Student Evaluations of Teaching Effectiveness and the Leniency Hypothesis: A Literature Review

Steven E. Gump

University of Illinois, Urbana-Champaign

This review presents an overview of selected articles on the leniency hypothesis: the idea that students give higher evaluations to instructors who grade more leniently. Such articles comprise a small subset of the voluminous research on student evaluations of teaching (SETs). In this diverse literature, research methods and aims have frequently affected the outcomes and conclusions, since SETs are typically context-specific instruments whose results, in isolated instances, do not generalize well. Thus this review questions the very generalizability of the massive and often contradictory SET-related literature on the leniency hypothesis and argues that future research must be designed and carried out in light of the implicit problems existing in the majority of earlier studies.

Background and Introduction to the Review

Although formal instructor evaluations at the postsecondary level are also carried out by peers, administrators, and the individual instructors themselves (as self-evaluations) (McGee, 1995), evaluations by students are the most ubiquitous. This ubiquity has resulted, according to Marsh (1984, p. 749), in student evaluations of college teaching effectiveness being "probably . . . the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research." Indeed, Centra (2003) reported that a query to the ERIC database returned references to over 2,000 studies on student evaluations of teaching (SETs). Although dozens of new studies continue to inundate the

literature every year, does the perennial popularity of this topic necessarily imply that all the scholarship and energy devoted to understanding student evaluations better is well articulated, well grounded, and worthwhile? Considering this question from an historical perspective, for example, Marsh and Dunkin (1992) argued that numerous methodologically unsound studies on SETs were published in the 1970s, a period which Centra (1993, p. 9) termed the "golden age of research on student evaluations." Marsh and Roche (1997, p. 1190) virtually condemned much of the literature base as nearly worthless: "The voluminous literature on potential biases in SETs is frequently atheoretical, methodologically flawed, and not based on well-articulated operational definitions of bias, thus continuing to fuel (and be fueled or fooled by) SET myths."

This review considers one small sub-theme in the literature related to student evaluations of college teaching effectiveness: the so-called leniency hypothesis, which attests that "instructors with more lenient grading standards receive more favourable ratings" (Wachtel, 1998, p. 200). Underscoring this hypothesis is the notion that instructors can "buy" better evaluations by giving out higher grades (Nimmer & Stone, 1991, p. 196). That notion, in turn, stems from what Petress (1996, p. 387) has termed the "quasi economic model" of education, wherein students are viewed as customers, consumers, or clients (see, for example, Bowen, 2001; Colby, Ehrlich, Beaumont, Rosner, & Stephens, 2000; Dowd, 2003; Eiszler, 2002; Muller, 1994; Stimpson, 2004). Interestingly, such a conception is not new. In his 1888 inaugural address, Francis L. Patton, the twelfth president of the College of New Jersey (later Princeton University), remarked that "college administration is a business in which Trustees are partners, professors the salesmen and students the customers" (Wertenbaker, 1946, p. 347). According to this rhetoric, then, which "ignores pedagogic reality" (Brookfield, 1995, p. 21), SETs have become "a form of customer satisfaction survey" (Eiszler, p. 499). And in (common) situations where SET ratings are used by administrators to determine merit raises or to support promotion and tenure decisions, faculty are, justifiably,

quite concerned about their teaching evaluations.

Uses, Methods, and Analyses of Student Evaluations of Teaching Effectiveness

Almost as soon as student evaluation procedures were introduced at several major universities in the U.S. in the 1920s (Marsh, 1987), research on SETs and the various factors that may affect or otherwise bias them appeared (e.g., Brandenburg & Remmers, 1927; Remmers, 1928, 1930; Remmers & Brandebnurg, 1927). Over the past seven decades, then, ratings have been shown to be of potential value to several constituents in addition to administrators, who may also use SET ratings for the identification of excellence in teaching (or for identification of instructors who could use help developing necessary classroom skills). SETs are of potential value to instructors for the improvement of instruction, to students for the selection of courses or instructors (when ratings are published), to researchers as a source for data for research on teaching, and to other interested and invested parties for addressing issues of accountability.

