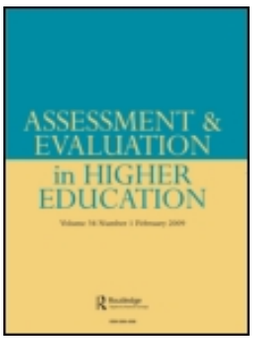

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Student Evaluation of College Teaching Effectiveness: a brief review

Howard K. Wachtel

To cite this article: Howard K. Wachtel (1998) Student Evaluation of College Teaching Effectiveness: a brief review, Assessment & Evaluation in Higher Education, 23:2, 191-212, DOI: 10.1080/0260293980230207

To link to this article: https://doi.org/10.1080/0260293980230207

Student Evaluation of College Teaching Effectiveness: a brief review

HOWARD K. WACHTEL, *Department of Science & Mathematics, Bowie State University, Maryland, USA*

ABSTRACT *This paper presents a brief review of the existing research on student written evaluations of the teaching performance of college and university instructors. First, a short historical background is given. Next, arguments are presented which have been advanced for and against the use of student written evaluations as a valid indicator of teaching effectiveness, followed by a discussion of various background characteristics which have been thought to influence student ratings. Student and faculty reaction to the use of student ratings is discussed, and finally suggestions for further investigation are given.*

Introduction

This paper attempts to present a brief review of the existing research on student written evaluations of the teaching performance of college faculty. There is an enormous amount of literature on student evaluations of instruction, literally thousands of papers according to Marsh and Dunkin (1992). Cashin (1988) writes that there are probably more studies of student evaluations than of all the other means used to evaluate college teaching combined.

Historical Background

Research on student evaluations of teaching and the factors which may affect them dates back to the 1920s and the pioneering work of Remmers (1928, 1930; Brandenburg & Remmers, 1927; Remmers & Brandenburg, 1927). In his series, Remmers confronted some of the major issues in the area of student evaluation research, such as whether the judgments of students agree with those of peers and alumni. Spencer and Flyr (1992) report that the first teacher rating scale was published in 1915; Marsh (1987) notes that student evaluation procedures were introduced at several major US universities in the

0260-2938/98/020191-21

1920s. Centra (1993, p. 49) divides the period of student evaluation research into four smaller periods as follows:

(1) The period from 1927 to 1960, which was dominated by Remmers and his colleagues at Purdue University.
(2) The 1960s, in which use of student evaluations was almost entirely voluntary.
(3) The 1970s, which he calls the 'golden age of research on student evaluations'.

During this decade, Centra states, a new wave of research ensued, including studies which demonstrated the validity and utility of student ratings and supported their use for both formative and summative purposes.

(4) The period from the early 1980s to the present day, during which time followed continued clarification and amplification of research findings, including meta-analyses which synthesised the results of many other studies.

Marsh and Dunkin (1992) contend that many methodologically unsound studies on students' evaluations of teaching were accepted for publication in the 1970s, while in the 1980s there were fewer, better quality articles published in major research journals. Marsh (1984) states that the validity of student ratings has been sufficiently well established that the focus of student evaluation research has shifted more recently to methodological concerns and the study of specific background characteristics which might harm validity.

Support for and Opposition to the Use of Student Evaluations

Faculty opinion on the utility of student evaluations of teaching ranges from 'reliable, valid, and useful' to 'unreliable, invalid, and useless' (Aleamoni, 1981). However, after nearly seven decades of research on the use of student evaluations of teaching effectiveness, it can safely be stated that the majority of researchers believe that student ratings are a valid, reliable, and worthwhile means of evaluating teaching (for example, Centra, 1977, 1993; Cohen, 1981; Koon & Murray, 1995; Marsh, 1984; 1987; Marsh & Dunkin, 1992; McKeachie, 1990; Murray *et al.*, 1990; Ramsden, 1991; Seldin, 1984; 1993). In fact, Marsh (1987) contends that student evaluations are the *only* indicator of teaching effectiveness whose validity has been thoroughly and rigorously established. Further arguments supporting the use of student ratings include:

- Feedback from student ratings can help to improve instruction (Cohen, 1980; Marsh & Roche, 1993; Menges, 1991; Overall & Marsh, 1979). However, we note that some authors who are supportive of the use of student ratings nonetheless argue that they alone will not automatically improve teaching and sustain that improvement without other types of feedback (Seldin, 1989, 1993; Tiberius *et al.*, 1989; Wilson, 1986). L'Hommedieu *et al.*, (1990) argue that methodological weaknesses in existing studies generally attenuate rather than exaggerate effects of feedback, i.e. the effect of feedback on teaching improvement may be even greater than that posited in the literature.
- The use of student ratings increases the likelihood that excellence in teaching will be recognised and rewarded (Aleamoni, 1981; McKeachie, 1979).
- Student ratings have been shown to be positively correlated with student learning and achievement, i.e. students rate most highly those instructors from whom they have learned the most (Aleamoni & Hexner, 1980; Centra, 1977; Cohen, 1981; McKeachie,

1990; Murray, *et al.*, 1990). Nonetheless, Derry (1979) and McCallum (1984) state that critics of student ratings cite the fact that these correlations are only moderate (or widely varying) in arguing against their validity.

- Students and faculty generally agree on what are the components of effective teaching and their relative importance (Feldman, 1976b, 1988). This is used to counter the view that students cannot accurately evaluate teaching because students and faculty cannot agree on what constitutes good teaching (Marsh & Dunkin, 1992, p. 181).
- Student ratings are positively correlated with ratings by alumni (Centra, 1974, 1979; Feldman, 1989; Howard *et al.*, 1985; McKeachie, 1979; Overall & Marsh, 1980). This runs counter to the argument by critics of student ratings that long after graduation, students with the benefit of additional years of wisdom, will hold a different view of an instructor (particularly one who is demanding) than at the time they were enrolled in that instructor's course, a view which is still commonly held (Wachtel, 1994, p. 86). However, it is noted that Braskamp & Ory (1994) feel that it maybe useful to collect alumni ratings anyway.

In spite of the extensive research supporting the validity of student evaluations of teaching, many writers still express reservations about their use (particularly for personnel and tenure decisions) or even oppose them outright (for example, Chandler, 1978; Dowell & Neal, 1982; Goldman, 1993; Hills, 1974; Koblitz, 1990; Menefee, 1983; Miller, 1984; Powell, 1978; Rutland, 1990; Sheehan, 1975; Small *et al.*, 1982; Vasta & Sarmiento, 1979; Zoller, 1992). In one extreme case the mathematics department at one university refused to participate in that university's student evaluation programme (Heller, 1984).

Furthermore, there is abundant anecdotal evidence of faculty hostility and cynicism toward the use of student ratings (Franklin & Theall, 1989). An example of such cynicism is the viewpoint that much of the research supporting the validity of student evaluations comes from those who developed rating scales and offer their services for a fee (Powell, 1978). Some of the caveats which have been advanced regarding the use of student evaluations include: (We emphasise that not all of the authors mentioned in the following discussion are against the use of student ratings.)

