STUDENT EVALUATIONS OF TEACHING: A LITERATURE REVIEW OF EVALUATION INSTRUMENTS

David B. Whitlark, Brigham Young University,
Marriott School of Management, Provo, UT 84602; (801) 422-4994.

Michael D. Geurts, Brigham Young University,
Marriott School of Management, Provo, UT 84602; (801) 422-2398.

Gary K. Rhoads, Brigham Young University,
Marriott School of Management, Provo, UT 84602; (801) 422-2198.

ABSTRACT

This paper looks into the variables that are thought to impact university teacher evaluations. Also the paper looks at an evaluation instrument for one university to determine its statistical validity. The literature has produced long lists of items associated with favorable student ratings of college courses and professors. This research should be used in designing evaluation instruments.

FINDINGS FROM PRIOR RESEARCH

Researchers find that high teaching evaluations are due to many different characteristics or dimensions. Wortuba and Wright (1974) identify nine characteristics of good teachers. Feldman (1976) describes nineteen characteristics of ideal and/or best college teachers drawn from a review of forty-nine separate studies. Marsh (1982) identifies nine dimensions associated with his Students' Evaluation of Educational Quality (SEEQ) instrument. Cashin and Downey (1992) single out six dimensions of SEEQ associated with the Instructional Development and Effectiveness Assessment (IDEA) rating system. The summarized findings of these studies are that the components driving student perceptions of effective and ineffective teaching can be divided into four categories:

(1) Instructor characteristics,
(2) Course characteristics,
(3) Classroom characteristics, and
(4) Evaluation bias.

Instructor Characteristics

Wortuba and Wright (1974), Feldman (1976), Marsh (1982), and Cashin and Downey (1992) identify instructor enthusiasm as an important influencer of student perceptions. Marsh (1982) defines instructor enthusiasm as being enthusiastic about teaching, dynamic and energetic, humorous, and having a teaching style that holds ones interest. Degree of organization, individual rapport with students, and ability to administer and grade exams fairly are other qualities that these authors found influencing student evaluations.

Wortuba and Wright (1974) also cite a professor's willingness to experiment and the professor's efforts to encourage student thinking and communication skills as being important. Feldman (1976) enlarges the list with instructor knowledge, intellectual depth, elocution, clarity, availability to students, ability to encourage discussion, providing timely and meaningful feedback, keeping classroom discussions focused, and being sensitive to class level and progress as attributes that influence student perceptions. Kulik and Kulik (1974) report that having the ability to communicate effectively is strongly associated with good ratings. Freeman (1994) reports that students prefer androgynous instructors to either dominantly male or female instructors. That is, students prefer instructors who are both affectionate and forceful; sensitive to others needs while willing to defend their own beliefs; being very compassionate, but having a strong personality.

Light (1990) describes the characteristics of highly respected courses from the perspective of undergraduate students at Harvard University. The

students report that the single most important factor for making a course effective is getting quick and detailed feedback on class assignments and examinations. For example, they recommend instructors hand out an excellent example of a completed assignment at the same time they turn in their assignment for grading. A second key factor is being able to submit an early version of an assignment, get detailed feedback from the instructor, and then hand in a final version for grading. Frequent opportunities for evaluation and good course organization also are named as key factors.

Course Characteristics

Marsh (1982) mentions learning value, clear and prepared course materials, stated and pursued objectives, lectures that facilitate note taking, breadth of course coverage, tests that emphasize course content, assignments that add to course understanding, and appropriate workload as factors linked with favorable student ratings. Feldman (1976) also reports intellectual challenge and perceived value of material as key factors. Brightman, Elliott, and Bhada (1993) identify course content that discusses major developments, contrasts theories, presents the origin of ideas, and discusses new points of view as being particularly important.

Classroom Characteristics

Small class sizes usually lead to favorable teacher evaluations (Marsh and Dunkin, 1992; Marsh, 1987; Marsh 1984; McDaniel and Feldhusen, 1971). Elective courses tend to earn better evaluations than required courses (Lovell and Haner, 1955). Marsh (1982) identifies classroom interaction in which "students share ideas and knowledge" as a positive factor.

Evaluation Bias

There are many sources of evaluation bias. Marsh and Dunkin (1992) report that a student's prior interest in the field of study significantly influences student evaluations. Feldman (1976) finds that a student's expected grade has a significant influence on ratings. Jacobs and Kozlowski (1985) find that halo error, that is the power of general positive or negative attitudes towards a person to influence performance ratings, increases as students observe an instructor over longer periods of time.

STUDENT RATINGS VERSUS STUDENT ACHIEVEMENT

Research findings are mixed regarding the association between student ratings and student achievement. Marsh and Dunkin (1992), Abrami, d'Apollonia, and Cohen (1990), and Feldman (1989) show that student learning is positively correlated with favorable teacher evaluations. On the other hand, Rodin and Rodin (1972) report a negative relationship between ratings and student learning. In a group of studies regarding the "Dr. Fox" effect, Abrami, Leventhal, and Perry (1982) report that instructor expressiveness not student achievement drives student ratings. The best predictor of student achievement, they find, is course content.

