Grade Inflation And Student Individual Differences as Systematic Bias in Faculty Evaluations

Marie-Line Germain and Terri A. Scandura

The media has recently exposed that grade inflation is a concern for higher education in North America. Grade inflation may be due to consumerism by universities that now compete for students. Keeping students happy (and paying) may have been emphasized more than learning. We review the literature on faculty evaluation and present a model that incorporates students' individual differences and grade inflation as sources of bias in teaching evaluations. To improve teaching effectiveness, and avoid consumerism in higher education, faculty evaluations must begin to focus on students and the reciprocal role of grade inflation in teaching evaluation.

Today, faculty are being held accountable for how well they serve the U.S. student population, and it has become common practice in universities and colleges for students to "grade" the professors that grade them. Grade inflation has become an issue in higher education; students' grades have been steadily increasing since the 1960's (Astin, 1998). In June 2001, a record 91 percent of Harvard seniors graduated with honors, and 48.5 percent of grades were A's and A-minuses (Boston Globe, 2001). Grade inflation has been under scrutiny, and there is a need to address exponential grade inflation (Bérubé, 2004). Several studies have linked grade inflation with students' ratings of faculty (Greenwald, 1997; Stumpf & Freedman, 1979). According Pfeffer and Fong (2002): "Grade inflation is pervasive in American higher education, and business schools are no exception" (p. 83).

Students' ratings of management faculty now serve dual purposes. First they provide faculty with feedback on teaching effectiveness. They are also used for faculty reappointment, promotion and/or pay increase decisions (Jackson, Teal, Raines, Nansel, Force, & Burdsal, 1999). Yet, Scriven (1995) identified several construct validity problems with student ratings of instruction, one of them being student consumerism. Consumerism results in bias due to information not relevant to teaching competency, but important to students such as textbook cost, attendance policy, and the amount of homework. Due to the impact on tenure and career, faculty might try to influence student evaluations, a phenomenon referred to as "marketing education," or even seduction (Simpson & Siguaw, 2000). Some have become alienated from the process of teaching evaluation entirely. Professors who have become hostile to evaluations (Davis, 1995) often do not use the feedback they receive in constructive ways (l'Hommedieu, Menges & Brinko, 1997).

Faculty Evaluations as Performance Appraisals

Since student ratings of faculty teaching effectiveness are used as one component of

Marie-Line Germain, Department of General Education, City College, Miami, FL. Terri A. Scandura, Department of Management, School of Business Administration, University of Miami.

Correspondence concerning this article should be addressed to Marie-Line Germain, Department of General Education, City College, 9300 South Dadeland Boulevard, Suite 700, Miami, FL 33156.

A previous version of this paper was presented at the Society of Industrial and Organizational Psychology meetings, Orlando, FL. (April 2003). The authors would like to thank Chris Hagan and Clara Wolman for their helpful comments on an earlier version of this paper.

faculty evaluation, it seems reasonable to consider these instruments as performance ratings. As such, they are subject to a number of possible biases, as has been shown in the literature on rating accuracy in Industrial and Organizational Psychology (Campbell, 1990; Murphy & Cleveland, 1995). A number of studies have indicated problems with the reliability of performance ratings (Christensen, 1974; Wohlers & London, 1989). As noted by Viswesvaran, Ones & Schmidt (1996), "...for a measure to have any research or administrative use, it must have some reliability. Low-reliability results in the systematic reduction in the magnitude of observed relationships..." (p. 557). The accuracy of performance evaluation ratings has been challenged as well (Murphy, 1991). This research has led to recommendations for improvement of rating accuracy. For example, Murphy, Garcia, Kerkar, Martin and Balzer (1982) reported that the accuracy of performance ratings is improved when ratings are done more frequently. However, faculty evaluations, in most cases, are only at the end of the course, leaving greater possibility for error. Other research has reported problems due to individual differences such as leniency or stringency (Bernardin, 1987; Borman, 1979; Borman & Hallam, 1991).