- The idea that evaluation presumes a consensus which does not yet exist. That is, how can we evaluate teaching effectiveness adequately if we cannot even agree on what constitutes effective teaching? (Marques *et al.*, 1979; Meeth, 1976; Monroe & Borzi, 1989; Spencer, 1992)
- Teaching is art and feelings, involves nurturing and similar qualities which cannot be easily assessed by evaluation instruments (Ornstein, 1990).
- Faculty may resent the loss of class time for the administration of the evaluation forms, and may be less motivated to experiment in their instructional methods (Centra, 1993, p. 93).
- Faculty and administrators have little knowledge of existing research in this area (Franklin & Theall, 1989) and therefore may administer the evaluations improperly or engage in some kind of abuse (for example, according too much significance to the last decimal place in a class average score).
- The use of student evaluations of teaching reduces faculty morale and job satisfaction. This is discussed below under faculty reaction.
- Teaching so as to promote favorable evaluations may be at odds with good educational practice (Baxter, 1991, p. 156; Zelby, 1974). Faculty may tend to reduce

standards and/or course workloads as a result of mandatory evaluation (Ryan *et al.*, 1980). (See also discussion of leniency hypothesis below.)
- Instructors who share students' attitudes, or appear to do so, have an advantage in procuring higher ratings (Hofman & Kremer, 1980).
- Many student evaluation instruments contain inappropriate items. Tagomori and Bishop (1995) found that the vast majority of the instruments collected in their survey contained evaluation items which were ambiguous, vague, subjectively stated, or did not correlate with classroom teaching behaviour.
- Ratings may be affected by one or more characteristics which have nothing to do with the instructor's behaviour or effective teaching. This is discussed below under the heading of background variables.

Background Variables Thought to Influence Student Ratings

This section considers the existing knowledge on characteristics other than instructional effectiveness which could conceivably exert an influence on student evaluation results. Some authors, including Brandenburg *et al.* (1977) and Tatro (1995), have argued that background variables can produce significant shifts in instructor ratings. It must be noted here that the mere existence of a correlation between a background variable and rating scores does not necessarily constitute a bias or a threat to the validity of student ratings. For example, if the student's expected grade in the course is found to be associated with the rating which that student gives to the instructor, it does not necessarily follow that an instructor can obtain higher ratings merely by giving higher grades. Alternative explanations include the possibility that more effective teaching will inspire students to work harder and earn better grades. Careful research is necessary in order to determine the nature of the relationship between the background variable and student ratings.

Characteristics Associated with the Administration of Student Evaluations

Seldin (1993) warns against invalidating a good evaluation form with poor administration procedures, timing or instructions. Such administrative characteristics are considered in the next sections.

Timing of Evaluation

Feldman (1979) found that the time at which student evaluations are administered, whether in the middle of the course, or before, during, or after the final examination, has no effect on the results. Frey (1976) found that ratings collected during the last week of classes were not significantly different from those collected in the first week of the succeeding term, but nonetheless is concerned about the effect that the thought of an impending final examination may have on ratings. Marsh and Overall (1980) found that mid-term ratings were highly correlated with end-of-term ratings, but L'Hommedieu *et al.* (1990) argue for the non-equivalence of mid-term and end-of-term ratings and call for further research into the effect of timing on student evaluations. Aleamoni (1981) asserts that results may be distorted if administered before or after an exam. In our experience it is not uncommon for a department to require that the evaluations be administered on a given day or week, which could be unfair to instructors who are giving an examination during that week. Braskamp *et al.* (1984) contend that ratings may be lower if administered during the final examination, and recommend the last two weeks of the

term (but not the last day of class) for administration of the evaluations. Seldin (1989) agrees with this suggestion when ratings are to be used summatively, but maintains that appraisal should take place early in the term, about four to six weeks after the course begins, when ratings are to be used formatively.

Anonymity of Student Raters

Feldman (1979) and Blunt (1991) report that students tend to give somewhat higher ratings when they identify themselves compared to those when they remain anonymous. However, Feldman states there are other circumstances which may interact with anonymity, such as whether the ratings are given before or after the students know their grades, whether the ratings are done in "special experimental sessions", whether the students are told that the ratings will be used for research purposes only, and whether the students believe that there is a possibility of a 'confrontation' with the instructor (Abrami *et al.* 1976). Blunt (1991) also expresses concern as to whether or not students feel that they can trust faculty and administration assurances of anonymity and confidentiality. Most authors (for example, Braskamp *et al.* 1984; Centra, 1993; McCallum, 1984) recommend that student raters remain anonymous.

Instructor Presence in Classroom

Feldman (1979) also reports that ratings are somewhat higher when the instructor being evaluated is present in the room while the evaluation forms are being completed. Most authors, including Braskamp and Ory (1994), Centra (1993, p. 78), Eble (1970, p. 24) and Scriven (1981, p. 255), recommend the use of a third party in charge of collecting forms along with assurances that the student's anonymity is protected. Another consideration is the possibility that there may be a latent effect on the ratings by having the professor being evaluated pass out the evaluation forms, even if (s)he is not the person collecting the forms. Pulich (1984) contends that even if the instructor leaves the room while the students are filling out the forms, some students may still be inhibited by the fact that the instructor him/herself distributed them. She therefore suggests that a separate evaluator come to distribute and collect the forms, as well as answer any questions which the students may have during the evaluation. The instructor could thus be absent during the entire process.

Stated Purpose of Evaluation

Although some studies have found that student ratings are somewhat higher if the stated purpose is for promotion and tenure (Aleamoni & Hexner, 1980; Braskamp *et al.*, 1984; Centra, 1976; Feldman, 1979), Frankhouser (1984) concluded that the stated purpose of the evaluation had no significant effect on ratings. Braskamp *et al.* (1984) recommend that students be told if the ratings are to be used for personnel decisions.

Characteristics of the Course

Electivity

Researchers have found that teachers of elective or non-required courses receive higher ratings than teachers of required courses. More specifically, the 'electivity' of a class can

be defined as the percentage of students in that class who are taking it as an elective (Feldman, 1978); a small to moderate positive relationship has been found between electivity of a class and ratings (Brandenburg *et al.*, 1977; Feldman, 1978; McKeachie, 1979; Scherr & Scherr, 1990). This may be due to lower prior subject interest in required versus non-required courses.

Class Meeting Time

From the limited amount of research on this topic, the consensus is that no consistent relationship exists between ratings and the time of day at which a class meets (Aleamoni, 1987; Centra, 1993; Feldman, 1978). An exception to this is a study by Koushki and Kuhn (1982) which found that very early morning classes, very late afternoon classes, and classes shortly after lunch receive the lowest ratings. The effect of the time of day found in this study was greater than that of other background variables such as gender, year in school, field of study, and expected grade.

Level of Course

Most studies have found that higher level courses tend to receive higher ratings (Feldman, 1978; Marsh, 1987, p. 324). However, no explanation for this relationship has been put forth, and Feldman also reports that the association between course level and ratings is diminished when other background variables such as class size, expected grade and electivity are controlled for. Therefore the effect of course level on ratings maybe direct, indirect, or both. Also, surprisingly the literature has generally neglected another factor which may be relevant here; namely, the age of the students. It is conceivable that the difference in the average age and/or maturity of the students at the time the ratings are administered might be a greater cause of the effect on ratings than the properties of the course itself.

Class Size

Considerable attention has been paid to the relationship between class size and student ratings. Most authors report that smaller classes tend to receive higher ratings (Feldman, 1978; Franklin *et al.*, 1991; McKeachie, 1990). Marsh (1987, p. 314; Marsh & Dunkin, 1992) reports that the class size effect is specific to certain dimensions of effective teaching, namely group interaction and instructional rapport. He further argues that this specificity combined with similar findings for faculty self-evaluations indicated that class size is not a 'bias' to student ratings (see also Cashin, 1992). However, Abrami (1989b) in his review of Marsh's (1987) monograph counters that this argument cannot be used to support the validity of ratings, and instead demonstrates that interaction and rapport, being sensitive to class size, are dimensions which should not be used in summative decisions. Scott (1977) found that instructors who felt that their class was too large for them to exhibit the course material adequately obtained lower ratings than other instructors. This suggests the possibility that instructors' feelings about class size may have some effect on their performance and ratings (Feldman, 1978).

