Valid Faculty Evaluation Data: Are there any?

An interactive symposium exploring issues in evaluation and student ratings.

AERA Annual Meeting, Montreal, CA, April 14, 2005

Michael Theall, Youngstown State University

Background

Forty years of research on faculty evaluation, student ratings, and related issues have demonstrated that there are many ways to assess faculty performance (e.g., Arreola, 2000; Centra, 1993; Feldman, 1998; Glassick et.al., 1997; Scriven, 1994; Seldin; 1991; Theall, Abrami, & Mets, 2001; Theall & Franklin 1990). These writers suggest that a variety of sources can provide a range of data (quantitative and qualitative) that illuminate the breadth of faculty work and the extent to which it can be fairly reviewed for both formative and summative purposes. These writers also essentially agree that good evaluation practice requires multiple sources of data; that the purposes of the evaluation affect the kinds of data collected and the nature and distribution of reports of results; that contextual factors must be taken into account; and that good practice demands rigor, resources, and a high degree of confidence in the systems used and in the people who use them.

The ratings debate and the validity of practice

Nonetheless, when the topic is faculty evaluation, the debate seems to focus on student ratings of teaching and the host of arguments used against ratings as valid and/or reliable measures of teaching performance. Here again, there is a wealth of evidence that when ratings instruments are properly constructed, properly administered, properly analyzed, properly reported, and used in a comprehensive systematic way, they can provide information felt by teachers, administrators, and students to be useful (Marsh, 1987). Given this possibility, why has the debate raged and why has practice not kept up with the research? Some writers (Feldman, 1998; Theall & Franklin, 1990) have argued that the issue of validity has much more to do with process and day-to-day practice than with the psychometric properties of the instruments used or the sophistication of the analysis. Several reasons are offered for this. First is that those who use the data are ill-prepared for the task in terms of a lack of knowledge about evaluation and ratings and in terms of strongly-held personal beliefs that are in conflict with the literature (Franklin & Theall, 1989). Another reason is that while the research is substantial, it has been done using large databases of ratings data where the weight of the data has over-ridden occasional variance in conflict with the generalizations that are made. In other words, contextual factors can influence both quantitative results and the interpretation and use of the data in ways which violate the established principles.

Anomalies and their effects

For example, Franklin & Theall (1992) studied gender issues and ratings at two large universities. The overall analysis showed what the broad literature (e.g., Feldman's, 1992a, 1992b meta-analyses) stated: that gender of the teacher or the students does not have a consistent or significant effect on ratings. However, deeper analysis at the first institution, using academic discipline as a variable, found that in certain departments there appeared to be a gender bias favoring male faculty. Yet further analysis of course assignments showed that female teachers had been assigned a predominantly heavy load of lower-level, introductory, required, large courses: all factors associated individually with depressed ratings. In effect, women were assigned to teach courses in which one might expect reduced ratings, while men were teaching courses where the contextual factors might predict higher ratings. Thus, if gender bias existed, it might have been in course assignment rather than in the minds of the students. For verification, the same analysis was used in the same departments at the second institution but there, course assignments were evenly distributed among men and women, and women had a slightly higher average ratings than men. The context at the first institution appeared to have influenced the ratings in a direction opposite to what the literature suggested and to what the overall analysis of that institution's data showed. To state it differently, the ratings were valid and reliable, but the underlying administrative practices created a situation that was invalid and indirectly influenced the data. Such results could be simplisticly and mistakenly interpreted as student bias against women. Instances such as this give rise to many of the complaints heard about ratings and lead to further confusion about the legitimate place and use of ratings data. This reaction distracts from the real possibility that gender bias existed in course assignment, and it does nothing to remedy the situation. Left unchecked, the inequitable course assignments could continue, the ratings of women could continue to be lower, and an incorrect conclusion about ratings could be perpetrated.

Evaluation and instructional changes

Another emerging issue is the extent to which new technologies and instructional practices have created teaching and learning situations that are sufficiently different from traditional instruction that old methods of evaluation may be largely invalid (Sorenson & Johnson, 2003). At least in part, the differences between face-to-face instruction and on-line instruction are obvious, but in one small study (Ohler & Theall, 2004), computer science faculty made only one distinction between the two situations: they reduced the amount of direct lecture, replacing it with text delivery. If faculty do not take into account the changes in context and instructional environment, and if they continue to use methods more suitable to traditional situations, evaluation with instruments designed with traditional situations in mind can result in lower ratings simply because those faculty do not have the opportunity to capitalize on the skills and experience that have allowed them to succeed in the classroom. Evaluation of these new situations is often done using the same instruments (even validated ones) that were devised in response to literature about college teaching (e.g., Feldman, 1989). That literature relied on studies conducted in largely traditional classrooms where lecture and discussion were the primary methods used. Thus, while the instruments were valid for that context, the question is whether they continue to be valid for new contexts.

Portfolios

Other data have been suggested as replacements for ratings. Portfolios (Seldin, 1991) have gained popularity and contain an array of information that can be very complete and very useful, but some research (Centra, 1993; Robinson, 1993) has found that the portfolio process is too time-consuming and that users of portfolio data have difficulty in interpreting and using the wide array of data for summative purposes. They find it difficult to arrive at summative decisions. This is encouraging in one sense: namely that it means that faculty recognize the complexity of the situation. However, the range of portfolio materials is wide, and much of what can be submitted is material that can not be validated in the same way or as reliably as quantitative data from ratings. How, for example, should one's philosophy of teaching and learning be judged? Are there standards that apply consistently and accurately? How does the inclusion of innovative instructional methods get judged? And aside from the innate quality of the instruction, what effect does the quality of the documentation have? Does a well-written description of a marginally effective instructional technique get judged to be superior to a less-well-written description of a technique that had great success? Who determines success? What evidence is acceptable? What is the validity of a portfolio?

Is testing the answer?