Surveys are virtually uncontested as the chief instruments for collecting student evaluations of teaching effectiveness; the acronym "SET," in fact, seems to be taken as synonymous with paper evaluations in survey form. The inherent quantitative nature of survey data, which, on an institution-wide scale, are more easily analyzed and compared than would be similar volumes of qualitative data, has perpetuated the mindset of such measurements as the most effective means for assessing student views. But both Sommer (1981) and Millea and Grimes (2002), however, recommended a multi-method approach for gathering student evaluations of teaching—an approach that would enlarge the conventional insinuation of the term "SET." And Greenwald and Gillmore (1997, p. 1215) admitted that expert appraisals of instructor effectiveness "might provide more valid assessments" (than traditional SETs)—but that that alternative "greatly

exceed[s] student ratings in cost."

Empirical studies of SETs regularly focus on one of five areas (as identified in Wachtel, 1998): (1) characteristics related to the administration of evaluations (e.g., anonymity of ratings, timing of evaluations, presence of the instructor during evaluation); (2) course characteristics (e.g., class size, selectivity); (3) instructor characteristics (e.g., gender, reputation); (4) student characteristics (e.g., age, expectations, prior subject interest); and (5) reactions to the use of evaluations (e.g., by faculty or students). In some research on SETs, experimental and quasi-experimental methods have been used to investigate correlations between some of the various normative characteristics identified by Wachtel (independent variables) and the quantitative evaluations (dependent variables). More common in the literature, though, appear to be studies on SETs that were based on "nonexperimental" research (Nimmer & Stone, 1991, p. 198), where conclusions are drawn from passive observation of aggregated data. Very rarely have qualitative methodologies been appropriated for the purposes of gaining a deeper understanding of SETs.

Supporting (or Contesting) the Leniency Hypothesis

According to Wachtel's (1998) typology, studies addressing the leniency hypothesis are a subgroup of studies of student characteristics that specifically consider grade expectations. The literature has addressed the leniency hypothesis from at least two opposing perspectives and through association with a number of related issues. First, researchers have typically approached the issue either in acceptance or with skepticism; and this "experimenter bias" (Centra, 2003, p. 497), while frequently inferable from the conclusions of articles cited in the introduction or literature review, is sometimes directly expressed at the outset of an article, as in Greenwald and Gillmore (1997, p. 1209), who began their presentation by stating unequivocally that "it is well established that students' evaluative ratings of instruction correlate positively with expected course grades." In fact, Greenwald and Gillmore were so convinced of the existence of

leniency bias in grading that they routinely (and uniquely) referred not to the "leniency hypothesis" but to the "leniency theory." Experimenter bias is also an issue when the researcher is the instructor in a class whose SETs are used as data (Centra, 2003). Generally, opponents of the administrative use of SETs have mobilized correlation with grades as an argument to the invalidity of SETs. Furthermore, faculty members-as-researchers may have attempted, through their research, to vindicate their own lower-than-average scores on SETs by "proving" that their ratings are simply a "result" of their higher standards in grading. Such faculty members generally have little faith in the concept of SETs; yet they might support a more digestible hypothesis that explains high SET ratings as follows: Students have learned more, earned higher grades, and thus provided higher evaluations (Eiszler, 2002; Nimmer & Stone, 1991). On the other hand, when coupled with low grades, low SET ratings can be interpreted to imply not that the instructor has higher standards but rather that the students have learned less. These underlying perspectives have complicated the literature; articles that have been forthright about investigator or author biases seem surprisingly rare.

Perhaps due in part to these biases, the literature on teaching evaluation is, in toto, "ambiguous" and "contradictory" (Sommer, 1981, p. 223) and offers "mixed results" (Nimmer & Stone, 1991, p. 196) that can be "subject to multiple interpretations" (Sojka, Gupta, & Deeter-Schmelz, 2002, p. 44). Researchers may (and frequently seem to) pick and choose from previous studies, presenting results that support their hypotheses or rationalize the needs for their particular studies. For example, Greenwald (1997, p. 1183) turned a well-nigh blind eye to research that was clearly in opposition to his thesis, stating that "the hypothesis that grading leniency–strictness affects ratings . . . has been supported with some clarity in virtually all published experimental texts." Alternatively, what has often appeared to be eclecticism on the part of the researcher may simply be a consequence of the unwieldy number of studies that have been carried out: most likely, only a manageable sampling of the

related research was consulted in the preparation of any study.