Another hypothesis is that the relationship between class size and student ratings is not a linear one, but rather, a U-shaped or curvilinear relationship, with small and large classes receiving higher ratings than medium-sized ones (Centra & Creech, 1976; Feldman, 1978, 1984; Koushki & Kuhn, 1982). (Feldman also suggests that even more

researchers may have found a curvilinear relationship if they had been aware of this possibility.) Some explanations which have been offered for this relationship (unexpectedly higher ratings for very large classes) include: Departments may assign known superior teachers to large lecture classes; superior teachers may attract more students to their classes by virtue of their reputation, thereby making the class sizes larger; instructors may feel an increased challenge of teaching a large lecture course and may prepare themselves more thoroughly or tailor their pedagogy to that type of course, including the more frequent use of audio-visual aids; finally, many courses with over one hundred students often have small discussion sections with teaching assistants.

Subject Area

Researchers have found that subject matter area does indeed have an effect on student ratings (Ramsden, 1991), and furthermore, that ratings in mathematics and the sciences rank among the lowest (Cashin, 1990, 1992; Cashin & Clegg, 1987; Centra & Creech, 1976; Feldman, 1978). Ramsden (1991) feels that the differences among disciplines are sufficiently large that comparisons in student ratings should not be made across disciplines. Centra (1993) proffers some possible explanations for the lower ratings in mathematics and science courses, namely that such courses and their instructors are less student oriented, the courses are less effective in presentation and faster paced, and the faculty are required to invest more time in research and seeking grants than their colleagues in other disciplines. Centra also cites an earlier study (Centra, 1973) in which he found that students and teachers in the natural sciences have much different perceptions of appropriateness of pace and workload, which can adversely affect ratings. Cashin (1988) argues that if in fact teaching is less effective in subjects rated lower, then subject area would not represent a bias to ratings; if, however, subjects requiring quantitative reasoning skills are rated lower because students are less competent in those skills, then that would constitute a bias to ratings. Perry *et al.* (1974) note that a 'poor' teacher presenting interesting material is rated consistently higher on some dimensions of effective teaching than a 'good' teacher presenting boring material. This may provide further reason to eschew comparing teacher ratings between unrelated subject areas.

Workload of Course

Ryan *et al.* (1980) reported that the introduction of mandatory student ratings at one US Midwestern university led faculty to reduce course workloads and make examinations easier. Dudley and Shawver (1991) found that student ratings in a marketing course without homework were improved by the introduction of relevant homework assignments; however, this result may not be general to all subject areas. (Also, the marketing instructors in this study originally had rating scores below the average for the business school and the university in question; were this not the case, it is possible and perhaps even likely that the rating scores would not have improved as a result of the homework assignments. It is interesting that the rating form used in this study does not include a question such as "Was the workload for this course appropriate?") Marsh (1987, p. 316) cites studies which have found a *positive* correlation between 'workload/difficulty' and student ratings (more difficult courses were rated more favourably), and on that basis rejects this factor as a possible bias to ratings. However, course level and student age might be confounding factors in more difficult courses. Also, difficulty and workload are not entirely the same thing. For example, Franklin *et al.* (1991) treated workload and

difficulty as distinct variables; they found that difficulty, but not workload, had a slight positive correlation with instructor ratings. As noted in the previous section, student perceptions of workload and pace may differ markedly from that of the instructor. The conclusion that courses with greater workloads are rated higher may not be general to mathematics and science; research to test this hypothesis in these areas would be most helpful. It is conceivable that the *pace* of a course, or the students' perceptions thereof, may have an adverse effect on the student ratings.

A Final Reflection on Course Characteristics

A study by Cranton and Smith (1986) shed further light on the effect of course characteristics such as course level and class size on student ratings. While the study confirmed the overall trends documented in previous studies, the authors found that within individual departments the effect of course characteristics on ratings varied drastically. In some departments the course characteristic did not have a significant effect on student ratings, while in some cases the effect was in the opposite direction to that predicted. The authors conclude that one cannot establish campus-wide norms for teaching effectiveness, and even extending such norms to an entire department should be done with caution. Since the relationship of course characteristics to student ratings does not hold equally in different departments, we feel that it would be useful to conduct research on the effects of course characteristics on ratings in individual subject areas.

Characteristics of the Instructor

Instructor Rank and Experience

Not surprisingly, where ratings of professors and teaching assistants have been compared, professors are rated more highly (Brandenburg *et al.*, 1977; Centra & Creech, 1976; Marsh & Dunkin, 1992). First-year teachers receive lower ratings than those in later years (Centra, 1978). Aside from the issue of teaching assistants, Feldman (1983) reviewed the literature concerning the relationships between seniority and ratings, and found that the majority of studies concerning academic rank found no significant relationship between rank and teaching evaluations. Among those studies which did find a significant (though weak) relationship, nearly all found that instructors of higher rank received more favourable ratings. Regarding the instructor's age and experience, Feldman warns that those characteristics should not be confused with academic rank. Of studies which Feldman reviewed concerning the relationship between age/experience and ratings, again the majority of studies found no significant relationship. However, among those studies in which a significant relationship was found, nearly all found an *inverse* relationship; that is, instructors of greater age and instructional experience received *lower* ratings. Feldman maintains that the minority of studies which have found significant associations are nonetheless too many to be ignored. Centra (1993, p. 73) warns that we can draw only limited conclusions regarding the effect of rank and experience owing to the fact that most studies have been cross-sectional rather than longitudinal.

Reputation of Instructor

There is relatively little discussion of this characteristic in the literature. Perry *et al.* (1974) found that prior student expectations of teaching performance based on instructor

reputation influenced ratings. These authors argue that this result has implications for the advisability of publication of student rating information to the entire academic community. Leventhal *et al.* (1976) found that students who used instructor reputation to select class sections gave higher ratings to their teachers than did their classmates. Perry *et al.* (1979) found an interaction between instructor reputation and expressiveness, namely that highly expressive instructors with a negative reputation received lower ratings than other highly expressive instructors with more positive reputations. Instructors exhibiting a lower degree of expressiveness were not significantly affected by the reputation factor.

Research Productivity

Arguments are often advanced that research productivity helps to improve teaching effectiveness because it enables one to remain current in one's field, or that it hinders teaching effectiveness because faculty spending more time on scholarship have less time to spend on teaching. The literature on student ratings does not support either of these hypotheses; on the contrary, most studies have shown that there is no relationship or a very weak positive relationship between research productivity and teaching (Aleamoni, 1987; Centra, 1983; Feldman, 1987; Marsh, 1979, 1987; Marsh & Dunkin, 1992). An exception is a recent study by Allen (1995) which found a small but significant positive correlation between research productivity and teaching evaluations. The study by Centra (1983) also found that, although a moderate positive correlation existed between research and teaching effectiveness in the social sciences, it was nearly zero in the sciences. In spite of this, there is evidence that institutions continue to use research and publication as a factor in evaluating teaching effectiveness (Seldin, 1984), a practice which has been widely criticised (Aubrecht, 1984; Centra, 1993, p. 74).

Personality of Instructor

This factor has been examined surprisingly infrequently in the literature. Feldman (1986) in his review found that when instructor personality was inferred from self-reports by instructors, very few of the 14 clusters of traits showed significant correlations with overall student evaluations. On the other hand, when instructor personality was inferred from the perceptions of students or colleagues, most of the trait clusters showed significant correlations with overall student evaluations. Feldman does not, however, distinguish whether or not this relationship is a valid influence on student ratings or a possible source of bias (see also Marsh, 1987). A more recent study (Murray *et al.*, 1990) found that to some extent teaching effectiveness could be predicted from colleague perceptions of personality. Murray and his colleagues also determined that instructor effectiveness varies substantially across courses, and personality traits which may be helpful in a particular course could conceivably be a liability in another course. (This supports the view that ratings used for personnel and tenure decisions should be taken from several courses rather than a single one.) These authors argue that the relationship between personality and student ratings supports rather than opposes the validity of student ratings.