A recent discussion thread on the professional listserves of AERA's Division J, the American Evaluation Association (AEA), the Professional and Organizational Development Network in Higher Education (POD), and various listserves devoted to science education has focused on the use of student learning data as the most appropriate measure of both learning and teaching. Critics of ratings insist that the established relationship between ratings and learning (established by Cohen, 1980; Abrami, d'Apollonia, & Cohen, 1990, and others) is spurious for two reasons. First, they say that Cohen's ratings/achievement correlation of .43 is not sufficiently strong to be useful evidence. Second, and perhaps more appropriate, is the argument that ratings do not measure learning. This is a legitimate point since ratings instruments were not and are not intended to measure subject knowledge.

One approach usable in certain kinds of courses is to use a standardized, validated measure of commonly accepted basic knowledge in the discipline. For example, the "Force Concept Inventory" is such a measure in physics, and it is accepted by most educators as a valid and reliable indicator of basic student knowledge in that discipline. However, most student learning is assessed with teacher-developed tests. These tests have not undergone any form of technical validation and few have gone beyond face validity with respect to content. There is no evidence of construct validity, predictive validity, or concurrent validity. A classic example of poor testing is the use of multiple choice exams to measure conceptual understanding. Such tests measure recall ability at best and they exacerbate the problems associated with grade orientation and surface learning (Entwistle & Tait, 1994; Ramsden, 1988; Weimer, 2002).

Even if classroom testing were as valid and reliable as ratings have been shown to be, a question would remain about the extent to which individual teachers should be held accountable for student learning. If there were a sudden move to eliminate ratings and replace them with tests of subject knowledge, it is entirely likely that the same arguments used against ratings would appear as reasons not to use tests results. Chief among these would be the charge that "grade inflation" would result (for differing opinions, see Greenwald & Gillmore, 1997a, 1997b; and Abrami & d'Apollonia, 1998). Currently, many try to make a grade inflation case against ratings, but if tests were used, there would be no standardization, and teachers could even more easily be charged with creating and using tests that make it easy for students to get 'A's. How would the quality of tests be assessed on a campus-by-campus basis? Who would undertake this huge task? How could summative judgments about the quality of the tests be made, even within departments where there is often strong disagreement about standards, the nature and extent of the content "covered", and the extent to which "good teaching" is thought to be synonymous with a grade profile that must include only a few 'A's, some 'B's, many 'C's, some 'D's, and a few mandatory 'F's? There is the major conundrum for those who oppose the use of ratings: while they want to measure effective teaching using test performance, they also demand that grades include a normal distribution with some students failing. The teachers who most loudly proclaim "standards" would now be near the low end of the scale when grades were used to determine merit, promotion, and tenure.

Summary

The evaluation of faculty performance is a complex task. Even if the discussion is restricted to teaching performance, there are wide gulfs between research and practice and between research findings and what the users of the data are willing to believe (Franklin & Theall, 1989). It is possible that the common complaints evaluators hear are not entirely based in fiction. Certainly, stories of administrative misuse of ratings and other data abound when evaluators or faculty developers share professional experiences. So, if there are validity problems with evaluation in general, the question is whether these problems can be solved methodologically or through improved practice.

The answer must involve not only improvements in both methodology and operational practice. It must involve a reconsideration of the full range of roles that faculty fill, and a careful examination of the work that faculty do and the skills they need to do that work well. Arreola, Theall, and Aleamoni (2003) have outlined a structure for doing this work, defining the professoriate as a "meta-profession" that encompasses both disciplinary ("base profession") skills and a large array of other, necessary skills required to undertake professional responsibilities in teaching, scholarly and creative activities, service, and administration. Embedded in this conceptualization is the notion that the meta-profession has certain characteristics including the need to be self-defining, self-controlling, and self-evaluative. (Theall, 2002) has suggested that the primary responsibility for leadership in faculty evaluation lies with the faculty themselves. Only when faculty are willing to undertake this responsibility and to work to make evaluation truly valid and reliable, will the situation change, and the status of the professoriate improve.

References

Abrami, P. C., & d'Apollonia, S. (1998). The positive relationship between course grades and course ratings: What is the cause and what, if anything can be done about it? Debate presented at the annual meeting of the American Educational Research Association. San Diego: April 17.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A.(1990)."The validity of student ratings of instruction: What we know and what we don't". *Journal of Educational Psychology, 82,* 219-231

Arreola, R. A.(2000) (2nd ed.) *Developing a Comprehensive Faculty Evaluation System*. Bolton, MA: Anker Publishing Company.

Arreola, R. A., Theall, M. & Aleamoni, L. M. (2003) "Beyond Scholarship: Recognizing the Multiple Roles of the Professoriate." Paper presented at the annual meeting of the American Educational Research Association. Chicago: April 22. Available at: http://www.cedanet.com/meta/Beyond%20Scholarship.pdf

Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness.* San Francisco: Jossey-Bass

Centra, J. A. (1993). "The use of teaching portfolios for summative evaluation" Paper presented at the 74th annual meeting of the American Educational Research Association. Atlanta: April 13.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Entwistle, N. & Tait, H. (1994) "Approaches to studying and preferences for teaching in higher education: implications for student ratings." *Instructional Evaluation and Faculty Development, 14* (1&2), 2-9.

Feldman, K. A. (1998) "Reflections on the effective study of college teaching and student ratings: one continuing quest and two unresolved issues." In J. C. Smart (Ed.) *Higher education: handbook of theory and research.* New York: Agathon Press.

Feldman, K. A. (1992a). College students views of male and female college teachers. Part 1: Evidence from the social laboratory and experiments. *Research in Higher Education,* 33 (3), 317-375.

Feldman, K. A. (1992b). College students views of male and female college teachers. Part 2: Evidence from students' evaluations of their classroom teachers. *Research in Higher Education,* 33 (4), 415-474

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583-645.

Franklin, J. & Theall, M. (1989) "Who reads ratings: knowledge, attitudes, and practices of users of student ratings of instruction" Paper presented at the 70th annual meeting of the American Educational Research Association. San Francisco: March 31. ERIC # ED 306 241

Glassick, C. E., Huber, M. T. & Maeroff, G. I. (1997) *Scholarship assessed.* San Francisco: Jossey Bass.