Second, studies that consider the leniency hypothesis often have done so only in part or even tangentially. A few of the main articles selected for this review, for example, approached the leniency hypothesis through its association with other issues, notably grade inflation, student learning outcomes, and instructor or administrator uses of SET results. A recent study published in this journal, in fact, yielded results that support the leniency hypothesis, although the main focus of the research was to investigate class (and classroom) sizes with respect to SETs (Safer, Farmer, Segalla, & Elhoubi, 2005). Inspiring and perpetuating what may be the overproduction of mediocre studies—and reflecting mixed conclusions when their works are situated within the body of previous literature, the authors of practically all articles consulted for this review stated that additional research on SETs is warranted.

How serious is this lack of consensus within the literature? Of the major articles selected for this literature review, three presented results that do not support the leniency hypothesis (Centra, 2003; Nasser & Fresko, 2002; Sommer, 1981), and five presented results that support it (Eizsler, 2002; Greenwald & Gillmore, 1997; Millea & Grimes, 2002; Nimmer & Stone, 1991; Sojka et al., 2002). Taken collectively as a sampling of the entire corpus of literature, these articles, which incorporated various quantitative and qualitative methodologies, embody the ambiguity of the research on SETs. Some of this ambiguity may, in fact, be due to methodological weaknesses in some of the studies (as previously suggested by Marsh and Roche, 1997), especially those drawn from passive observation of, typically, the investigators' own SET data. Often the desire seemed to be to generalize based on results from samples that were not representative of the populations for which the generalizations were being constructed and applied. Indeed, that the research on SETs has remained so inconclusive supports the following statement by Campbell and Stanley (1963, p. 17): "the problems of external validity are not logically solvable in any neat, conclusive way."

Student Evaluations of Teaching Effectiveness and Perceptions of Validity

Just as there has been a lack of consensus among studies that have considered the leniency hypothesis, there has also been a lack of consensus involving both student and faculty *perceptions* of the validity of SETs. Nasser and Fresko (2002) and Sojka et al. (2002) specifically focused on perceptions of SET validity, an important issue in light of the fact that instructors would need to accept the leniency hypothesis, for the most part, in order to grade more easily (or to lower their expectations) in attempts to raise their evaluations. Marsh (1987) had previously reported that 68% of faculty members at a major research university, when asked, believed grading leniency is a likely bias of student evaluations.

Nasser and Fresko (2002) and Sojka et al. (2002) couched their studies of perceptions of SETs in awareness of the ongoing debate in the literature over the psychometric quality (reliability and validity) of SETs. Nasser and Fresko, who analyzed the results of questionnaires returned by 101 instructors (out of a queried sample of 447) at a school of education in Israel, found a lack of consensus among the faculty with respect to perceptions of the validity of SETs. This lack of consensus, fittingly, mirrors what has been reported in the literature as a lack of consensus among researchers on the validity of SETs. Since Nasser and Fresko found that changes to grading procedures were rarely made as a result of SETs, however, their research did not support the leniency hypothesis. Instead, their discussion addressed the ramifications of faculty rejection of the hypothesis with respect to faculty members' own instructional approaches.

Sojka et al. (2002) also used survey methodology to consider ramifications of the leniency hypothesis for faculty; but they found, contrary to Nasser and Fresko (2002), at least, that the faculty members in their sample assumed that grading more leniently would lead to higher SET ratings. Unfortunately, however, the Sojka et al. study is methodologically limited, thus fitting the stereotype described by Marsh and Roche (1997). The

report of the study, based on a survey of students and faculty at a mid-sized university in the Midwestern United States, provides no information on the method of selection, the number or nature of questions on the survey, or student response rates. If, as it appears, a convenience sample was used, this important detail should not have been omitted from the description of the methodology, in order to allow for better interpretation and application of the results.

Millea and Grimes (2002) also based their support of the leniency hypothesis on a surveyed convenience sample (of 149 undergraduates in an economics class); but they were at least more thorough than Sojka et al. (2002) in presenting details about the sample population. Indeed, the restricted nature of the samples used by Nasser and Fresko (2002), Sojka et al., and Millea and Grimes raises questions of generalizability (à la Campbell & Stanley, 1963). In all three cases, the authors' results were certainly applicable to the populations at the schools from where the samples were drawn; but the extent to which their findings would be replicated elsewhere is largely a function of the sizes, locations, and student and faculty bodies at the schools. Thus the implications and conclusions that assume applicability to all instructors or to any group of students must be approached with care. Of the studies examined for this review, for example, only Nimmer and Stone (1991), whose study used a true experimental design (i.e., was a carefully controlled laboratory experiment) to test the leniency hypothesis as well as the corollary of learning effects, extended a concerted effort to mention difficulties underlying the assumption of generalizability of their findings to other populations.