Construct validity relates to the level of correspondence between performance evaluation and the actual performance of an individual on the job. The construct validity of performance ratings has rarely been examined in the literature (Austin & Villanova, 1992; Lance, 1994). Recently, Scullen, Mount and Judge (2003) examined the construct validity of ratings of managerial performance using two samples and four different rating sources (boss, peer, subordinate and self). Their results indicated that lower order factors (technical and administrative) skills were better supported in by their data than higher order factors (contextual performance: human skills and citizenship behavior). They conclude "... that the structure of ratings is still not well understood" (p. 50). One might argue that teaching effectiveness is as complex or perhaps even more complex as the contextual performance aspect of managerial performance. Construct validity must start with a clear definition of the construct of interest (Murphy, 1989). In the case of faculty evaluations, there is no clear definition of the criterion of effective teaching upon which to develop rating instruments.

The Criterion Problem

Research has shown that there is no one correct way of teaching (Joyce & Weil, 1996). Marsh (1982) found that the single most important factor affecting student evaluations was the amount learned, and the least important was the course difficulty. Researchers seem to agree that good faculty evaluations should reflect the amount learned in a class. However, not all students agree that learning is the most important factor in evaluating an instructor. Affect or likeability for example, may be more important than knowledge imparted. Faculty evaluations' imperfections are perhaps due to the fact that they utilize fallible measures (Guilford, 1954; Nunnally, 1978). Previously identified biases can be grouped into four main categories: teaching effectiveness, student grading practices, teacher characteristics, and the format of evaluation forms. Perhaps of even more concern is that teaching evaluations have been linked to students' course grades (McKeachie, 1979) which suggests that the criterion has been contaminated.

Despite these concerns, there seems to be no other reliable alternative method of evaluating faculty, which explains their continued use. The use of ratings as evaluation tools increased by 57% between 1973 and 1993 (Seldin, 1993). However, research on the integrity of the measures and the evaluation process has not kept pace with this increased use, despite concerns raised about the issue for several decades. Some researchers support their validity (Greenwald, 1997), but some

challenge their validity and usefulness. For example, d'Apollonia & Abrami (1997) suggest that student ratings are unsophisticated and provide little guidance for teaching improvement noting that "only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable) [should be made on the basis of student ratings] (p. 1202)." Cashin (1995) reports that generally speaking, students' ratings of faculty are reliable and relatively free from bias. McKeachie (1997) concludes that students' ratings have some validity but should be supplemented with other evidence. Although Cashin (1995) believes that students' ratings of faculty, generally speaking, are reliable and relatively free from bias, McKeachie (1997) opposes them and suggests that students' ratings can be biased by variables other than teaching effectiveness. Research conclusions range from "valid, reliable, and useful to invalid, unreliable, and useless" (Aleamoni, 1981, as cited by Gordon, 2001: 6). Thus, there is no clear consensus regarding the construct validity and usefulness of faculty evaluations.

A Shift in Focus: From Faculty to Students

This paper takes a different perspective on faculty evaluations. Many studies fail to look at the relationship between grade inflation and the construct validity of faculty evaluation, concentrating only on arguments of their reliability. Only a few attempts have been made to relate attributes of students to faculty evaluations. An exception is McKeachie (1997) who reported that student ratings in first- and second-year courses may have lower validity than student ratings in more advanced courses (in which students have broader experience as a basis for their ratings). First, we review the research on the relationship of faculty evaluations and grade inflation. We then present a model that attempts to explain systematic sources of bias including grade inflation, but also student individual differences. We incorporate student characteristics such as learning style into existing models of grade inflation and faculty evaluations.

Faculty Evaluations and Grade Inflation

Research has supported the premise that one element in faculty evaluations is grades expected or obtained by students (Snyder & Clair, 1976). Grade fairness (Jackson et. al, 1999) also referred to as "examination and grading" (Marsh, 1982) or "grading quality" (Burdsal & Bardo, 1986; Worthington & Wong, 1979). Greenwald (1997) reports that students' ratings of instruction correlate positively with expected course grades. There may be a process of reciprocity operating (Aronson & Linder, 1965); when an instructor praises a student via good grades, in return, the student will praise the instructor by giving good evaluations.