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52(11),* 1209-1217.

Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-751.

Marsh, H. W. (1987). Students evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253-388

Ohler, S, & Theall, M. (2004) "An exploratory study of teacher opinions about teaching on-line and face-to-face: the instructional choices of a sample of faculty in computer science. *Instructional Evaluation and Faculty Development,* Spring. Available at: http://www.umanitoba.ca/uts/sigfted/iefdi/spring_2004/spring.pdf

Ramsden, P. (1998) "Studying learning: improving teaching." In P. Ramsden (Ed.) *Improving learning: new perspectives."* London: Kogan Page.

Robinson, J. E. (1993). "Faculty orientations toward teaching and the use of teaching portfolios for evaluating and improving university-level instruction." Paper presented at the 74th annual meeting of the American Educational Research Association. Atlanta: April 13

Scriven, M. (1994) "Duties of the teacher." *Journal of Personnel Evaluation in Education, 8,* 151-184.

Seldin, P. (1991) *The teaching portfolio: a practical guide to improved performance and promotion/tenure decisions.* Bolton MA: Anker Publications.

Sorenson, D. L. & Johnson, T. D (2003) (Eds.) "Online student ratings of instruction." *New Directions for Teaching and Learning #96*. San Francisco: Jossey Bass.

Theall, M. (2002) "Leadership in faculty evaluation and development: some thoughts on why and how the "meta-profession" can control its own destiny." Paper presented at the annual meeting of the American Educational Research Association. New Orleans: April 3. Available at: http://www.cedanet.com/meta/meta_leader.pdf

Theall, M, Abrami, P. A. & Mets, L. (Eds.) (2001) "The student ratings debate. Are they valid? How can we best use them?" *New Directions for Institutional Research No. 109*. San Francisco: Jossey Bass.

Theall, M., & Franklin, J. L. (Eds.) (1990). "Student ratings of instruction: Issues for improving practice." *New Directions for Teaching and Learning #43*. San Francisco: Jossey Bass.

Weimer, M. (2002) *Learner – centered teaching."* San Francisco: Jossey Bass.

Critical issues in faculty evaluation: valid data and the validity of practice

Michael Scriven, Western Michigan University

In several past discussions (Scriven, 1987, 1988a, 1988b, 1990, 1994, 2004), I have described threats to effective faculty evaluation, and discussed the validity issues that surround the debate over the use of student ratings of teaching. My stance on the validity issues has not changed substantially and this short paper compresses my views and adds some new thinking in order to add the mix to this symposium discussion.

Basic issues

There are three issues at the forefront of this discussion. First, is the distinction between valid and invalid evidence with respect to the historic sources of evaluation data and some alternatives (Scriven, 1988). Second is the determination of the subject of teacher evaluation, that is, evaluation based on the "duties" I have described elsewhere (Scriven, 1994). Third is the inclusion of valid evidence of student learning as discussed in several recent forums (1). I will begin with a restatement of essential validity problems.

In Scriven (1994) I listed nine potential sources of validity for student ratings. All of these are predicated on the use of appropriate forms, process, analysis, and interpretation. In abbreviated form, these sources were:

1. The correlation of ratings to learning
2. The unique position of students as raters of their own learning
3. The unique position of students as raters of changes in their motivation
4. The unique position of students as raters of observable fact
5. The unique position of students as raters of style indicators
6. The position of students as raters of matters such as the face validity of tests
7. The position of students as consumers
8. Ratings as participation in a "democratic" process
9. The "best available alternative" argument

I proceeded to demonstrate that summative decisions based on even perfect correlations with learning (# 1 above) were inappropriate because they presumed something about performance rather than being based on sound decision rules. Thus, being highly rated on one dimension of teaching would not guarantee overall high quality nor would being poorly rated on one dimension suggest overall poor quality. This "guilt by association" factor is particularly dangerous because it is a form of stereotyping that can lead to incorrect or unfair decisions.

I considered items 2 through 6 as appropriate sources of validity because they provide data from what I called "cognitive witness", "affective witness", "eyewitness", and "well-placed witness" perspectives (p. 11). Students are the most frequent observers of many facets of teaching and learning and their collective opinions as witnesses can provide useful information, particularly when they are asked to observe specific behaviors or materials.

Though the discussion of students as consumers (# 7 above) often sheds more heat than light, I proposed that students' opinions were relevant because they represent the tastes and preferences of the students, this despite the fact that they do not represent the merit of the teaching involved. I stated, "It would be naïve to think that such considerations are not relevant to the value of a teacher, even if that is demarcated from merit." (pp 11-12)

I enumerated several issues pertaining to democratic process (# 8 above), pointing out that even though there is little support for the direct validity of the 'voting' argument, there are at least five areas in which students opinions can provide information that has positive potential. For example, ratings provide an outlet that can encourage more participation in teaching and learning and provide a voice that can lead to "…reduction of alienation or *anomie,* and increasing the student's sense of part-ownership of the institution and his or her own destiny." (p 13) I urged then, and urge now, that we explore the possible contributions of this aspect of ratings to validity.

Finally, with respect to alternatives (# 9 above), I suggested two possibilities. One was the careful use of highly qualified visiting observers as supplements to gathering student opinions. The other was to explore learning in a more rigorous way so as to estimate the quality of teaching on the basis of the amount learned. This "gain score" approach would have to be predicated on the development and use of particularly well-designed and developed tests of understanding, and it carries with it, the requirement that teachers must be skilled in test construction and related areas. Current evidence does not suggest that in general, we are at the point of accepting the proposition that all teachers and their tests are at this elevated stage of sophistication. However, recent work suggests that we are making progress.

