Students Evaluations of Instructors' Performance in Higher Education Institutions: What Research Has to Say

Wossenu Yimam*

Abstract: The evaluation of instructors' teaching performance by their students has been used since the early 1920's. There has been a tremendous increase in interest regarding students' evaluations of teaching and this topic has been the subject of a substantial body of research spanning over 70 years with nearly 2000 studies. Most of these researches show that student evaluations are generally reliable and valid methods for gathering data on teaching. However, student evaluations are certainly not a perfect measure of teaching. To help substantiate and extend data from student evaluations, the evaluation process should include the triangulation of results from student evaluations, colleague evaluations, and supervisor evaluations. In light of this view, this paper attempts to discuss the following issues that are mostly raised in higher education research: (1) Are student evaluations reliable and valid? (2) Are students able to make correct judgments prior to having been away from the course, and possibly from the university, for a number of years? (3) Are student evaluations a popularity contest? (4) Do grades students receive or expect to receive affect their evaluations of the course and instructor? (5) Do extraneous variables bias student evaluations? (6) Can student evaluations be used to improve instruction and/or make personnel decisions? Finally, the paper tries to make conclusions and forward recommendations based on critical review of available literature.

Introduction

Students' evaluations of their instructors' teaching performance were first used in the early 1920's when students at the University of Washington were asked to fill out questionnaires about their professors (d'Apollonia and Abrami, 1997). Ever since that time, there has been a tremendous increase in interest regarding student evaluations of instruction and this topic has been the subject of a substantial body of research in higher education spanning over 70 years (Arreola, 1995). There are nearly 2000 published articles dealing with research on student evaluation of instruction. Most of these articles demonstrate that student evaluations are generally a

*Assistant. Professor, Institute of Educational Research, Addis Ababa University

reliable and valid method for gathering data on teaching much more so than any other teaching evaluation method.

With this as a background, the paper proceeds to treat the following research issues that are most commonly raised in student evaluations of instruction in higher education institutions.

Research Issues in Students' Evaluation of Instruction

Are student evaluations reliable and valid?

Concerns regarding issues of reliability and validity are germane for evaluation forms that have not gone through the rigorous psychometric testing necessary to produce professional evaluation forms. According to a review of the literature conducted by Aleamoni (1987) and Arreola (1995), well-developed and tested student evaluation forms are both reliable and valid. Before delving into the details of this issue, here it will be useful to define what reliability and validity mean.

a. Reliability: this indicates how consistently a set of items measure a particular construct or set of constructs. This can refer to consistency across evaluators (e.g. all students evaluate an instructor as a "4"), termed inter-evaluator reliability; across time (e.g. an instructor receives the same evaluations every semester) termed test-retest reliability; across items (e.g. an instructor is consistently evaluated highly on all the "organization" items) called internal consistency. In short, reliability provides information on the extent to which a given measurement will give similar information in different contexts or times of measurement. Inter-evaluator reliability or "inter-evaluator agreement" is a key indicator of the reliability of student evaluation forms. Marsh and Roche (1997) state that the reliability of student evaluation forms "is most appropriately determined from studies of inter-evaluator agreement that assess agreement among different students within the same course" (p. 1188). In a study conducted by Marsh (1987), he found that while correlations indicative

of reliability between two evaluators were low (.20s), the reliability of the class average response was high. He also noted that the reliability of the class average response depends on the number of students in a class. Reliability correlations (where 1.0 indicates a perfect correlational relationship) were: .95 for 50 students, .90 for 25 students, .75 for 10 students, and .60 for 5 students (Marsh, 1987). The findings of other researchers (e.g. Costin, Greenough and Menges 1971; Marsh, 1984) also support the reliability of student evaluation forms reporting reliabilities for professionally constructed forms to be approximately 0.90. Aleamoni (1987) cites several evaluation forms with a reliability of 0.90 and above. Centra (1993) maintains that reliability estimates for student evaluations of instruction are "good" (p. 58). Finally, Costin, Greenough and Menges (1971:513) comment:

> It would appear, then, that students can evaluate classroom instruction with a reasonable degree of reliability. In particular, the evidence cited concerning the stability of students' evaluations argues against the connection (sometimes made by opponents of student evaluations) that opinions of instruction are difficult to interpret, since they might be made after a particularly good or bad atypical experience.