Marsh (1982) reported that student evaluations reflect the effects of the teacher, not necessarily the course. Murray et al. (1990) discovered that different teacher personality traits that contribute to effective teaching differed markedly for different courses. Marsh (1993) considered a host of 'background characteristics' such as prior subject interest, overall GPA, teacher rank, workload, grade leniency, class size (McKeachie, 1997), sex of instructor, academic discipline, reputation (Griffin, 2001), fashion (Morris, Gorham, Cohen, & Huffman, 1996), and even instructor enthusiasm. The latter was illustrated quite vividly by the "Dr. Fox" experiment (Naftulin, Ware & Donnelly, 1973). In this experiment, the research team hired a professional actor to lecture enthusiastically and expressively to a group of graduate students. The lecture was exciting but completely devoid of content. Despite the lack of content, the actor received favorable ratings. This study suggests that students react more to faculty acting skills more than any other factor in their ratings (Sherman & Blackburn, 1975).

Faculty are also evaluated on other dimensions, including rapport with students (Jackson et al., 1999), which consists of

showing respect, allowing questions to be asked in the classroom, also referred to as "enthusiasm" by Marsh (1982). d'Apollonia et al. (1997) argue that instructor expressivity and grading practices can unduly influence ratings of instruction. Also, faculty interaction, and more specifically competition and cooperation for favorable evaluations and for enrollment in their course using "game theory" may play an important role in ratings. For example, a faculty member may talk negatively about other faculty members (Correa, 2001) thus creating comparison and competition. This affects student/faculty interaction and may affect the workload that a course demands and the willingness of a student to attend class. Finally, preparedness of the instructor may be related to faculty evaluations. Some refer to this as course organization and design (Jackson et al., 1999; Burdsal & Bardo, 1986); others have labeled it organization and clarity (Marsh, 1982). Additional factors have been identified, such as workload and difficulty of the course (Jackson et al., 1999; Marsh, 1982; Burdsal & Bardo, 1986). For instance, many first-year science courses are used to weed-out the weakest students (Greenwald & Gilmore, 1997). Powell (1977) suggests that reduction of work in class or giving more paper assignments or quizzes can raise students' grades thus improving ratings.

Other researchers have identified additional biases that may inflate ratings. One is the motivation for courses, which can affect both grades and ratings (Howard & Maxwell, 1980; Marsh, 1984). For example, it may make a difference if the course is required or is an elective. The format of the student's evaluation form itself has also been questioned. Most universities use standardized forms that don't recognize individual course content (Divoky, 1995). Content of the rating forms as well as the number of items might affect ratings. It may not be possible to evaluate effectively with just four or five questions. Per measurement theory, longer inventories can be more precise than shorter ones and results are therefore more reliable (Nunnally, 1978). Evaluations are multifaceted and ratings should reflect this (Marsh, 1984; 1993).

Some remedies have been suggested in order to avoid or reduce bias factors. One of them is to train student raters to reduce halo effects and leniency and reduce psychometric error in student evaluations of instructor performance (Cook, 1989). Also, by weighting items, some believe that factor scores could result in improved rating (Abrami, 1989; Marsh, 1993). Additionally, involving faculty in the creation of rating forms may reduce skepticism and improve use of the feedback. Some have even suggested an alternative to student evaluations which would be to assess a teaching portfolio, which would be updated annually (Defina, 1996). Finally, the idea of providing midcourse evaluations has been suggested, which could increase rapport with students and treat students as partners in the teaching/learning process.

Individual Differences and Student Ratings of Faculty

Figure 1 depicts our model of the role that student individual characteristics play in determining both grade inflation and faculty evaluations. A reciprocal relationship is shown between grade inflation and faculty evaluations as previous research has indicated. Student individual characteristics and grade inflation represent sources of systematic bias in faculty evaluations. In the following sections, we review specific student characteristics that may be sources of systematic bias in faculty evaluations.