Interactive instruction and learning gains

Hake (2002, 2004) discusses the use of a highly interactive model that incorporates pre-testing, in-class review, the use of various technologies such as remote response devices sending data to computers & software that can immediately display the distributions of student responses to questions, and follow-up activities designed to address problem issues. This interactive strategy requires more work of the teacher and the students and there is some evidence that ratings are reduced as a result of the increased workload (e.g., Greenwald & Gillmore, 1997). Nonetheless, Hake (2004) reports increases of as much as triple the gain scores over other methods of instruction. Three issues are important in considering this work. First, is the need to develop validated measures (e.g., the "Force-concept inventory" in physics). Second is to have teachers who know how to carry out the instruction and how to administer and use appropriate measures. Third, is the determination of the extent to which learning gains will be used in overall summative decisions. It is also important to consider the nature of the content and the level of the course involved. The interactive methodology would seem to work best in lower division courses that require student acquisition of basic knowledge, understanding of major concepts, and the ability to apply those concepts to the solution of problems. One ignores this issue of instructional context at one's own risk.

The skills and duties of the teacher

As noted above, success in teaching and learning and the valid evaluation of teaching performance require a combination of factors. First, is the clear delineation and definition of the duties and performance that will be evaluated. In “Duties of the teacher” (Scriven, 1994, p. 154) I proposed a hierarchy of three types of duties. “Generic duties” are those like subject matter knowledge, that are common to all teaching positions. “Job-specific duties” are those that are part of typical areas of responsibility, for example a course load and service responsibilities. “Site-specific interpretations of each duty” are the demands (and criteria?) imposed by the supervisor and the particular context in which the teacher works. Thus, a department chair can extend the duty of teaching ‘Psychology 101’ to include coordinating it as a multi-section course, supervising graduate assistants, and collecting and reporting assessment data. What is important, is to clearly describe both the nature of the duties AND the associated performance expectations. Just as teachers must provide a course syllabus or other documentation of the objectives, criteria, and conditions for student performance, supervisors and institutions must provide teachers with analogous documentation of duties, expectations, and evidence requirements.

A second step, since not all duties have equal weight, is to assign proportionate weights to these duties or expected performance areas. Next, is the determination of what kinds of evidence are appropriate and what sources of that evidence will be used. Appropriate analysis, interpretation, and use of the data follow, and finally, all the factors must be assembled into a coherent process with appropriate policies for operational decision-making. This sequence is quite like that proposed by Arreola (2002) for the development of a comprehensive evaluation system and it represents a more valid approach than is found at most institutions today.

But the discussion must extend beyond even these parameters because faculty are expected to perform in areas other than teaching and the scholarship of their disciplines. Arreola, Theall & Aleamoni (2003) have referred to these additional areas as the “meta-professional” skills of the faculty (2). And as with teaching, these additional areas require skills that more than likely have NOT been acquired in undergraduate or graduate education. Can an institution rightly expect faculty to possess all these skills at entry? Not currently, and thus, the expectation of fully capable faculty can not be met in the hiring process. As a result, what obligation does the institution have to assist its faculty in developing these skills and abilities? The implication is that the evaluation of faculty performance must be linked with institutional programs that support professional development as a necessary element in improving overall institutional performance.

Summary

There are numerous issues that impact on the validity of the faculty evaluation process and the use of data provided by students. While technical validity concerns are important, the validity of day-to-day practice is of equal concern and valid evaluation must follow a series of other, equally valid activities. In effect, valid faculty evaluation proceeds from a much larger requirement for the valid determination of overall faculty duties and the development of policy and process that not only define the necessary requirements, skills, and performance criteria, but also provide the resources that allow the development and refinement of these skills. Without this comprehensive approach, evaluation becomes a threatening and often punitive process that leads to suspicion and resentment: the antitheses of effective programs for both evaluation and development.

Reference Notes

(1) The listserves sponsored by AERA's Divisions, AIR, POD, STHLE, and other professional organizations, and subject-specific lists such as Phys-L, PhysLrnR, & Physhare have carried extended discussions of evaluation and ratings and issues. Some of these lists (as in Scriven, 2004 below) require subscription to access archived messages.

(2) Copies of several papers about the "meta-profession of the faculty" and an interactive set of matrices outlining faculty roles and skills can be found at: http://www.cedanet.com/meta

References

Arreola, R. A. (2000) (2nd ed.) *Developing a Comprehensive Faculty Evaluation System*. Bolton, MA: Anker Publishing Company.

Arreola, R. A., Theall, M. & Aleamoni, L. M. (2003) "Beyond scholarship: recognizing the multiple roles of the professoriate." Paper presented at the 83rd annual meeting of the American Educational Research Association. Chicago: April 22. Available at: http://www.cedanet.com/meta/Beyond%20Scholarship.pdf

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743-751

Hake, R. R. (2002) RE: Problems with student evaluations: Is assessment the remedy? Available as ref. 18 at: http://www.physics.indiana.edu/~hake

Hake, R.R. (2004) Design-based research: a primer for physics education researchers. Availabla as ref. 34 at: http://www.physics.indiana.edu/~hake

Scriven, M. (1987) Validity in personnel evaluation. *Journal of Personnel Evaluation in Education 1,* 9-23.

Scriven, M. (1988a) Duty based teacher evaluation. *Journal of Personnel Evaluation in Education 1,* 309-334.

Scriven, M. (1988b) The validity of student ratings. *Instructional Evaluation, 9(2),* 5-18.

Scriven, M. (1990) Can research-based teacher evaluation be saved? *Journal of Personnel Evaluation in Education 4,* 19-32.

Scriven, M. (1994) Duties of the teacher. *Journal of Personnel Evaluation in Education 8,* 151-184.

Scriven, M. (2004) RE: Is success dependent on technique – Hawthorne effect?" EvalTalk post, May 3, 2004. Available at: http://bama.ua.edu/cgi-bin/wa?A2=ind0405a&L=evaltalk&T=0&O=D&X=14AD7F14CF0B2527AF&Y=rrhake@earthlink.net&P=963

Using Teacher Ratings Forms to Evaluate Teaching: Doing a Better Job With What We've Got

Philip C. Abrami, Concordia University

Teacher rating forms (TRFs) completed by students are often used by promotion and tenure committees to arrive at summative decisions concerning teaching effectiveness. TRFs are often the major source and sometimes the only source of information available concerning a faculty member's teaching performance.