b. Validity: this refers to the degree to which a test actually measures what it is supposed to measure. Most researchers agree that validity is more difficult to determine than reliability. Nonetheless, numerous researchers (Abrami, d'Apollonia and Cohen, 1990; Cohen, 1981; Feldman, 1989; Marsh, 1987) all conclude that student evaluations of instruction are, indeed, valid. Much of the evidence that supports the validity of student evaluation forms arises from studies in which student evaluations are "correlated with other indicators of teacher competence" (Arreola, 1995). For example, student evaluations are often correlated with colleague evaluations, trained observers' evaluations, alumni evaluations or measures of student learning. Aleamoni and Hexner (1980) cited 14 studies in which student

evaluations were compared to the above indicators. Moderate to high positive correlations were found which, in turn, support the validity of student evaluations of instruction. Similarly, Murray (1984: 119) summarized several general reviews (Marsh, 1983 cited in Murray, 1984; McKeachie, 1979; Murray, 1980 cited in Murray, 1984) and concluded that:

> Student evaluations of classroom teaching correlate moderately to highly (0.50 to 0. 90) with comparable evaluations made by supervisors, colleagues, alumni, and paid classroom observers, indicating that student perceptions of good and poor teaching are similar to those of more expert, more mature, and more neutral observers.

McKeachie(1997) also noted that researchers agree that student evaluations are the single most valid source of data on teaching effectiveness and there is little evidence of the validity of any other sources of data.

Generally, in terms of reliability and validity, data provided by students are the most investigated aspect of faculty evaluation with the greatest weight of consistent, positive supporting evidence. Unfortunately, because it often is emphasized as the sole measure of teaching and is not carefully collected and reported, it remains the most suspect aspect of faculty evaluation (Theall, 1997, cited in Sorcinelli, 1999).

Are students able to make correct judgments prior to having been away from the course, and possibly away from the university, for a number of years?

Since student evaluations are usually carried out anonymously, it is highly problematic to compare evaluations of a given student or a group of students, years after they have graduated. Therefore, most of the research in this area looks at the relationship between student

evaluations by alumni, or graduating seniors, and those made by current students. Research in this area was conducted by Drucker and Remmers at Purdue University early in 1950 and 1951 (cited in Arreola, 1995, Aleamoni, 1987). High positive correlations were found between evaluations of graduates of 5 and 10 years and currently enrolled students. Similar studies were conducted at the University of Illinois (Aleamoni and Yimer, 1974 cited in Aleamoni, 1987), and at the University of California, Los Angeles (Marsh, 1977) and produced similar results (Aleamoni, 1987), as did studies by Marsh and Overall (1979) and McKeachie, Lin and Mendelson (1978).

In general then, the evidence suggests considerable stability in the evaluations of courses and instructors; those evaluated most highly by currently enrolled students are also likely to be highly regarded when considered retrospectively.

Are student evaluations a popularity contest?

Many faculty believe this statement to be true, and accordingly, much has been written about this issue. In what has come to be termed the "Dr. Fox" studies, there have been results that suggest that instructors who are enthusiastic and expressive will receive good student evaluations regardless of the content they deliver in their classes. In the original Dr. Fox study, a professional actor gave a lecture to educators and graduate students in a dynamic and enthusiastic manner, but devoid of meaningful content. Nevertheless, he received favourable evaluations (Nauftin et al., 1973 cited in Centra, 1993). This original study, however, has been soundly criticized for methodological weaknesses (See Marsh and Dunkin, 1997).

In a re-evaluation of subsequent "Dr. Fox" studies, Marsh and Ware (1982) discovered that when students are given an incentive to learn, (i.e. students know that they will be tested on the material), a situation much closer to the real university setting, the "Dr. Fox" effect did not occur. In other words, the instructor who is expressive, yet does not deliver the appropriate content, is evaluated highly only in those

categories directly related to enthusiasm (i.e. "Instructor Enthusiasm") and receives appropriately lower scores in categories such as "Instructor Knowledge" and "Organization and Clarity" (Marsh and Roche, 1997). In a review of the "Dr. Fox" research, Abrami, Leventhal and Perry (1982) comment that much of this research has been fraught with inconsistencies in findings from various studies, which, in turn, has led to disagreement among reviewers.

Costin et al., (1971) reviewed Guthrie as cited in Costin et al., (1971) study in which he found that instructors who were highly evaluated were considered to be "substance teachers" and not simply "entertainers" (p. 518). Furthermore, in Murray's 1983 study, in which he employed neutral observers, he concluded that student evaluations seemed "determined more by the actual classroom behaviours of the instructor than by extraneous factors such as "personality" or "popularity" (p. 146). Murray also reasons that "expressive teaching behaviors serve to communicate the lecturer's enthusiasm for the subject matter, and thereby elicit and maintain student attention to lecture material" (p. 147). This, in turn, assists students in remembering the material which they have learned-and consequently and appropriately, also affects the evaluations students give their instructor (Murray, 1983).