Seductiveness: The 'Dr. Fox' Effect

This section is concerned with the issue of whether an instructor who is highly

entertaining or expressive can procure excessively high ratings. In the original 'Dr. Fox' study (Naftulin *et al.*, 1973), a professional actor, referred to as 'Dr. Fox', gave a highly expressive and enthusiastic lecture which was devoid of content, and received high overall ratings. This was thought to demonstrate that a highly expressive and charismatic lecturer can seduce the audience into giving undeservedly high ratings. However, this study has been severely criticised and the issue of seductiveness reanalysed by many later studies (Marsh, 1987), and now the so-called 'Dr. Fox' effect has very little support in the literature (Perry, 1990). A review of studies on educational seduction by Abrami *et al.* (1982a) found that student ratings were much more sensitive to expressiveness than to lecture content. (In a later article, however, Abrami (1989a) wrote "... expressiveness affects ratings more than student learning and thus represents a bias".) Marsh and Ware (1982) in their reanalysis found that when students are not given incentive to learn, expressiveness had a much greater effect on student ratings than content coverage; however, when incentive was given, a situation more like a college classroom, expressiveness was less important and there was essentially no 'Dr. Fox' effect.

Gender of Instructor

Discussion of the effect of teacher gender on student evaluations of teaching appears to be quite varied. Many authors contend that student ratings are biased against women instructors (for example, Basow, 1994; Basow & Silberg, 1987; Kaschak, 1978; Koblitz, 1990; Martin, 1984; Rutland, 1990). A few studies (Bennett, 1982; Kierstead *et al.*, 1988) have found that female instructors need to behave in stereotypically feminine ways in order to avoid receiving lower ratings than male instructors. In view of this, Koblitz (1990) sees a difficulty for women instructors who need to adopt a 'get tough' approach. On the other hand, Tatro (1995) found that female instructors received significantly higher ratings than male instructors. In a two-part meta-analysis, Feldman (1992, 1993) reviewed existing research on student ratings of male and female teachers in both the laboratory and the classroom setting. In his review of laboratory studies, Feldman (1992) reports that the majority of studies reviewed showed no difference in the global evaluations of male and female teachers. In the minority of studies in which differences were found, male instructors received higher overall ratings than females. Very few studies in this review reported interaction effects between student and teacher gender, and such effects when reported were inconsistent. Subsequently, in his review of classroom studies, Feldman (1993) again reported that the majority of studies reported no significant differences between the genders. However, in the few studies where differences were reported, this time the female instructors received slightly higher overall ratings than the males. An interaction effect was found in that students tended to rate same-gender teachers slightly higher than opposite-gender teachers.

Minority Status of Instructor

No studies have yet investigated whether there exists a systematic racial bias in student evaluations of teaching (Centra, 1993, p. 76). However, a more recent paper by Rubin (1995) examines differences in perceptions of non-native speaking instructors of various nationalities.

Physical Appearance of Instructor

A study by Buck and Tiene (1989) found that physical attractiveness of the instructor did not have an effect by itself on perceptions of teacher competence, but there was a significant interaction between gender, attractiveness and authoritarianism; namely, teachers with an authoritarian philosophy were rated less negatively if they were attractive and female. Rubin (1995) found that students' judgements of teaching ability of non-native speaking instructors were affected by judgements of physical attractiveness.

Characteristics of the Students

Personality Characteristics

No consistent relationship exists between student personality characteristics and evaluations of teaching (Abrami *et al.*, 1982b).

Prior Subject Interest

Evidence suggests that students with greater interest in the subject area prior to the course tend to give more favourable teacher ratings (Feldman, 1977, p. 236; Marsh & Cooper, 1981; Prave & Baril, 1993). Marsh and Dunkin (1992) assert that the influence of prior subject interest on student ratings does not constitute a bias, although they admit that when ratings are used summatively this influence can be a source of unfairness in that it is a function of the course and not the teacher. Controlling for prior subject interest may be especially important in areas which have a large percentage of students enrolled in service courses, such as mathematics and English, for example. Prave & Baril (1993) discuss means of measuring and controlling for this factor.

Gender of Students

Conflicting findings have resulted from studies of the relationship between student gender and ratings (Aleamoni, 1987; Aleamoni & Hexner, 1980; Feldman, 1977). Many studies have reported that there is essentially no difference in ratings by male and female students, but quite a few have also come to a different conclusion. Feldman (1977) reports that among those studies which did find significant relationships between student gender and ratings, most found that female students gave higher ratings than males. Tatro (1995) also found that female students gave higher ratings than male students. However, Koushki and Kuhn (1982) found in their study that male students gave slightly higher ratings than female students. In addition, some studies have reported a tendency for students to rate same-sex instructors slightly higher than opposite-sex instructors (Feldman, 1993; Centra, 1993).

Expected Grade and the Leniency Hypothesis

The effect of a student's expected grade in a course on the student's evaluation of the teacher of that course has been one of the most controversial topics in the literature on student evaluation of teaching. At this time the consensus is definitely that there is a moderate positive correlation between expected grade and student ratings (students expecting higher grades will give more favourable ratings) (Braskamp & Ory, 1994;

Centra, 1979, p. 32; DuCette & Kenney, 1982; Feldman, 1976a; Howard & Maxwell, 1980; Marsh, 1987; Marsh & Dunkin, 1992). The controversy concerns the interpretation of this association. Marsh (1987; Marsh & Dunkin, 1992) suggests three plausible interpretations:

- The leniency hypothesis. Instructors with more lenient grading standards receive more favourable ratings, i.e. a teacher can 'buy' better evaluations by assigning higher grades. This would suggest that the grades-ratings relationship constitutes a bias and a threat to the validity of student ratings.
- The validity hypothesis. More effective instructors cause students to work harder, learn more and earn better grades. This implies that the grades-ratings relationship supports the validity of student ratings.
- The student characteristic hypothesis. Pre-existing student characteristics such as prior subject interest affect both teaching effectiveness and student ratings.

Numerous authors have argued in favour of the leniency hypothesis (Chacko, 1983; Koshland, 1991; Nimmer & Stone, 1991; Powell, 1978; Snyder & Clair, 1976; Vasta & Sarmiento, 1979; see also DuCette & Kenney, 1982) and against it (Abrami *et al.*, 1980; Howard & Maxwell, 1980; Marsh, 1987; Theall & Franklin, 1991; see also Franklin *et al.*, 1991). Chacko (1983), for example, showed that more strict grading standards led students to rate the instructor lower even on components of instruction unrelated to grading fairness, such as humor, self-reliance and attitude toward students. Goldberg and Callahan (1991) in their study found that adjunct faculty tended to give higher grades and receive higher ratings than full-time faculty, even though most students were not aware of their instructor's status; they call for further study of the relationship between adjunct status and student ratings. Nimmer and Stone (1991) found that leniency bias was a significant problem, and in addition, that the degree of leniency bias was affected by the time at which the ratings were administered, i.e. the bias was greater when the evaluations were administered after an examination, even if the students had not yet received the results of the examination. At the present time the dispute over the possibility of leniency bias is not resolved.

Another view on the expected grade controversy is provided by Hewett *et al.* (1988), who suggest that what matters is not so much the students' expected grades, but the students' impressions of the course as dictated by their examination performance. These authors found that correlation between student scores on a *single* examination and end-of-course evaluations were strongest with the first examination, weaker with each succeeding examination, and weakest with the last examination given before the administration of the student evaluation. Hewett and her colleagues offer this as a possible explanation for mixed results in expected grade effect studies, as the effect on ratings of the overall grade is weaker than that of the first examination grade only. They also conclude that student ratings may not measure overall teaching effectiveness, but only the students' first impressions of the instructor's effectiveness. This discussion of the 'first impression effect' also raises the question of whether an instructor can procure higher student ratings not by giving higher grades, but by giving an easy first examination.