Promotion and tenure committees have a great responsibility; their decisions often determine the course of academic careers and the quality of departments. Mistakes, either favoring a candidate or against a candidate, are costly. How, then, should evidence on teaching effectiveness by weighed so that correct decisions are made?

Anecdotal reports suggest there is wide variability in how promotion and tenure committees use the results of TRFs. At one extreme, there are reports of how discriminations between faculty and judgments about teaching are based on decimal point differences in ratings. Experts in the area are often shocked to learn of such decisions but do not have sufficient means to prevent such abuses. At the other extreme, there are reports of how discriminations between faculty and judgments about teaching fail to take into account evidence of teaching effectiveness (i.e., instructors are assumed to teach adequately), so the importance of instructional quality is substantially reduced when assessing faculty performance. The correct use of TRFs lies somewhere in between these two extremes.

Drawing on my previous work (Abrami, 2001 a & b) concerning the use of TRF scores for summative decisions, I will describe ways for: improving the reporting of results; improving the decision-making process; and incorporating TRF validity estimates into the decision process. My recommendations for improving judgments about teaching effectiveness using TRFs include:

1. Report the average of several global items or a weighted average of specific items, if global items are not included in the TRF.

2. Combine the results of each faculty member's courses together. Decide in advance whether the mean will reflect the average rating for courses (i.e., unweighted mean) or the average rating for students (i.e., weighted mean).

3. Decide in advance on the policy for excluding TRF scores by choosing one of the following alternatives: a) include TRFs for all courses; b) include TRFs for all courses after they have been taught at least once; c) include TRFs for all courses but those agreed upon in advance (e.g., exclude small seminars); or d) include TRFs for the same number of courses for all faculty (e.g., include best ten rated courses).

4. Choose between norm-referenced and criterion-referenced evaluation. If norm-referenced, select the appropriate comparison group and relative level of acceptable performance in advance. If criterion referenced, select the absolute level of acceptable performance in advance.

5. Follow the steps in statistical hypothesis testing: a) state the null hypothesis; b) state the alternative hypothesis; c) select a probability value for significance testing; d) select the appropriate statistical test; e) compute the calculated value; f) determine the critical value; g) compare the calculated and critical values in order to choose between the null and alternative hypotheses.

6. Provide descriptive and inferential statistics and illustrate them in a visual display which shows both the point estimation and interval estimation used for statistical inference.

7. Incorporate TRF validity estimates into statistical tests and confidence intervals.

8. Since we are interested in instructor effectiveness and not student characteristics, consider using class means and not individual students as the units of analysis.

9. Decide whether and to what extent to weigh sources of evidence other than TRFs.

References

Abrami, P.C. (2001a). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P.C. Abrami, & L. A. Mets (Eds.), *New directions for institutional research: No.109. The student ratings debate: Are they valid? How can we best use them?* (pp. 59-87). San Francisco: Jossey-Bass.

Abrami, P.C. (2001b). Improving judgments about teaching effectiveness: How to lie without statistics. In M. Theall, P.C. Abrami, & L. A. Mets (Eds.). *The student ratings debate: Are they valid? How can we best use them? 109, 27*(5), 97-102. San Francisco: Jossey-Bass.

A Fractal Thinker Looks at Student Evaluations*

Edward Nuhfer, Idaho State University

"Dr. X's performance in the area of teaching meets expectations. However, if overall student evaluations do not improve during 2005, teaching may not meet performance expectations."

The quote is real, recent, and illustrates thinking that equates satisfactory teaching performance as synonymous with high student ratings. It refers specifically to tabulated agreement with a single item: "Overall this was an excellent course." Nothing else was taken into account; indeed nothing else truly counts in this unit's evaluative practice. The quote is from a report from within a College of Education. It represents a paradox: a culture with awareness of the necessity for multiple measures, which nevertheless cannot overcome its own entrenched practices of making personnel decisions based upon single-measure convenience. The practice of this particular College is far from an aberration. Even units that collect data from multiple sources usually fail to incorporate it in meaningful ways and instead resort to "evaluate" through reactions to single measures.

Fractals provide important insights to the student evaluation problem. A fractal form has the illusion of being randomly irregular when, in fact, order is present. Order includes the complex forms built from recursive operations on a small unit called a generator, similarity when viewed at different scales, and predictable growth of a dimension such as length in accord with decreasing length of measuring tool. Nature is full of fractal forms: trees, clouds, blood vessels, and landforms, to name a few. Education is replete with fractal characteristics in both space and time. Neural networks, like blood vessels, are fractal networks. In space, physical brain changes include growth of such networks in the process of becoming educated. Many natural temporal patterns in time are fractal, and learning, too, is a product of a series of events in time. Some transitions, for example the Perry stages (1968) change between stage 4 and stage 5, are punctuated events that are similar to the temporal events in fractal rainfall and flood patterns. The act of learning disciplinary content, learning to teach effectively, mastering good assessment practices, gaining ability to think at higher levels and making evaluative personnel decisions are all part of the same package that goes with development of appropriate synaptic networks. These manifestations of neural activity reflect the fractal quality of the physiological entities that produce them.

The fractal dimension is one of the essential manifestations required to describe a fractal form. Derivation of a fractal dimension requires multiple measures taken at different scales. There is no substitute for so doing; no single measure can capture the quality of a fractal form. Such measures require one to begin thinking in multiple scales. In education, these scales across an institution run from individual class sessions through signature traits of institutional degrees. The assessment movement appears to understand such thinking. Assessment refuses to accept convenient single measures as adequate, and assessment gets educators out of the rut of thinking myopically at just the level of courses and individuals. It recognizes that individuals are part of something larger, that education is more than courses, and that successful educators do much of profound value that cannot be captured by student satisfaction ratings.