Winer (1999) and Wiesenfeld (1996) suggest that our recent emphasis on earning high student evaluations has had negative consequences for education and society. One must ask whether it is ethical for instructors to teach to student ratings when doing so may not prepare students for succeeding in an increasingly competitive business environment. For example, consider the implications of minimizing the factors that generate classroom and individual stress, a key barrier to creating student perceptions of having a peak learning experience. In the business world, we frequently face problems that are ambiguous, demand personal initiative, and whose solution depends more on the application of a dynamic thinking process than on recalling what has been done in the past in similar situations. By minimizing learning-related stress, we may be minimizing the chances for a student's real-world success.

Is learning an inherently painful process with the satisfaction to follow once one reflects on personal achievement? Is there a way in which instructors can challenge students to attain a high level of scholastic achievement, yet make the process pleasant and comfortable? As we think about business education today, the role student

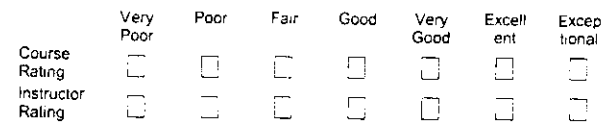
ratings play in shaping curriculum and teaching styles, and particularly as we move to outcomes-based business education (Bush and Sjolander, 1996), we must consider these questions and perhaps be willing to experiment with the balance between student comfort and student achievement.

Desai, Damewood, and Jones (2001) compare importance ratings of fifty-one evaluation items across a sample of college students and faculty. For many items there was general agreement between students and faculty. However, there are other items such as "setting up individual meetings or special sessions when needed," for which students consider to be much more important than do faculty. The authors suggest one way to balance the needs of faculty with the satisfaction of students is through applying a simple model that assigns each evaluation item to one of four quadrants and then taking the appropriate action. That is, items easy for faculty to accommodate and important to students (strategic), items difficult for faculty to accommodate and important to students (hold the line), items easy for faculty to accommodate and unimportant to students (manage to advantage), and items difficult for faculty to accommodate and unimportant to students (avoid).

INSTRUMENT VALIDITY

One concern is that instruments given to students to evaluate the learning process may not produce statistically valid results. If the instruments are not valid is it appropriate to use them to evaluate faculty and learning experiences?

Validity is often defined as "measuring what was intended to be measured." The first question is what should be measured. Some instruments may really measure teacher expressiveness rather than student development. Professors with low evaluations may point out that they are providing students with a great amount of useful information and that students need pain to attain. Consider the instrument used at one university. The instrument is one page. It has been in use for more than twenty years. It is organized as follows:



After the overall ratings for course and instructor, the instrument lists several items related to the course and then the instructor such as "assigned workload is appropriate for credit hours," "assigned homework is not just busy work," "assignments are appropriately distributed throughout the semester," etc. Each item is rated on an asymmetric seven-point scale anchored at "strongly disagree," "disagree," "somewhat disagree," somewhat agree," "agree," "strongly agree," and "very strongly agree."

One of the first potential problems with the instrument is the questions are not consistent with the items that research identifies as being related to an excellent educational experience. For example, there is no mention in the research that "assignments are appropriately distributed throughout the semester " is tied to quality education. The same is true for many of the items in our example and may be true for many of the items used at other universities. Another potential problem is that all of the scales used are ordinal rather than interval so the only appropriate statistical analysis is to calculate percent of responses in each category rather than mean scores. Also, many of the questions are double-barreled and include ambiguous terms. For example, one item reads, "exams are a good measure of my knowledge, understanding, or ability to perform. Which question should the student answer? The exams are a good measure of knowledge. The exams are a good measure of my understanding. The exams are a good measure of my ability to perform.

Tenure and promotion committees almost always ignore everything but the overall instructor rating. In our example, the university assigns a one to seven numerical value to each adjective. Then a mean score is calculated for each professor for each class that is taught. A "very good" is assigned a five. Students when quizzed about the evaluations say that the "semantic space" between the terms "good" and "very good" is smaller than the "semantic space" between the terms "excellent" and "exceptional." Consequently, the validity of

means scores calculated from such scales is questionable. In our example university, a mean score of five is often considered as not being good enough to promote a professor. In fact a mean score of five is termed as not being very good even though in the instrument the term "very good" is assigned a value of five. One may say that the promotion and tenure process is flawed because an invalid instrument is combined with an inappropriate analysis to evaluate teaching performance.

One example case is too limited to make an inference to all universities. However, we recommend taking a careful look at the evaluation instrument at your institution. Perhaps the research cited in this paper will suggest a redesign for the instrument measuring your teaching effectiveness.