Student Expectations

McKeachie (1979) asserts that the single most important student characteristic affecting student ratings is student expectations, i.e. students who expect an instructor to be good

usually find this to be so. Research on classroom expectation is considerable, but much less emphasis has been placed on student expectation than on teacher expectation (Gigliotti, 1987; Koermer & Petelle, 1991; Perry *et al.*, 1979). Student expectations, a factor outside the control of the instructor, has been shown to influence course ratings; specifically, students with high expectations and high experiences give higher ratings than those with low expectations and high experiences or with low experiences (Koermer & Petelle, 1991). Also Gigliotti (1987) found that negative expectancy violations resulted in unfavorable course evaluations for the items in his study.

Emotional State

We know of only one study which addresses this variable; Small *et al.* (1982) found that the emotional state of students at the end of the semester, when student ratings are usually collected (but not earlier in the term) were associated with their ratings of instruction. Specifically, the more hostile, anxious and depressed the students were feeling at the end of the term, the lower the ratings they gave to their instructor. Although the authors believe that this is a serious threat to the validity of end-of-semester student ratings, we have not seen their results corroborated more recently.

Student Age

Surprisingly, to our knowledge there have not been any recent studies which examine specifically the effect of the age of the students (as opposed to closely related variables such as course level and year in school) on student ratings of instruction. We do not know, for instance, whether higher ratings in upper-level courses are a result of the more advanced level of subject matter, or the students being older and more mature. A study by Klann and Hoff (1976) of mathematics students at a community college found no difference in the ratings of students under 20 years old and students at least 20 years old. Further studies at both two-year and four-year college with more age brackets would be indicated.

Reaction to the Use of Student Evaluations

Reaction by Faculty

In a study by Ryan *et al.* (1980), it was found that the introduction of *mandatory* use of student evaluations led to a significant reduction in faculty morale and job satisfaction. It also motivated faculty to reduce standards and workloads for their students and to make examinations easier. A similar study in which the use of student ratings was not compulsory (Baxter, 1991) found that faculty were generally satisfied with the process. Such a study would undoubtedly contain some non-response bias due to self-selection. Ryan and his colleagues also believed that the imposition of mandatory student ratings may contribute to grade inflation. This relates to the above discussion of the leniency hypothesis; however, even if that hypothesis is incorrect, faculty may nonetheless believe it to be accurate (Centra, 1993, p. 75) and reduce standards anyway. Jacobs (1987) found that a majority of faculty responding to her survey did not feel that student ratings have a negative effect on faculty morale and did believe that the ratings should be required; it is interesting to note, however, that her survey was conducted at a more highly research-oriented university than that of Ryan *et al.* (1980). Rich (1976) found that

faculty are more favourably disposed toward the use of student ratings for summative purposes at research oriented institutions where teaching holds a position of lesser importance. Avi-Itzhak and Kremer (1986) found that senior and tenured faculty are most opposed to the use of student ratings for summative purposes, while non-tenured faculty, both tenure-track and non-tenure-track, were most supportive of this use of student ratings. They believe that this is due to the fact that the senior faculty devote the most time to research and less time to their teaching, and are less 'student oriented' than junior faculty. Also, Cranton and Knoop (1991) argue that job satisfaction should be taken into account in a thorough consideration of teaching effectiveness. Spencer and Flyr (1992) found that only 23% of faculty responding to their survey reported making changes to their teaching based on student evaluation results, and when such changes did occur they usually took the form of altering handouts, presentation habits, and assignments. We note here that faculty who familiarise themselves with existing research on student ratings of instruction tend to have more positive attitudes about their use (Franklin & Theall, 1989). Other authors (for example, Kronk & Shipka, 1980) urge institutions to require the use of student ratings, and in fact the vast majority of institutions in the US do require them in some form (Wachtel, 1994, p. 65).

Reaction from Students

The nature of students' reactions to the common use of student rating forms has also attracted surprisingly little attention (Marlin, 1987). One study which did examine student satisfaction with the process was carried out by Abbott *et al.* (1990). The authors emphasised the importance of student satisfaction, since that may affect students' willingness to participate in the evaluation process. Abbott and his colleagues state that students often complain about the frequency with which they are asked to fill out rating forms and the degree to which faculty are perceived to adjust their courses in response to student feedback. They concluded that students are more satisfied with mid-term rather than end-of-course evaluations, and with extended instructor reaction to their opinions, though those two factors did not have a cumulative effect. The authors also express concern about the fact that the most common method of administering student evaluations is also that which is least satisfactory to students, namely end-of-course questionnaires. An earlier study (Wulff *et al.*, 1985) found that students prefer interviews to end-of-course questionnaires. (Faculty also rated evaluative information obtained from group interviews more highly in terms of accuracy, utility and believability than that obtained from student rating objective items or written comments, according to Ory & Braskamp, 1981.) Jacobs (1987) found in her study that 82% of instructors use student evaluations only at the end of a term, and 28% administered the evaluation on the last day of class. Brandenburg *et al.* (1979), concerned about the possible overuse of student rating forms, wonder whether students take the evaluations seriously, a question which to our knowledge has hardly been addressed by existing research since then. Marlin (1987) in his survey found that over half of the students sampled stated that they took sufficient time and attempted to be fair and accurate in the rating of their instructors; however, Marlin also states that students tend to view evaluations as a "vent to let off student steam" and that they complain that faculty and administrators pay scant attention to student opinions, and that teachers do not alter their behaviour based on the comments on student rating forms. Jacobs (1987) also reports that 40% of the students responding to her survey said that they have heard of students plotting to get back at an instructor by collectively giving low ratings. An overwhelming majority of the students reported

that they have never heard of an instructor who tried to manipulates students into giving higher ratings.

Publicizing Student Ratings

Another purpose which is occasionally cited for student evaluations of teaching is to provide assistance to students in making course selections. This remains the subject of controversy. McKeachie (1969) states that publication of student ratings may inhibit the improvement of poor teachers by provoking a feeling of defensiveness and anxiety. Aleamoni (1981) states that publication has created antagonism between students and faculty. On the other hand, Scriven (1981) argues that it is unethical to deny students the opportunity to view the results of ratings which they have engendered. Indeed, at one Midwestern university the Student Association organised a boycott of teacher evaluations in response to the faculty's refusal to share the results with them (Flaherty, 1993).

Concluding Remarks

Although the use of student ratings of instruction is well-entrenched in North American universities, there are still some useful areas for further research. A few background characteristics have not yet been sufficiently investigated as to whether there is a significant effect on student ratings, e.g. class meeting time, minority status and physical appearance of the instructor. Also, it is felt that the effect, if any, of the average *age* of the students on ratings could be studied, provided that other characteristics such as course level, class size and prior student interest can be controlled for. In view of the results cited in the section 'A Final Reflection on Course Characteristics' above, educators in individual disciplines may wish to study further the question of whether or not course characteristics would be likely to affect student ratings in the manner in which existing research has predicted, or in a different manner, or not at all, in courses in their own subject area. Further studies of whether or not students are inclined to take written rating forms seriously (whether or not they are formally instructed to do so) would be worthwhile. Perhaps it might also be worthwhile to do further surveys of faculty to determine how often and what types of changes they make in their instruction based on the results of student ratings. Finally, it might be interesting to examine whether the effect of certain background characteristics on student ratings may change depending on whether the course is taught at a highly research-oriented university, or a four-year institution which is more teaching-oriented, or a two-year institution such as a community college.

Note on Contributor

DR HOWARD WACHTEL received his Bachelor's degree in mathematics at Washington University in St. Louis, Missouri, his Master's degree in mathematics from the University of Wisconsin-Madison, Wisconsin, and his Doctorate in mathematics from the University of Illinois at Chicago, Illinois. His research interests are in the area of the evaluation of college teaching in general and mathematics teaching in particular. He is an assistant professor of mathematics at Bowie State University in Bowie, Maryland, USA.