College administrators who practice in accord with our example above do not comprehend the need to capture a complex entity such as "good teaching" through multiple measures. Such leadership is likewise prone to failing to appreciate the need to capture student learning through multiple measures, and it may account for the fact that assessment persists as the most vulnerable area to criticisms from accreditation agencies. Student learning outcomes are the most important assessable measure of institutional effectiveness. It makes sense that part of personnel evaluation of "good teaching" should require a direct look at learning changes produced in the classes of individuals within the institution. The relationship between student learning and evaluative ratings is positive and sound, but the trends taken on populations prove too weak to use to predict the learning that occurs in an individual's classrooms.

Rating of faculty by college students is an evaluative challenge—the highest level of challenge in Bloom's (1956) taxonomy. Ability to handle well the evaluative challenge in the special case of rating professors should be in accord with that to handle other evaluative challenges. The ability of students to do evaluative thinking rests upon their ability to use evidence and to meet a high Bloom-level challenge with a high level thinking response on the Perry scale. Compilations show that the average high school graduate reasons at Perry level 3 2/3, the average new baccalaureate holder reasons at level 4. In short, level 4 thinkers have difficulty in using evidence effectively and objectively. The outcome of thin-slices research is very much in accord with what we expect, based on the results that characterize thinking levels of undergraduates.

We know that evaluative ratings result from a mix of cognitive and affective factors. The importance of the latter is under appreciated, even though correlations reported between ratings and affective first impressions (thin slices) are higher than those reported between ratings and learning performance. Affective satisfaction is so important that it should be one multiple measure. However, satisfaction measures are less important than, and are no substitutes for, demonstrated promotion of learning. Student ratings alone cannot capture "good teaching," prove that learning occurred or serve to show outcomes were met.

Research, particularly that by Kenneth Feldman (1989), confirms benefits of particular instructional practices on both student learning and student satisfaction. This provides reasonable basis for including, as one essential multiple measure, a profile of pedagogical practices and the degree to which beneficial ones are present in the classroom. In short, there are reasons to require formative data be used in summative evaluative decisions.

*This presentation will describe the nature of fractals and show applicability of fractal thinking to education with emphasis on student evaluations. It results from a dozen years' progressive design of the Boot Camp for Profs™ program used in helping faculty to become more successful educators. It summarizes content from many articles on the "Educating in Fractal Patterns" series published by the author in National Teaching and Learning Forum between 2002 and 2005, and that used as a basis for "fractal" short courses at AAHE Assessment Conferences and Lilly Conferences. The presentation will outline the nature of fractal thinking, obviate the need for multiple measures and show parallels between educational design, learning assessment, and student evaluations. I'll show examples in which formative surveys and knowledge surveys successfully filled gaps inherent in student evaluations.

References

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956) *Taxonomy of educational objectives. The classification of educational goals. Handbook I: Cognitive Domain.* New York: David McKay.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education, 30(2)*, 137-194.

Perry, W. G. (1968) *Forms of intellectual and ethical development in the college years: a scheme.* New York: Holt, Reinhart, & Winston.

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

Jennifer Franklin, University of Arizona

Over the course of my career as a researcher and consultant on rating-related issues and faculty development, I've read hundreds of questionnaires, thousands of questions -- most were cobbled together from other questionnaires. As a researcher and administrator, my main concerns were poor construction of questions and response sets (items) and administrative misuse of the data they yielded. Over the last decade, faculty have become increasingly aware of new ways to understand and practice teaching. Now as a teacher myself and as an instructional development consultant in a university learning technologies center, I observe that effective teaching itself has become a moving target.

Many faculty use student ratings of instruction to get feedback to assess their teaching practices and course designs. 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 used in lectures, seminars, labs and discussion, how can we use ratings to get feedback when we adopt new teaching methods unexamined by ratings research? How do well-crafted but otherwise typical "diagnostic" items work in new settings? For example, do ratings questions used to ask students about face to face lecturing skills apply to online Powerpoint presentations with audio? Do questions about classroom discussion activities apply to asynchronous threaded discussion boards? to an internet chatroom? to instant messaging exchanges? What kinds of questions best get at effective collaborative learning experiences?

For this essay, set aside the issues related to using ratings in high stakes processes such as hiring, retention, promotion or tenure or administrative misuses of ratings data. Assume that experts generally agree that ratings as a measurement method are imperfect indicators of teaching effectiveness, but are sufficiently valid and reliable to provide useful information if they are properly used. (Two useful sources for guidance: Arreola (1999), & Doyle (1975))

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 (LCE) 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 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 can 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 at 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 in time (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.

The same research did reveal persistent patterns of relationship among diagnostic ratings items which in turn revealed stable dimensions of teaching in conventional face to face courses (e.g., presentation, rapport, feedback interaction, workload and difficulty). For recent sources see Feldman, 1989; & Marsh and Dunkin, 1997. Do such dimensions translate to fully distance internet courses, to 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 website 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, respectively) 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 developed “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. The current focus is whether online ratings data equate to classroom collected data. In strictly and empirically defensible terms, the literature of research and practice has not yet developed sufficiently to provide or validate generalizations that support practical guidance when adapting old items or writing new ones to explore the new instructional contexts. 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 we should proceed with due caution.

What faculty can do until the doctor, that is, 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 to capture specific and observable teaching behaviors that constitute broader, more abstract dimensions of teaching such as clarity, enthusiasm, organization, interaction, pace, and rapport. It is instructive that Murray began with live observation of instruction using replicable observation protocols and trained observers. Once vetted, those observations of effective and ineffective teaching behaviors were translated to ratings items for students, in effect making the students the observers. In Murray’s Teaching Behavior Inventory, students are ask to respond (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 constituent observable behaviors is the goal of writing “low inference” items. Extrapolating further, if I want to get feedback to improve the way I set up my threaded discussion assignments, and if I only want 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 whether the students knew what they were supposed to do; if there was sufficient time to complete the assignment, if the rules of the road for discussion allowed everyone a chance to participate, etc, 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 it 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 rating, 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 how we were teaching in terms of how it affected our 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 members of a learning community instead of consumers of ratings research factoids and the rantings of ratings opponents. Over the years I've heard a number of faculty comment on ratings studies they were reading, saying that they could do better. Maybe, maybe not. But somebody should.