Studies of Student Evaluations of Teaching: Evaluating an Alternative Approach

Research presented by Sommer (1981) integrated a methodology uncommon in studies of SETs—autobiography—to describe the ways in which "one instructor's teaching evaluations over a 20-year period" (1961 to 1981) "affected the form and content" of the instructor's teaching (p. 224). Sommer offered a

"personal experience story" (as described by Denzin, 1994, p. 541) to address what Nasser and Fresko (2002) and others have attempted to do via surveys. But although the grounds for which interpretations of survey results are reached are not always presented in explicit detail, Sommer's personal case study offered rich details on how both his teaching and experiences as an administrator had been affected by his personal SETs and the evaluations of others. As the subject of the narrative, on one level (Denzin, p. 541), Sommer offered the reader a glimpse into his personal life story, presenting what he had learned, had felt, and had come to believe about SETs over the course of twenty years. In fact, the conclusions presented by Sommer were framed as personal judgments that can be viewed as expert analysis (since the "subject-as-author is given an authority over the life that is written about," Denzin, p. 541). The reader is enlightened about how and why Sommer reached the conclusions he did, and, ultimately, how and why he responded to the evaluations of his teaching in the ways he did. Sommer did not explicitly state that the low evaluations he occasionally received ever affected the method in which he graded his students. Instead, in response to SETs, he adjusted his teaching style or modified the course content covered.

Although Sommer (1981) used an idiographic approach (which, as described by Allport, 1961, involves the study of individuals—and is the opposite of a nomothetic approach, which involves the comparison of groups), his presentation was nonetheless related to actual course evaluations accumulated in the "field" (as opposed to a laboratory setting). The use of "field" data is typical of most studies on SETs that involve passive observation of data. Centra (2003), Eiszler (2002), Greenwald and Gillmore (1997), and Millea and Grimes (2002) all analyzed "authentic" SET data in their studies, with all but Millea and Grimes basing their findings on aggregations of vast quantities of historical data. (Centra, for example, analyzed data from approximately 55,000 classes over a five-year period, and Eiszler analyzed data from more than 37,000 courses over a twenty-year

period. Although impressive in scope, however, post-hoc analyses of such large samples can encounter problems with statistical power.)

Even when methodologies may be similar, differences in definitions of key terms (e.g., "bias," "workload") or in the questions asked on various SET instruments complicate comparisons across studies. The substantial data analyzed by Centra (2003), for example, was based on SETs that asked students to use a 5-point Likert-type scale to rate the "quality of instruction as it contributed to their learning" (p. 501), not simply to rate the teacher or the course, as is typical of most other SET forms. (The 5-point Likert-type scale, too, seems to be a taken-for-granted characteristic of SETs.) Thus the key variable of student learning, addressed also by Nimmer and Stone (1991), complicates the relationship between grading leniency and SETs. Indeed, a perennial problem affecting studies on SETs appears to be the difficulty with which many studies have in unpacking or untangling what could be effects of the leniency hypothesis from other variables affecting SETs.

Fortunately, authors of some of the more methodologically sound studies (Centra, 2003; Nimmer & Stone, 1991) were aware of (and made necessary adjustments for) the various interactions and complications. For example, and contrary to Sommer's (1981, p. 224) comment that "experimentation, including random assignment, is not feasible" with studies of SETs, Nimmer and Stone carried out two laboratory experiments on undergraduate psychology students, using a between-subjects, completely randomized factorial design involving a videotaped lecture that was viewed by all participants in order to examine relationships between grading practices and SETs while controlling for learning effects. They found that student ratings were directly affected by grading practices—but that the particular effect was, in turn, a function of the timing of the evaluations with respect to the students' knowledge or anticipation of their grades.

Implications and Suggestions for the Future

Although studies considering the leniency hypothesis by nature offer implications for college and university instructors, another common theme among studies of SETs, regardless of methodology, has been the consideration of the "serious" implications for administrative and institutional use of such ratings (Nasser & Fresko, 2002, p. 189). Sommer (1981), for example, spoke to these implications from personal experience, recommending that SETs be used routinely only for the "*selection* and *retention* of nontenured faculty and the *assignment* of tenured faculty" or longitudinally for signs of "incipient burnout" (p. 226, emphases in original). And Eiszler's (2002) study linked the use of SET data in personnel decisions to grade inflation (thus assuming, a priori, an underlying relationship between faculty desire for higher evaluations and the higher grades that are given in expectation of—or exchange for—them).