Individuals have different learning styles. Learning styles are different approaches or ways of learning. Accordingly to Kolb (1976; 1984) there are four types of learners: "Concrete Experience" learners are hands-on individuals who rely on intuition rather than logic; "Reflective Observation" learners make careful observations from

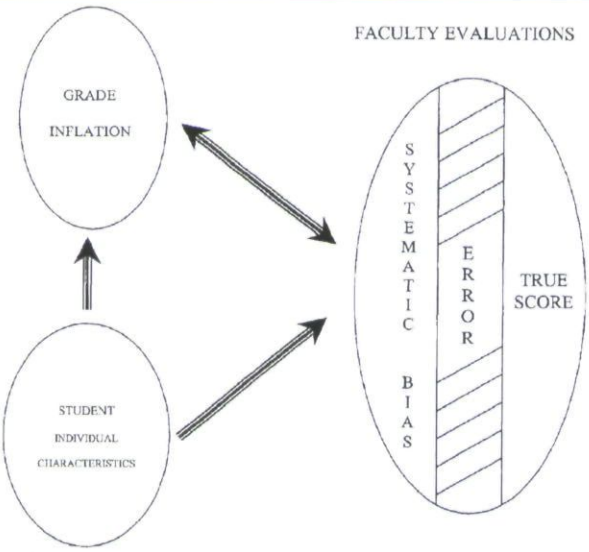

Figure 1. Revised model of current approaches for faculty evaluation.

many points of view; "Active Experimentation" learners like solving problems and finding practical solutions and uses for their learning; and finally, "Abstract Conceptualization" learners consider abstract ideas and concepts to be important.

McKeachie (1990) related learning styles and faculty ratings and reported that students may rate speech and material presented differently. Therefore, there is probably not a single criterion for teaching effectiveness. Although it is rare that students fall into one learning style, most people have a clear and prominent preference. It makes sense that an Active Experimentation student will not learn as much from a professor's presentation if he/she solely bases lectures on theory. Therefore, what students learn in a specific class may be a function of their learning style, and not the effectiveness of the professor.

Unless the instructor assesses students' learning styles at the beginning of each new class and adapts his or her lecturing style accordingly, there is no sure way that the lecture content will reach the students equally. This could explain the discrepancies in the amount learned in class as well as the perception of effectiveness of the instructor. Thus, learning style may contaminate student ratings of faculty teaching.

Another factor that may affect ratings is the existence of learning disabilities. These are usually hidden disabilities because they are not apparent to the outside observer, and are frequently overlooked when initiatives are undertaken for people with disabilities. Those who live with a learning disability experience its impact on a daily basis; this impact often has ramifications, not only in academic settings, but also in other facets of life, including vocational as well as social settings. In an academic setting, unspoken disabilities may make it more challenging for a student to learn and/or understand. Students may be reticent to inform faculty of these disabilities, and it may be difficult for an instructor discern them and make accommodations. Ultimately, if a student has difficulty understanding due to a personal handicap (whether it is a known handicap or not), the grade obtained in that class may be affected. Similarly, the student with a disability may see the instructor negatively because he/she may have not met the student's special needs. The instructor may have talked too fast, inaudibly, or may have simply not

repeated the concepts enough times.

Students' stage of personal development should also be considered. Individuals pass through certain stages of development in their life, and their needs tend to change (Maslow, 1954). How a person acts in a given situation depends partly on the demands and uniqueness of the moment and partly on the general developmental level at which he or she is functioning. In any classroom, there are individuals who are controlled to some degree by their own needs. In addition, depending on the student's developmental stage, his/her needs may differ (Levinson et al., 1978). Some students may still be in need of reinforcement and security whereas others feel safer and look for to growth opportunities without the need for support. Others have a high need for achievement and want the course to be challenging. This might make a difference on how the student will perceive both the course and the instructor.

