1997), who first described the use of "low inference" items which describe specific and observable teaching behaviors that also point to broader, more abstract dimensions of teaching such as clarity, enthusiasm, organization, interaction, pace, and rapport. In Murray's Teaching Behavior Inventory (TBI), students respond on a scale ranging from "almost always" to "never" to the frequency of those behaviors. Although Murray's work was done during a time when lecture, lab, and discussion were the rule (and the TBI remains excellent for assessing conventional classroom teaching), it is the "low inference" measurement approach that I am advocating.

If your students tell you that you are not communicating clearly, you will need more information to remedy the problem. Asking whether you are enunciating intelligibly, presenting information at an appropriate pace, signaling transitions between topics, repeating explanations of difficult concepts, or using clarifying examples will yield action items for you to improve. Getting at those more specific teaching behaviors is the goal of

writing "low inference" items. Let's say I want to get feedback to improve the way I set up my threaded discussion assignments. If I only wanted an overall assessment, I might ask, "How effective were the discussions in facilitating your learning?" But if I wanted feedback for improvement, I would ask things such as "Did you know what you were supposed to do?," "Was there sufficient time to complete the assignment?," "Did the rules of the road for discussion allow everyone a chance to participate?," which taken together would tell me how well the activity was working. In the interest of keeping the questionnaire short, I would reserve this detailed view for a few facets of the course and use more general questions for aspects I am not likely to be working on soon.

This approach can be applied to observable characteristics of any instructional interaction, including new teaching methods and applies whether you are adopting, adapting, or writing your own ratings items. At the same time, do not assume because you are experienced at writing quizzes or because you've taken a lot of surveys that you know everything you need to know about writing ratings items. Get an orientation to the unique characteristics of ratings starting with the sources I mentioned earlier. With such focused information and some basic skills in item writing, you can dramatically increase the value of student ratings feedback and more effectively assess the impact of your instructional innovations.

And back to research: a parting thought

It was that forty-year span of ratings research that helped us discover stable dimensions of teaching in the world of teaching as we knew it and gave us new instruments to help us understand students' perceptions and attitudes toward our work. As we find new ways to ask our students questions about how we are teaching, making a commitment to share what we have learned with each other and the research community makes us active participants in a learning community instead of consumers of ratings research factoids.

References & Resources

Arreola, R.A. (1999). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system,* (2nd Ed). Bolton, MA: Anker Publishing.

Doyle, K. O. (1975). *Student evaluation of instruction*. Lexington, MA: D.C. Heath and Co.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry and J.C. Smart (Eds.) *Effective teaching in higher education: Research and practice* (pp. 241-367). New York, NY: Agathon Press.

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R.P. Perry and J.C. Smart (Eds.*) Effective teaching in higher education: Research and practice* (pp.171-204). New York, NY: Agathon Press.

Jennifer Franklin (Ph.D., Indiana University) is Instructional Development and Assessment Specialist, Learning Technologies Center, University of Arizona.