Correspondence: Howard K. Wachtel, 9215J Livery Lane, Laurel, Maryland 20723–1614, USA. Tel: (301) 464 6015.

REFERENCES

ABBOTT, R. D., WULFF, D. H., NYQUIST, J. D., ROPP, V. A. & HESS, C. W. (1990) Satisfaction with processes of collecting student opinions about instruction: The student perspective, *Journal of Educational Psychology*, 82, pp. 201–206.

ABRAMI, P. C. (1989a) How should we use student ratings to evaluate teaching?, *Research in Higher Education*, 30, pp. 221–227.

ABRAMI, P. C. (1989b) SEEQing the truth about student evaluations of instruction, *Educational Researcher*, 18, 43–45.

ABRAMI, P. C., LEVENTHAL, L., PERRY, R. P. & BREEN, L. J. (1976) Course evaluation: How?, *Journal of Educational Psychology*, 68, pp. 300–304.

ABRAMI, P. C., DICKENS, W. J., PERRY, R. P. & LEVENTHAL, L. (1980) Do teacher standards for assigning grades affect student evaluations of instruction?, *Journal of Educational Psychology*, 72, pp. 107–118.

ABRAMI, P. C., LEVENTHAL, L. & PERRY, R. P. (1982a) Educational seduction, *Review of Educational Research*, 52, pp. 446–464.

ABRAMI, P. C., PERRY, R. P. & LEVENTHAL, L. (1982b) The relationship between student personality characteristics, teacher ratings, and student achievement, *Journal of Educational Psychology*, 74, pp. 111–125.

ALEAMONI, L. M. (1981) Student ratings of instruction, in: J. MILLMAN (Ed.) *Handbook of Teacher Evaluation*, pp. 110–145 (Beverly Hills, Sage).

ALEAMONI, L. M. (1987) Student rating myths versus research facts, *Journal of Personnel Evaluation in Education*, 1, pp. 111–119.

ALEAMONI, L. M. & HEXNER, P. Z. (1980) A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation, *Instructional Science*, 9, pp. 67–84.

ALLEN, M. (1995) Research productivity and positive teaching evaluations: Examining the relationship using meta-analysis. Paper presented at the Annual Meeting of the Western States Communication Association, Portland, Oregon.

AUBRECHT, J. D. (1984) Better faculty evaluation systems, in: P. SELDIN (Ed.) *Changing Practices in Faculty Evaluation* (San Francisco, Jossey-Bass).

AVI-ITZHAK, T. & KREMER, L. (1986) An investigation into the relationship between university faculty attitudes toward student rating and organizational and background factors, *Educational Research Quarterly*, 10, pp. 31–38.

BASOW, S. A. (1994) Student ratings of professors are not gender blind. Paper presented at the meeting of the Society of Teaching and Learning in Higher Education, Vancouver, British Columbia.

BASOW, S. A. & SILBERG, N. T. (1987) Student evaluations of college professors: Are female and male professors rated differently?, *Journal of Educational Psychology*, 79, pp. 308–314.

BAXTER, E. P. (1991) The TEVAL experience, 1983–88: The impact of a student evaluation of teaching scheme on university teachers, *Studies in Higher Education*, 16, pp. 151–178.

BENNETT, S. K. (1982) Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation, *Journal of Educational Psychology*, 74, pp. 170–179.

BLUNT, A. (1991) The effects of anonymity and manipulated grades on student ratings of instructors, *Community College Review*, 18, pp. 48–54.

BRANDENBURG, D. C., BRASKAMP, L. A. & ORY, J. C. (1979) Considerations for an evaluation program of instructional quality, *CEDR Quarterly*, 12, pp. 8–12.

BRANDENBURG, D. C., SLINDE, J. A. & BATISTA, E. E. (1977) Student ratings of instruction: Validity and normative interpretations, *Research in Higher Education*, 7, pp. 67–78.

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

BRASKAMP, L. A., BRANDENBURG, D. C. & ORY, J. C. (1984) *Evaluating Teaching Effectiveness* (Newbury Park, CA, Sage).

BRASKAMP, L. A. & ORY, J. C. (1994) *Assessing Faculty Work* (San Francisco, Jossey-Bass).

BUCK, S. & TIENE, D. (1989) The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations, *Journal of Educational Research*, 82, pp. 172–177.

CASHIN, W. E. (1988) *Student Ratings of Teaching: A Summary of the Research* (Manhattan, KS, Center for Faculty Evaluation and Development, Kansas State University).

CASHIN, W. E. (1990) Students do rate different academic fields differently, in: M. THEALL & J. FRANKLIN (Eds) *Student Ratings of Instruction: Issues for Improving Practice*, New Directions for Teaching and Learning, No. 43 (San Francisco, Jossey-Bass).

CASHIN, W. E. (1992) Student ratings: The need for comparative data, *Instructional Evaluation and Faculty Development*, 12, pp. 1–6.

CASHIN, W. E. & CLEGG, V. L. (1987) Are student ratings of different academic fields different? Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

CENTRA, J. A. (1973) Self-ratings of college teachers: A comparison with student ratings, *Journal of Educational Measurement*, 10, pp. 287–295.

CENTRA, J. A. (1974) The relationship between student and alumni ratings of teachers, *Educational and Psychological Measurement*, 34, pp. 321–326.

CENTRA, J. A. (1976) The influence of different directions on student ratings of instruction, *Journal of Educational Measurement*, 13, pp. 277–282.

CENTRA, J. A. (1977) Student ratings of instruction and their relationship to student learning, *American Educational Research Journal*, 14, pp. 17–24.

CENTRA, J. A. (1978) Using student assessments to improve performance and vitality, in: W. R. KIRSCHLING (Ed.) *Evaluating Faculty Performance and Vitality*, pp. 31–49 (San Francisco, Jossey-Bass).

CENTRA, J. A. (1979) *Determining Faculty Effectiveness* (San Francisco, Jossey-Bass).

CENTRA, J. A. (1983) Research productivity and teaching effectiveness, *Research in Higher Education*, 18, pp. 379–389.

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

CENTRA, J. A. & CREECH, F. R. (1976) *The relationship between student teachers and course characteristics and student ratings of teacher effectiveness.* Project Report 76–1 (Princeton, NJ, Educational Testing Service).

CHACKO, T. I. (1983) Student ratings of instruction: A function of grading standards, *Educational Research Quarterly*, 8, pp. 19–25.

CHANDLER, J. A. (1978) The questionable status of student evaluations of teaching, *Teaching of Psychology*, 5, pp. 150–152.

COHEN, P. A. (1980) Using student ratings feedback for improving college instruction: A meta-analysis of findings, *Research in Higher Education*, 13, pp. 321–341.

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

CRANTON, P. A. & KNOOP, R. (1991) Incorporating job satisfaction into a model of instructional effectiveness, in: M. THEALL & J. FRANKLIN (Eds) *Effective Practices for Improving Teaching, New Directions for Teaching and Learning*, No. 48, pp. 99–109 (San Francisco, Jossey-Bass).

CRANTON, P. A. & SMITH, R. A. (1986) A new look at the effect of course characteristics on student ratings of instruction, *American Educational Research Journal*, 23, pp. 117–128.

DERRY, J. O. (1979) Can student's ratings of instruction serve rival purposes? *Journal of Higher Education*, 50, pp. 79–88.

DOWELL, D. A. & NEAL, J. A. (1982) A selective review of the validity of student ratings of teaching, *Journal of Higher Education*, 53, pp. 51–62.

DUCETTE, J. & KENNEY, J. (1982) Do grading standards affect student evaluations of teaching? Some evidence on an old question, *Journal of Educational Psychology*, 74, pp. 308–314.