References

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. Anker Publishing. Bolton, MA.

Doyle, K. O. (1975) *Student Evaluation of Instruction*. DC. Heath and Co. Lexington, MA.

Feldman, K. A. (1989) The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583-645

Marsh, H. and Dunkin, (1997) Students' evaluations of university teaching: a multidimensional perspective. In (Eds. Perry, R.P. and Smart, J.C. *Effective Teaching in Higher Education: Research and Practice*. Agathon Press, N.Y. p 241-367.

Murray, H. G. (1997) Effective teaching behaviors in the college classroom. In (R. P.. Perry, & J. C.Smart, (Eds.) *Effective Teaching in Higher Education: Research and Practice*. Agathon Press, N.Y. p 171-204.

Validity, like Beauty is…..

Raoul A. Arreola, The University of Tennessee Health Science Center, Memphis

Introduction

The question before us is "Are there any valid faculty evaluation data?" The traditional, research-based answer to that question is "Yes." Even though decades of research have demonstrated that tools used in faculty evaluation systems, especially student rating forms, can be designed so as to be both valid and reliable, the question persists. So… why does the question as to whether faculty evaluation data is valid still reverberate throughout higher education after so much research? My experience with hundreds of colleges and universities facing the task of revising or building their faculty evaluation systems has led me to conclude that there are three main reasons:

1. Many of the tools used in faculty evaluation systems are 'home made' and are thus of dubious validity and reliability;
2. Most academic administrators are not conversant with the finer points of psychometrics; and
3. Higher education has yet to establish a universally accepted definition as to the characteristics and skills necessary for teaching excellence.

These three conditions lead to a cascading set of circumstances that profoundly affect the general professoriate's perception, and willingness to accept, the possibility that any faculty evaluation data is, or can be, valid. A brief examination of these conditions reveals the underlying problem.

Homemade Faculty Evaluation Tools

Faculty evaluation tools have, by and large, been dominated by the use of some form of student rating form. Student rating forms have been, and continue today, to be the one main common element of all faculty evaluation systems. In many cases student ratings constitute the only systematically gathered data used in a faculty evaluation system. Unfortunately, the great bulk of student rating forms in use across higher education in America today are 'home-made'. That is, they have been constructed by committees comprised of various combinations of faculty, academic administrators, and students. Owing to the lack of psychometric expertise or rigor in constructing these forms, they are of dubious (and often undetermined) validity and reliability. This common situation has, over the decades, resulted in a rich and voluminous storehouse of anecdotes and stories leading to a number of 'myths' including the ever-popular myths that student ratings are just a popularity contest and that faculty can 'buy' good ratings by giving easy grades. Unfortunately, these and many other such 'myths' likely have a basis in fact since they may possibly be true for poorly constructed forms. The anecdotes that have produced these and other 'myths' about faculty evaluation tools are so common, so voluminous, and so widespread throughout the culture of higher education that they have taken on an aura of 'common knowledge.' Thus, in the minds of the professoriate and academic administrators this 'common knowledge' is so pervasive that it far overshadows the 'truth' concerning student ratings and other faculty evaluation tools buried in the pages of psychometric journals. Which brings us to the next point.

What the heck is psychometrics?

It is unfortunate but true that the majority of academic administrators are unfamiliar with the finer points of psychometrics. Deans and Vice Presidents for Academic Affairs are usually the ones who are faced with the task of gathering some form of evaluative information concerning their faculty's performance for the purpose of making promotion, tenure, continuation, or similar personnel decisions. Of the many hundreds of academic administrators with which I have had the occasion to interact on the issue of faculty evaluation, I could number on one hand the ones that have been sufficiently conversant with psychometrics to understand the subtle differences between, say, content and construct validity. These individuals, biologists, musicians, historians, physicists, physicians, nurses, pharmacists, etc., not only are generally unfamiliar with the finer points of psychometrics but rarely, if ever, read articles in such publications as *Journal of Educational Psychology,* or *Review of Educational Research.* A personal case in point:

> When I first joined the University of Tennessee Health Science Center in 1983 as chairman of the Department of Education I reported directly to the Vice Chancellor for Academic Affairs – a microbiologist of some note. At our first meeting he asked me in what area I had received my doctorate. I replied "Educational Psychology – specializing in psychometrics". His response was, first, incredulity, then contempt, and then he said "Well that must be some kind of phony degree cause I've never heard of that."

This anecdote may seem to represent an extreme case, but I have encountered some form of that response among any number of academic administrators. From the perspective of many of the people in academe responsible for making significant decisions concerning the design, structure, and format of faculty evaluation systems, the entire field of research endeavor that speaks to the issues of psychological measurement is, at best, a little-known area and, at worst, a 'phony' or illegitimate area of study. Thus, when the issue of the validity of faculty evaluation data is raised the definition or conception of validity used is much more likely to be that of colloquial usage rather than technical precision. It is useful to look at these two definitions. Below are the dictionary definition of "valid" and the technical definition of 'validity' from a psychometric perspective:

Main Entry: val·id
Pronunciation: 'va-l&d
Function: adjective
Etymology: Middle French or Medieval Latin; Middle French valide, from Medieval Latin validus, from Latin, strong, from valEre

1. having legal efficacy or force; especially: executed with the proper legal authority and formalities <a valid contract>
2. well-grounded or justifiable: being at once relevant and meaningful <a valid theory> b: logically correct <a valid argument> <valid inference>
3. appropriate to the end in view: EFFECTIVE <every craft has its own valid methods>
4. of a taxon: conforming to accepted principles of sound biological classification

- va·lid·i·ty /v&-'li-d&-tE, va-/ noun
- val·id·ly /'va-l&d-lE/ adverb

VALID implies being supported by objective truth or generally accepted authority <a valid reason for being absent> <a valid marriage>

[REFERENCE: MIRRIAM-WEBSTER ONLINE DICTIONARY]

Validity The effectiveness of the test in representing, describing or predicting the attribute that the user is interested in.