In short, the literature on the leniency hypothesis has succeeded in demonstrating that there is little consensus on the validity of SETs, the perceptions of this validity (or lack thereof), and the ways in which SETs should (or should not) be used by the various constituents involved. A new wave of research on SETs seems to be in order: research that looks critically yet holistically at past studies with respect to their methodologies, conclusions, and implications in an attempt to discern the extent to which context-specificity renders the results of studies applicable to little more than the populations on which they were based. It is possible that instances of the leniency hypothesis, if it is to be accepted, are so closely entangled with (for example) specific academic disciplines, course difficulty levels, and individual instructors' personalities and idiosyncrasies that teasing out conclusions in support of any general hypothesis or theory becomes an interpretive art that is far removed from precise and methodologically rigorous educational experimentation. Instructors should continue to use the results of their own SETs to reflect on and improve their own teaching effectiveness; but they should not depend on the literature to reveal uncontested and foolproof panaceas, such as grading more leniently, for

attempting to improve their ratings.

References

Allport, G. (1961). *Pattern and growth in personality.* New York: Holt, Rinehart, & Winston.

Bowen, R. W. (2001, June 22). The new battle between political and academic cultures. *Chronicle of Higher Education*, p. B14.

Brandenburg, G. C., & Remmers, H. H. (1927). A rating scale for instructors. *Educational Administration and Supervision, 13*, 399–406.

Brookfield, S. D. (1995). *Becoming a critically reflective teacher.* San Francisco: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco: Jossey-Bass.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*, 495–518.

Colby, A., Ehrlich, T., Beaumont, E., Rosner, J., & Stephens, J. (2000). Introduction: Higher education and the development of civic responsibility. In T. Ehrlich (Ed.), *Civic responsibility and higher education* (pp. xxi–xliii). Phoenix, AZ: Oryx Press.

Denzin, N. (1994). Biographical research methods. In *The international encyclopedia of education* (2nd ed.) (Vol. 1, pp. 537–543). New York: Pergamon.

Dowd, A. C. (2003). From access to outcome reality: Revitalizing the democratic mission of the community college. In K. M. Shaw & J. A. Jacobs (Eds.), *Community colleges: New environments, new directions* (pp. 92–119). Thousand Oaks, CA: Sage.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*, 483–501.

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

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

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. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143–233). New York: Agathon.

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.

McGee, R. (1995). Faculty evaluation procedures in 11 western community colleges. *Community College Journal of Research and Practice, 19*, 341–348.

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36*, 582–590.

Muller, S. (1994). Presidential leadership. In J. R. Cole, E. G. Barber, & S. R. Graubard (Eds.), *The research university in a time of discontent* (pp. 153–178). Baltimore: Johns Hopkins University Press.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation*

in Higher Education, 27, 187–198.

Nimmer, J. G., & Stone, E. F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning. *Research in Higher Education, 32*, 195–215.

Petress, K. C. (1996). The dilemma of university undergraduate student attendance policies: To require class attendance or not. *College Student Journal, 30*, 387–389.

Remmers, H. H. (1928). The relationship between students' marks and students' attitudes toward instructors. *School and Society, 28*, 759–760.

Remmers, H. H. (1930). To what extent do grades influence student ratings of instructors? *Journal of Educational Psychology, 21*, 314–316.

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue ratings scale for instructors. *Educational Administration and Supervision, 13*, 519–527.

Safer, A. M., Farmer, L. S. J., Segalla, A., & Elhoubi, A. F. (2005). Does the distance from the teacher influence student evaluations? *Educational Research Quarterly, 28*(3), 28–35.

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching, 50*, 44–49.

Sommer, R. (1981). Twenty years of teaching evaluations: One instructor's experience. *Teaching of Psychology, 8*, 223–226.

Stimpson, C. R. (2004, June 18). Reclaiming the mission of graduate education. *Chronicle of Higher Education*, p. B6.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*, 191–211.

Wertenbaker, T. J. (1946). *Princeton, 1746–1896.* Princeton, NJ: Princeton University Press.