DUDLEY, S. & SHAWVER, D. L. (1991) The effect of homework on students' perceptions of teaching effectiveness, *Journal of Education for Business*, 67, pp. 21–25.

EBLE, K. E. (1970) *The Recognition and Evaluation of Teaching* (Salt Lake City, Project to Improve College Teaching).

FELDMAN, K. A. (1976a) Grades and college students' evaluations of their courses and teachers, *Research in Higher Education*, 4, pp. 69–111.

FELDMAN, K. A. (1976b) The superior college teacher from the students' view, *Research in Higher Education*, 5, pp. 243–288.

FELDMAN, K. A. (1977) Consistency and variability among college students in rating their teachers and courses: A review and analysis, *Research in Higher Education*, 6, pp. 223–274.

FELDMAN, K. A. (1978) Course characteristics and college students' ratings of their teachers: What we know and what we don't, *Research in Higher Education*, 9, pp. 199–242.

FELDMAN, K. A. (1979) The significance of circumstances for college students' ratings of their teachers and courses, *Research in Higher Education*, 10, pp. 149–172.

FELDMAN, K. A. (1983) Seniority and experience of college teachers as related to evaluations they receive, *Research in Higher Education*, 18, pp. 3–124.

FELDMAN, K. A. (1984) Class size and college students' evaluations of teachers and courses: A closer look, *Research in Higher Education*, 21, pp. 45–116.

FELDMAN, K. A. (1986) The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis, *Research in Higher Education*, 24, pp. 139–213.

FELDMAN, K. A. (1987) Research productivity and scholarly accomplishments: A review and exploration, *Research in Higher Education*, 26, pp. 227–298.

FELDMAN, K. A. (1988) Effective college teaching from the students' and faculty's view: matched or mismatched priorities, *Research in Higher Education*, 28, pp. 291–344.

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, pp. 137–189.

FELDMAN, K. A. (1992) College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments, *Research in Higher Education*, 33, pp. 317–375.

FELDMAN, K. A. (1993) College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers, *Research in Higher Education*, 34, pp. 151–211.

FLAHERTY, R. (1993) Protesting NIU students hold back faculty grades. *Chicago Sun-Times*, 6 December, p. 4.

FRANKHOUSER, W. M. (1984) The effects of different oral directions as to disposition of results on student ratings of college instruction, *Research in Higher Education*, 20, pp. 367–374.

FRANKLIN, J. & THEALL, M. (1989) Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco.

FRANKLIN, J., THEALL, M. & LUDLOW, L. (1991) Grade inflation and student ratings: A closer look. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

FREY, P. W. (1976) Validity of student instructional ratings: Does timing matter? *Journal of Higher Education*, 47, pp. 327–336.

GIGLIOTTI, R. J. (1987) Are they getting what they expect? *Teaching Sociology*, 15, pp. 365–375.

GOLDBERG, G. & CALLAHAN, J. (1991) Objectivity of student evaluations of instructors, *Journal of Education for Business*, 66, pp. 377–378.

GOLDMAN, L. (1993) On the erosion of education and the eroding foundations of teacher education (or why we should not take student evaluation of faculty seriously), *Teacher Education Quarterly*, 20, pp. 57–64.

HELLER, S. (1984) Math department balks at officials' effort to require forms for student evaluation, *Chronicle of Higher Education*, 28(8), p. 24.

HEWETT, L., CHASTAIN, G. & THURBER, S. (1988) Course evaluations: Are students' ratings dictated by first impressions? Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Snowbird, UT.

HILLS, J. R. (1974) On the use of student ratings of faculty in determination of pay, promotion, and tenure, *Research in Higher Education*, 2, pp. 317–324.

HOFMAN, J. E. & KREMER, L. (1980) Attitudes toward higher education and course evaluation, *Journal of Educational Psychology*, 72, pp. 610–617.

HOWARD, G. S., CONWAY, C. G. & MAXWELL, S. E. (1985) Construct validity of measures of college teaching effectiveness, *Journal of Educational Psychology*, 77, pp. 187–196.

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

JACOBS, L. C. (1987) *University Faculty and Students' Opinions of Student Ratings*. Indiana Studies in Higher Education, #55 (Bloomington, IN, Bureau of Evaluation and Testing, Indiana University).

KASCHAK, E. (1978) Sex bias in student evaluations of college professors, *Psychology of Women Quarterly*, 2, pp. 235–242.

KIERSTEAD, D., D'AGOSTINO, P. & DILL, H. (1988) Sex role stereotyping of college professors: Bias in students' ratings of instructors, *Journal of Educational Psychology*, 80, pp. 342–344.

KLANN, W. E. & HOFF, E. (1976) The use of judgment analysis in analyzing student evaluation of teachers, *Mathematical Association of Two-Year College Journal*, 10, pp. 137–139.

KOBLITZ, N. (1990) Are student ratings unfair to women?, *Newsletter of the Association for Women in Mathematics*, 20, pp. 17–19.

KOERMER, C. D. & PETELLE, J. L. (1991) Expectancy violation and student rating of instruction, *Communication Quarterly*, 39, pp. 341–350.

KOON, J. & MURRAY, H. G. (1995) Using multiple outcomes to validate student ratings of overall teacher effectiveness, *Journal of Higher Education*, 66, pp. 61–81.

KOSHLAND, D. E. (1991) Teaching and research, *Science*, 251, p. 249.

KOUSHKI, P. A. & KUHN, H. A. J. (1982) How reliable are student evaluations of teachers?, *Engineering Education*, 72, pp. 362–367.

KRONK, A. K. & SHIPKA, T. A. (1980) *Evaluation of Faculty in Higher Education* (Washington, DC, National Education Association).

L'HOMMEDIEU, R., MENGES, R. J. & BRINKO, K. T. (1990) Methodological explanations for the modest effects of feedback from student ratings, *Journal of Educational Psychology*, 82, pp. 232–241.

LEVENTHAL, L., ABRAMI, P. C. & PERRY, R. P. (1976) Do teacher rating forms reveal as much about students as about teachers?, *Journal of Educational Psychology*, 68, pp. 441–445.

MARLIN, J. W., JR. (1987) Student perception of end-of-course evaluations, *Journal of Higher Education*, 58, pp. 704–716.

MARQUES, T. E., LANE, D. M. & DORFMAN, P. W. (1979) Toward the development of a system for instructional evaluation: Is there consensus regarding what constitutes effective teaching?, *Journal of Educational Psychology*, 71, pp. 840–849.

MARSH, H. W. (1979) *Annotated Bibliography of Research on the Relationship Between Quality of Teaching and Quality of Research in Higher Education* (Los Angeles, Office of Institutional Studies, University of Southern California).

MARSH, H. W. (1984) Students' evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility, *Journal of Educational Psychology*, 76, pp. 707–754.

MARSH, H. W. (1987) Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research, *International Journal of Educational Research*, 11, pp. 253–388.

MARSH, H. W. & COOPER, T. L. (1981) Prior subject interest, students' evaluations, and instructor effectiveness, *Multivariate Behavioral Research*, 16, pp. 82–104.

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 Press).

MARSH, H. W. & OVERALL, J. U. (1980) Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria, *Journal of Educational Psychology*, 72, pp. 468–475.

MARSH, H. W. & ROCHE, L. (1993) The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness, *American Educational Research Journal*, 30, pp. 217–251.

MARSH, H. W. & WARE, J. E., JR. (1982) Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the 'Dr. Fox' effect, *Journal of Educational Psychology*, 74, pp. 126–134.

MARTIN, E. (1984) Power and authority in the classroom: Sexist stereotypes in teaching evaluations, *Journal of Women in Culture and Society*, 9, pp. 482–492.