The practical relevance of the course and whether the students can directly apply the concepts taught in class will also affect the way the student sees the course as well as the motivation level of the course. It may take a few years, once the student has actually graduated, before material taught in class can be applied. Until then, the course taken may simply be seen as a waste of time.

The reason why students are attending college may also influence how they rate faculty. Individuals who choose to go to school (or the particular school they attend) may be more likely to be positive about their experience. Individuals who feel forced to attend by parents or the company they work for may be less agreeable overall and more willing to complain.

Another critical factor that will influence the faculty ratings is students' previous relationships with instructors or their popularity with students on campus. This rapport differs from showing respect, allowing questions, and enthusiasm in the classroom as found with the Dr. Fox Experiment (Jackson et. al, 1999). Examples include the degree to which the instructor is well known on campus or whether a student has had the instructor in a previous course. The pre-existing relationship may affect the rapport during the semester. Because the student has pre-existing knowledge about the instructor, faculty evaluations may be biased compared with students who have not had the professor previously.

Student demographics, including gender, race/ethnicity and age may be important factors in faculty evaluations. For example, several studies reviewed by Koblitz (1993) have found that male students rate women instructors more harshly than female students. A student's race may also influence the way a faculty are rated. Minority students may be lenient with minority faculty as they may see them as role models. Finally, the age of the student may affect the way faculty are perceived. More mature students may be less judgmental with older faculty members. Similarly, younger students may be more likely to give good evaluations to young faculty members. Typical generation gap issues could explain this phenomenon.

Socioeconomic status may also affect how students relate to faculty. From an economic standpoint, wealth or hardship of a student could affect grades obtained in the class and subsequent evaluations. Students who are on grants or scholarships are obligated to obtain A's or B's to maintain their benefits. They may use this element to leverage and possibly negotiate their grades with the instructor. Also, fearing the loss of financial aid might make students grade professors that give C's on course assignments more harshly.

Finally, our model suggests that cultural beliefs strongly influence the values and behavior of the people who grow up in the culture, often without their being fully aware of it. Response to these influences varies among individuals. More specifically, a student's cultural background can influence

how they react to their learning environment. The effect of cultural context has been long recognized as having an effect on student learning (Holloway, 1988) and may similarly influence student evaluations of instruction. For instance, rating faculty is seen as a consumerism approach in Hong Kong and because of students' cultural background, the paternalistic image and therefore respect shown for teachers, they pay more attention to their evaluations. In this context, they may focus more on the personal qualities of their teachers and not the teaching content and grades obtained. As Ting (2000) notes, due to their cultural background, Chinese students "pay more attention in their faculty evaluations to the personal qualities of their teachers" (p. 637). In Europe, teaching evaluations are quite rare, and the implementation of faculty evaluations in Germany in the mid-1990's was a great source of controversy (Rindermann & Schofield, 2001). Most European universities do not have student evaluation of faculty, a practice that is perceived as unconceivable because of the formality of student/faculty interactions. Most studies on faculty evaluations are undertaken within a North American context. However, international students in these universities may respond very differently on faculty evaluation forms.

Suggestions for Future Research

Students' individual characteristics clearly need further study as a source of systematic bias in faculty evaluations. The impact of individual differences such as student gender, race/ethnicity, and learning style on faculty evaluations is yet unknown. Interesting research questions arise from this approach. For example, research is needed to examine the impact of student race and gender on rating black female instructors. Similarly, minority students might grade white instructors more harshly. Also, research on emotional state and/or mood of students while filling out the faculty evaluations should be conducted. The amount of effort students actually put into faculty evaluations should also be studied. Perhaps evaluation forms need revisions to encourage students to separate the quality of instruction from the grade they expect to receive in the class. Also, research is needed to examine how well students understand that faculty evaluations are used for faculty retention, promotion and salary increases. Finally, an international or cross-cultural study of faculty evaluation perceptions by students could be very revealing and these results could be used to better understand the role that national culture may play in the evaluation process, given the increasing number of international students in U.S. colleges and universities.