REFERENCES

Abrami, P. C. 1989. SEEQing the Truth about Student Ratings of Instruction. *Educational Researcher*, Vol. 43, 43-45.

Abrami, P.C. and S. d'Apollonia. 1991. Multidimensional Students' Evaluations of Teaching Effectiveness: Generalizability of "N=1" Research: Comment on Marsh (1991). *Journal of Educational Psychology*, Vol. 83, 411-415.

Abrami, P.C., S. d'Apollonia, and P. A. Cohen. 1990. Validity of Student Ratings of Instruction: What We Know and What We Do Not. *Journal of Educational Psychology*, Vol. 82, 219-231.

Abrami, P.C., L. Leventhal, and R.P. Perry. 1982. Educational Seduction. *Review of Educational Research*, Vol. 52, 446-464.

Bongiorno, L. 1994. The B-School Profs at the Head of Their Class. *Business Week*, October 24, 73-74.

Brightman, H. J., M. L. Elliott, and Y. Bhada. 1993. Increasing the Effectiveness of Student Evaluation of Instructor Data through a Factor Score Comparative Report, *Decision Sciences*, Vol. 24, 192-199.

Bush, R. F. and R. J. Sjolander. 1996. Outcomes-Based Education is Here; Are You Ready? *Marketing Educator*, Vol. 15, No. 2, 1-11.

Cashin, W. E. and R. G. Downey. 1992. Using Global Student Rating Items for Summative Evaluation. *Journal of Educational Psychology*, Vol. 84, 563-572.

Costa, P. T., Jr. and R. R. McCrae. 1992. Normal Personality Assessment in Clinical Practice: The NEO Personality Inventory, *Psychological Assessment*, Vol. 4, 5-13, 20-22.

Desai, S., E. Damewood and R. Jones. 2001. Be a good teacher and be seen as a good teacher. Journal of Marketing Education, Vol. 23, 2, 136-144.

Feldman, K. A. 1976. The Superior College Teacher from the Students' View. *Research in Higher Education*, Vol. 5, 243-288.

Feldman, K. A. 1989. 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*, Vol. 30, 583-645.

Fetterman, D. M. (ed.). 1988. *Qualitative Approaches to Evaluation in Education: The Silent Scientific Revolution*. New York, NY: Praeger.

Jacobs R. and S. W. J. Kozlowski. 1985. A Closer Look at Halo Error in Performance Ratings, *Academy of Management Journal*, Vol. 28, No. 1, 201-212.

Kulik, J. A. and C. C. Kulik. 1974. Student Ratings of Instruction, *Teaching of Psychology*, Vol. 1, No. 2, 51-57.

Light, R. J. 1990. The Harvard Assessment Seminars: Explorations with Students and Faculty about Teaching, Learning, and Student Life. *Harvard University Graduate School of Education and Kennedy School of Government*, Cambridge, MA.

Lovell, G. D. and C. F. Haner. 1955. Forced-Choice Applied to College Faculty Rating, *Educational and Psychological Measurement*, Vol. 15, 291-304.

Marsh, H. W. 1982. SEEQ: A Reliable, Valid, and Useful Instrument for

Collecting Students' Evaluations of University Teaching, *British Journal of Educational Psychology*, Vol. 52, 77-95.

Marsh, H. W. 1984. Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. *Journal of Educational Psychology*, Vol. 76, 707-754.

Marsh, H. W. 1987. Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Educational Research*, Vol. 11, 253-388.

Marsh, H. W. 1991. A Multidimensional Perspective on Students' Evaluations of Teaching Effectiveness: A Reply to Abrami and d'Apollinia (1991). *Journal of Educational Psychology*, Vol. 83, 416-421.

Marsh, H. W. 1994. Weighting for the Right Criteria in the Instructional Development and Effectiveness Assessment (IDEA) System: Global and Specific Ratings of Teaching Effectiveness and Their Relation to Course Objectives. *Journal of Educational Psychology*, Vol. 86, 631-648.

Marsh, H. W. and M. J. Dunkin. 1992. Students' Evaluations of University Teaching: A Multidimensional Perspective. *In J. Smart, ed., Higher Education: A Handbook of Theory and Research*, New York, NY: Agathon.

McDaniel, E. and J. F. Feldhusen. 1971. College Teaching Effectiveness, *Today's Education*, Vol. 60, 27.

Miles, R. E. The Future of Business Education. 1985. *California Management Review*, Vol. 27, 63-73.

Rodin, M. and M. Rodin, Student Evaluations and Teachers. 1972. *Science*, Vol. 177, 1164-1166.

Wiesenfeld, K. 1996. Making the grade. *Newsweek*, 17 June.

Winer, L. 1999. Pursuit of customer satisfaction ruins schools. *Marketing News* (Chicago), 2 August, 11.

Wortuba, T. R. and P. L. Wright. 1974. How to Develop a Teacher-Rating Instrument: A Research Approach. *Journal of Higher Education*, Vol. 46, 653-663.