Content validity refers to the faithfulness with which the test represents or reproduces an area of knowledge.
Construct validity refers to the accuracy with which the test describes an individual in terms of some psychological trait or construct.
Criterion-related validity, or *predictive validity* refers to the accuracy with which the test scores make it possible to predict some criterion variable of educational, job, or life performance.

[REFERENCE: Thorndike, R.L. & Hagen, E. *Measurement and Evaluation in Psychology and Education* (Third Edition), New York: John Wiley & Sons, New York, 1969, pp. 655]

Looking at the differences between these definitions, and realizing that the majority of faculty and academic administrators use the dictionary definition rather that the psychometric one, it is easy to see why the question "Are there any valid faculty evaluation data?" persists. There is a large body of stories and anecdotes that gives credence to the popular myths that faculty evaluation data are invalid and unreliable. Unfortunately, there is no similar body of anecdotes that support the position that faculty evaluation data (especially student ratings) are *"well-grounded or justifiable, being at once relevant and meaningful, logically correct, or supported by objective truth or generally accepted authority."* Which leads us to our third point.

Excellence – I know it when I see it.

The word 'excellence', like the word 'diversity' has become so overused (and often inappropriately used) in higher education that it has lost its meaning. Once used to designate the 'best' or 'superior,' the word 'excellence' has come to mean a sort of a minimal expectation in virtually every faculty evaluation system. There are many problems with the use of this word in faculty evaluation – the main one being that 'excellence' is a term of relative position. That is, in order to be 'excellent' a person must be *better than* some one or some group of individuals. 'Excellence' is a norm-referenced term and not a criterion-referenced one. Yet, in most faculty evaluation systems the expectation persists that *all* faculty are to achieve excellence. I call this, of course, the lake woebegone model of faculty evaluation. As a profession, higher education is stuck in the silly verbal knot of expecting everyone to be what, by definition, only one or a few can be.

The vast, underlying problem in the whole of faculty evaluation is the fact that the academy has not come forward with a universally accepted definition as to what constitutes an excellent teacher. If we had some list of characteristics, some specific description of the qualities and characteristics that constitute an excellent teacher, then faculty evaluation would be relatively easy. Many faculty and administrators consider the main component of teaching excellence to be content expertise. Others hold that teaching excellence is some sort of ephemeral, immeasurable characteristic that results in some long-term (and perhaps never seen by the instructor) effect on student lives. We may never solve this particular problem to the satisfaction of all but, as an informal effort based on a career-long assimilation of research literature and professional experience, I would suggest the following as a jumping off point:

Characteristics of an Excellent Teacher

- ***Content Expertise***
    - Obviously a faculty member must be knowledgeable in the content field in order to teach it. However, content expertise is a *necessary* but *insufficient* quality for teaching excellence
- ***Affective Traits/Skills***
    - Enjoy teaching as much or more than they enjoy working in their field.
    - Model the best characteristics of an accomplished *practitioner* in the fields they are teaching.
    - Model the best characteristics of a *life-long learner.*
    - Is demanding but fair.
    - Is ethical and honest.
    - Is comfortable admitting ignorance.
- ***Performance Skills***
    - Speaks clearly.
    - Is organized when making a presentation.
    - Uses personal examples when teaching.
    - Uses humor effectively.
    - Creates an appropriate psychological environment for learning.
- ***Cognitive Skills***
    - *Instructional design* – develops and uses learning objectives in designing effective learning experiences;
    - *Instructional delivery* – skilled in presenting information in a variety delivery modes
    - *Instructional assessment* – skilled in the design and use of a variety of tools and procedures to assess student learning.

In Conclusion

In addressing the problem of faculty and administrators continuing to question the validity of faculty evaluation data it is necessary to take a different approach than simply conducting more studies that result in articles published in journals that appeal to a specific subset of educators. Rather, it is necessary to deal with the *perception* of the invalidity of faculty evaluation data rather than the researchable fact. This perception arises from several situations, the main ones having been described above. Therefore, I would propose the following:

1. The educational research community should undertake a concerted effort to reach the popular press with positive (but true) anecdotes and stories concerning the effectiveness and positive use of faculty evaluation data. The Chronicle of Higher Education has a fairly obvious bias against student ratings and is quick to publish any negative anecdotal stories concerning their use. A concerted effort must be undertaken to counteract this rather powerful public press force.

2. Educational researchers, especially those who have been most instrumental in producing the research literature on faculty evaluation, should make a concerted effort to produce the kinds of articles that will be published in instruments that are generally read by *all* academic administrators rather than just professionals interested in psychometrics and the psychology of teaching and learning. Such outlets as Magna Publication's *Academic Leader* or Anker Publication's *Department Chair* are a couple of examples.

3. As faculty evaluation professionals we must learn to speak the language of the non-psychometrically sophisticated academic administrator. Those professionals in the field of educational research who have gone on to take high-level academic administrative positions should take the lead in this endeavor. We must become more actively involved in professional organizations to which academic administrators belong, presenting papers and conducting sessions that are non-technical in nature and respond to the needs of administrators faced with difficult personnel decisions, but which are still founded on research findings.

4. Finally, we may wish to turn our attention, as educational researchers, to the task of developing and promulgating in the popular press, a definition of the qualities and characteristics of teaching excellence. Although there is a body of research literature on this issue, it has not yet been packaged and 'sold' to the general academic community (and society in general) in a form that could ultimately reach the level of 'common knowledge'. A first attempt in moving in this direction is presented above in the brief discussion on teaching excellence. It is important to remember the extraordinary power of the popular press in affecting the function of higher education in America. Even incomplete models of excellence, such as that used by US News & World Report to rank colleges and universities, can have an extraordinarily powerful impact on the priorities, and thus the functioning, of higher education. As professionals concerned with the valid and reliable evaluation of faculty performance, we would be remiss in ignoring this reality.