Conclusion

To improve accuracy, it is necessary to include assessments of students' characteristics and grade inflation in the faculty evaluation process. The impact of this issue is undeniably important to all constituents: Students, faculty, and administrators. Improved evaluations should improve the quality of student education. Learning conditions might be improved, which might ultimately lead to improved student retention. For faculty, this approach will bring more relevance and understanding of how to interpret faculty evaluations, which could lead to fairer promoting, tenure and pay increase decisions for faculty members, improved job satisfaction, and ultimately a sense of justice. For management students, improving the faculty evaluation process may ultimately improve knowledge acquisition. As Pfeffer and Fong (2002) note "... neither grades in business school nor completion of the program may provide much evidence of learning" (p. 83). Rewarding faculty for teaching management with rigor should affect learning, and cease brokering and game theory approaches to good grades for good evaluations. Finally, incorporating individual differences of students into models of faculty evaluations will improve understanding of the implications

on grade inflation for the administrators who are concerned both with student recruitment and retention as well as retaining the most effective faculty.

References

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 80*, 221-227.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.). *Handbook of teacher evaluation*. (pp. 110-145). Beverly Hills, CA: Sage

D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198-1208.

Aronson, E. & Linder, D. (1965). Gain and loss of esteem as determinants of interpersonal attractiveness. *Journal of Experimental Social Psychology, 1*, 156-171.

Astin, A. W. (1998). The changing American college student: thirty-year trends, 1966-1996. *The Review of Higher Education, 21*, 115-135.

Austin, J. T. & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836-874.

Bernardin, H. J. (1987). Effect of reciprocal leniency on the relation between consideration scores from the leader behavior description questionnaire and performance ratings. *Psychological Reports, 60*, 663-682.

Bérubé, M. (2004). How to end grade inflation: a modest proposal. *New York Times Magazine*, May 2, p. 32.

Birbaum, R. (1977). Factors related to university grade inflation. *Journal of Higher Education, 48*, 5, 519-539.

Borman, W. C. (1979). Individual differences correlates of accuracy in evaluating others' performance effectiveness. *Applied Psychological Measurement, 3*, 103-115.

Borman, W. C. & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual differences correlates. *Journal of Applied Psychology, 76*, 11-18.

Boston Globe (The). (2001). Rampant grade inflation seen at Harvard. Other Institutions. February 7.

Burdsal, C. A., & Bardo, J. W. (1986). Measuring students' perceptions of teaching: Dimensions of evaluation. *Educational and Psychological Measurements, 56*, 63-79.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, 2nd ed., pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Christensen, L. (1974). The influence of trait, sex, and information accuracy of personality assessment. *Journal of Personality Assessment, 38*, 130-135.

Cook, S. (1989). Improving the quality of student ratings of instruction: A look at two strategies. *Research in Higher Education, 30*, 31-45.

Correa, H. (2001). A game theoretic analysis of faculty competition and academic standards. *Higher Education Policy, 14*, 175-182.

Davis, M. (1995). *Staging a pre-emptive strike: Turning student evaluations of faculty from threat to asset*. Paper presented at the annual meeting of the Conference on College Composition and Communication. Washington, D.C.

Defina, A. (1996). An effective alternative to faculty evaluation: The use of the teaching portfolio. (ERIC Document Reproduction Service No. ED394561).

DeMarrais, K. B. & LeCompte, M. D. (1999). *The way schools work: A sociological analysis of education*. New York: Longman.

Divoky, J. (1995). Eliciting teaching evaluation information interactively. *Journal of Education for Business, 70* (6), 317-332.

Felman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education, 5*, 243-288.

Fresco, B. & Nasser, F. (2001). Interpreting student ratings: consultation, instructional, modification, and attitudes towards course evaluation. *Studies in Educational Evaluation, 27*, 291-305.

Gordon, P. A. (2001). *Student evaluations of college instructors: An overview*. Working Paper, Valdosta State University.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182-1186.

Greenwald, A. G. & Gilmore, G. M. (1997).

Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26*, 534-552.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Holloway, S. D. (1988). Concepts of ability and effort in Japan and the United States. *Review of Educational Research, 58*(3), 327-345.

Howard, G. S., & Maxwell, S. E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*, 810-820.

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C. & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580-596.

Joyce, B., & Weil, M. (1996). *Models of teaching* (5th ed.). Boston: Allyn & Bacon.

Koblitz, N. (1993). Bias and other factors in student ratings. *Chronicle of Higher Education*, September 1.

Kolb, D. A. (1976). *The Learning Style Inventory: Technical Manual*. Boston, MA.: McBer.

Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Koon, J. & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education, 66*, 61-81.

L'Hommedieu, R., Menges, R. & Brinko, K. (1997). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*, 232-240.

Lance, C. E. (1994). Test of a latent structure of performance ratings derived from Wherry's (1952) theory of ratings. *Journal of Management, 20*, 757-771.

Levinson, D. J., Darrow, C. N., Klein, E. B., Levinson, M. A. & McKee, B. (1978). *The seasons of a man's life*. NY: Ballantine Books.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Psychology, 52*, 77-95.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

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

Marsh, H. W. (1993). Multidimensional students' evaluations of teaching effectiveness. *Journal of Higher Education, 64*, 1-18.

Marsh, H. W., & Dunkin, M. (1992). *Students' evaluations of university teaching: Handbook on theory and research* (Vol. 8, pp. 143-234). NY: Agathon Press.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187-1197.

Maslow, A. H. (1954). *Motivation and personality*. NY: McGraw Hill.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384-397.

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82*, 189-200.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

Morris, T. L., Gorham, J., Stanley, H. C., & Huffman, D. (1996). Fashion in the classroom: Effects of attire on student perceptions of instructors in college classes. *Communication Education, 45*, 135-147.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and students instructional ratings in six types of university courses. *Journal of Educational Psychology, 8*, 250-261.

Murphy, K. R. (1989). Dimensions of job performance. In R. F. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives* (pp. 218-247), NY: Praeger.

Murphy, K. R. (1991). Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy. *Organizational Behavior and Human Decision Processes, 50*, 45-50.

Murphy, K. R., Cleveland, J. N. (1995).

Understanding performance appraisal: Social, Organizational, and goal-based perspectives. Thousand Oaks, CA: Sage Publications.

Murphy, K. R., Garcia, M, Kerkar, S, Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*, 320-325.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). NY: McGraw-Hill.

Pfeffer, J. & Fong, C. T. (2002). The end of business schools? Less success than meets the eye. *Academy of Management Learning and Education, 1*, 78-95.

Powell, P.W. (1977). Grades, learning, and student evaluation of instructors. *Research in Higher Education, 7*, 193-205.

Preskill, H. (2000). Coming around again: renewing our commitment to teaching evaluation. *American Journal of Evaluation, 21*, 103-104.

Rindermann, H. & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education, 42*, 377-399.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 224-271). Beverly Hills, CA: Sage.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. (ERIC Reproduction Service No. ED39824).

Scullen, S. E., Mount, M. K. & Judge, T. A. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88*, 50-66.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 46*, A40.

Sherman, B. R., & Blackburn, R. T. (1975). Personal characteristics and teaching effectiveness of college faculty. *Journal of Educational Psychology, 67*, 124-131.

Simpson, P. M. & Siguaw, J. A. (2000). Student Evaluations of teaching: an exploratory study of the faculty response. *Journal of Marketing Education, 22*, 199-213.

Snyder, C. R., & Clair, M. (1976). Effects of expected and obtained grades on teacher evaluation and attribution of performance. *Journal of Educational Psychology, 68*, 75-82.

Stumpf, S. A. & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71*, 293-302.

Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637-661.

Viswesvaran, C., Ones, D. & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Wiesenfeld, K. (1996). Making the Grade. *Newsweek Magazine*, June 17, p.16.

Wohlers, A. J. & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42*, 235-261.

Worthington, A. G., Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71*, 764-775.