MCCALLUM, L. W. (1984) A meta-analysis of course evaluation data and its use in the tenure decision, *Research in Higher Education*, 21, pp. 150–158.

MCKEACHIE, W. J. (1969) Student ratings of faculty, *American Association of University Professors Bulletin*, 55, pp. 439–443.

MCKEACHIE, W. J. (1979) Student ratings of faculty: A reprise, *Academe*, 65, pp. 384–397.

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

MEETH, L. R. (1976) The stateless art of teaching evaluation. Report on teaching, Vol. 2, in *Change*, 8, pp. 3–5.

MENEFEE, R. (1983) The evaluation of science teaching, *Journal of College Science Teaching*, 13, p. 138.

MENGES, R. J. (1991) The real world of teaching improvement: A faculty perspective, in: M. THEALL & J. FRANKLIN (Eds) *Effective Practices for Improving Teaching, New Directions for Teaching and Learning*, Vol. 48, pp. 21–37 (San Francisco, Jossey-Bass).

MILLER, S. N. (1984) Student rating scales for tenure and promotion, *Improving College and University Teaching*, 32, pp. 87–90.

MONROE, C. & BORZI, M. G. (1989) Methodological issues regarding student evaluation of teachers: A pilot study, *ACA Bulletin*, 70, pp. 73–39.

MURRAY, H. G., RUSHTON, P. J. & PAUNONEN, S. V. (1990) Teacher personality traits and student instructional ratings in six types of university courses, *Journal of Educational Psychology*, 82, pp. 250–261.

NAFTULIN, D. H., WARE, J. E. & DONNELLY, F. A. (1973) The Doctor Fox lecture: A paradigm of educational seduction, *Journal of Medical Education*, 48, pp. 630–635.

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, pp. 195–215.

ORNSTEIN, A. C. (1990) A look at teacher effectiveness research—theory and practice, *National Association of Secondary School Principals Bulletin*, 74, pp. 78–88.

ORY, J. C. & BRASKAMP, L. A. (1981) Faculty perceptions of the quality and usefulness of three types of evaluative information, *Research in Higher Education*, 5, pp. 271–282.

OVERALL, J. U. & MARSH, H. W. (1979) Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes, *Journal of Educational Psychology*, 72, pp. 321–325.

PERRY, R. P. (1990) Introduction to the special section: Instruction in higher education, *Journal of Educational Psychology*, 82, pp. 183–188.

PERRY, R. P., NIEMI, R. R. & JONES, K. (1974) Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance, *Journal of Educational Psychology*, 66, pp. 851–856.

PERRY, R. P., ABRAMI, P. C., LEVENTHAL, L. & CHECK, J. (1979) Instructor reputation: An expectancy relationship involving student ratings and achievement, *Journal of Educational Psychology*, 71, pp. 776–787.

POWELL, R. (1978) Faculty rating scale validity: The selling of a myth, *College English*, 39, pp. 616–629.

PRAVE, R. S. & BARIL, G. L. (1993) Instructor ratings: Controlling for bias from initial student interest, *Journal of Education for Business*, 68, pp. 362–366.

PULICH, M. A. (1984) Better use of student evaluations for teaching effectiveness, *Improving College and University Teaching*, 32, pp. 91–94.

RAMSDEN, P. (1991) A performance indicator of teaching quality in higher education: The course experience questionnaire, *Studies in Higher Education*, 16, pp. 129–150.

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

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

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

RICH, H. A. (1976) Attitudes of college and university faculty toward the use of student evaluations, *Educational Research Quarterly*, 1, pp. 17–28.

RUBIN, D. (1995) Effects of language and race on undergraduates' perceptions of international instructors: Further studies of language and attitude in higher education. Paper presented at the International Communication Association, Albuquerque, NM.

RUTLAND, P. (1990) Some considerations regarding teaching evaluations, *Political Science Teacher*, 3, pp. 1–2.

RYAN, J. J., ANDERSON, J. A. & BIRCHLER, A. B. (1980) Student evaluation: The faculty responds, *Research in Higher Education*, 12, pp. 317–333.

SCHERR, F. C. & SCHERR, S. S. (1990) Bias in student evaluations of teacher effectiveness, *Journal of Education for Business*, 65, pp. 356–358.

SCOTT, C. S. (1977) Student ratings and instructor-defined extenuating circumstances, *Journal of Educational Psychology*, 69, pp. 744–747.

SCRIVEN, M. (1981) Summative teacher evaluation, in: J. MILLMAN (Ed.), *Handbook of Teacher Evaluation*, pp. 244–271 (Beverly Hills, Sage).

SELDIN, P. (1984) *Changing Practices in Faculty Evaluation* (San Francisco, Jossey-Bass).

SELDIN, P. (1989) Using student feedback to improve teaching, in: A. F. LUCAS (Ed.) *The Department Chairperson's Role in Enhancing College Teaching. New Directions for Teaching and Learning*, Vol. 37, pp. 89–97 (San Francisco, Jossey-Bass).

SELDIN, P. (1993) The use and abuse of student ratings of professors, *Chronicle of Higher Education*, 39(46), p. A40.

SHEEHAN, D. S. (1975) On the invalidity of student ratings for administrative personnel decisions, *Journal of Higher Education*, 46, pp. 687–699.

SMALL, A. C., HOLLENBECK, A. R. & HALEY, R. L. (1982) The effect of emotional state on student ratings of instructors, *Teaching of Psychology*, 9, pp. 205–208.

SNYDER, C. R. & CLAIR, M. (1976) Effects of expected and obtained grades on teacher evaluation and attribution of performance, *Journal of Educational Psychology*, 68, pp. 75–82.

SPENCER, P. A. (1992) *Improving Teacher Evaluation* (Riverside, CA, Riverside Community College).

SPENCER, P. A. & FLYR, M. L. (1992) *The formal evaluation as an impetus to classroom change: Myth or reality*? (Research/Technical Report, Riverside, CA.)

TAGOMORI, H. T. & BISHOP, L. A. (1995) Student evaluation of teaching. Flaws in the instruments, *Thought and Action*, 11, pp. 63–78.

TATRO, C. N. (1995) Gender effects on student evaluations of faculty, *Journal of Research and Development in Education*, 28, pp. 169–173.

THEALL, M. & FRANKLIN, J. (1991) Using student ratings for teaching improvement, in: M. THEALL & J. FRANKLIN (Eds) *Effective Practices for Improving Teaching*, New Directions for Teaching and Learning, No. 48, pp. 83–96 (San Francisco, Jossey-Bass).

TIBERIUS, R. G., SACKIN, D. H., SLINGERLAND, J. M., JUBAS, K., BELL, M. & MATLOW, A. (1989) The influence of student evaluative feedback on the improvement of clinical teaching, *Journal of Higher Education*, 60, pp. 665–681.

VASTA, R. & SARMIENTO, R. F. (1979) Liberal grading improves evaluations but not performance, *Journal of Educational Psychology*, 71, pp. 207–211.

WACHTEL, H. K. (1994) A critique of existing practices for evaluating mathematics instruction. Doctoral dissertation, University of Illinois at Chicago, *Dissertation Abstracts International*, 56, p. 0129.

WILSON, R. C. (1986) Improving faculty teaching: Effective use of student evaluations and consultants, *Journal of Higher Education*, 57, pp. 196–211.

WULFF, D. H., STATON-SPICER, A. Q., HESS, C. W. & NYQUIST, J. D. (1985) The student perspective on evaluating teaching effectiveness *ACA Bulletin*, 53, pp. 39–47.

ZELBY, L. W. (1974) Student-faculty evaluation, *Science*, 183, pp. 1267–1270.

ZOLLER, U. (1992) Faculty teaching performance evaluation in higher science education: Issues and implications (a 'cross-cultural' case study), *Science Education*, 76, pp. 673–684.