In other research, Grush and Costin (1975) found that the correlation between the personal attraction students held for their instructors and how highly they evaluated those instructors (i.e. "teacher skill") was low (Grush and Costin, 1975). Aleamoni as cited in Aleamoni (1987) reviewed thousands of written comments by students and discovered that while they praised instructors for their humour and enthusiasm, if their courses were not well-organized, for example, the students also criticized their professors on this point. As Aleamoni (1987: 17) puts it:

> ...the students are not easily fooled. In evaluating their instructors, students discriminate among various aspects of teaching ability: If a teacher tells great jokes and has the students in the palm of his

> or her hand in the classroom, he or she will receive high evaluations in humor and classroom manner, but these evaluations do not influence students' assessments of other teaching skills.

Research by Costin, Greenough and Menges (1971), Frey (1978), and Arreola as cited in Arreola (1995) also identifies students "as discriminating judges of instructional effectiveness" (Arreola, 1995, p. 84). Centra (1993: p. 77) summarizes the view of many researchers when he comments:

> Do these findings indicate that student evaluations are unduly affected by expressive instructors? Probably not. First, Abrami, Leventhal, and Perry (1982) mention that, a twenty- to thirty-minute videotaped lecture represents only a minuscule percentage of actual lecture time in a three-credit course. Second, such extreme manipulativeness is unlikely in real-life teaching situations. Few college teachers provide no content in their courses and instead substitute enthusiasm. For these reasons, generalizations from the laboratory experiments to actual classroom teaching are tenuous. But if we were to generalize, a reasonable lesson from seduction research would be that by teaching more enthusiastically, teachers will receive high evaluations and their students will retain more of the course content.

A study by Williams and Ceci (1997), however, found that instructor enthusiasm had a strong biasing effect on student evaluations. One of the authors (Ceci), took a faculty development seminar to improve his "presentational style" since he had consistently received average evaluations. When he taught the course again, he made the same main points, used the same text, syllabus and overheads but changed the level of his enthusiasm and used the presentational techniques

(voice inflection, gesturing) which he had learned in the seminar. The end result was that his student evaluations were significantly higher than his previous evaluations. Furthermore, his evaluations improved in areas not "directly related" to instructor enthusiasm (i.e. "knowledge," "accessibility outside of class"). In addition, there was no positive correlation (as might be expected), between instructor enthusiasm and student learning. The students in his more "enthusiastic" class did not do better on tests than his previous students.

Although this study shows an "enthusiasm effect" that appears to question the validity of the evaluations, a number of points should be noted about the study. For example, although students did not perform any better on the exams in the more enthusiastic condition, it is not clear how much lecture material was in the exams. If the exams were primarily on the text, enthusiasm should not be expected to exert much effect on performance. Also, the effect was evaluated only with one class, and no attempt was made to recheck the effect with a third class. Thus, it is not clear that the results observed did not simply reflect that the two classes were composed of different people at a different time. In fact, d'Apollonia and Abrami (1997b) severely criticize the study noting that Williams and Ceci's literature review is "selective, biased, and erroneous" and the research itself has a number of serious "methodological flaws" (p. 18). d'Apollonia and Abrami challenge the claim made by Williams and Ceci that student evaluations are invalid and biased and suggest that their study has little or no value. While their study may suggest some future research to define appropriate limits on the use of student evaluation data, they view the study as so poorly done that it offers no basis for strong conclusions.

Furthermore, as Brown (1998) reminds us, this is only one study in comparison to numerous others which offer opposite results. There are certainly numerous factors that can affect students' performances on exams. Brown (1998:6) does point out that enthusiasm alone will not help an instructor with serious "flaws":

> What the study [Williams and Ceci, 1997] shows, at a minimum, is that a well structured course with a well chosen text book and clear syllabus can be considerably down-graded by students if the instructor lacks enthusiasm. It does not show that a poor instructor can get better evaluations on a flawed course simply by being more enthusiastic.

It is also important that the advantages of an expressive and enthusiastic instructor for student outcomes beyond test performance should not be overlooked. These include such variables as class attendance, selection of courses and majors, and perceived approachability of the instructor. For example, Phillips (1998:9) conducted a study at York University in which he collected student opinion regarding student evaluations of teaching. He commented that:

> Students admitted that personality did enter into their assessment and that they would most likely evaluate the charismatic lecturer more highly. However, they insisted that this was relevant to the question of the effectiveness of the pedagogy. To quote one student "if I am bored I learn less...if I am constantly engaged by the teacher I learn more".

In conclusion then, with the exception of the study by Williams and Ceci (1997), the belief that student evaluations are based on popularity or personality variables has not been substantiated by the literature. As Braskamp and Ory, (1994:180) clearly put it, "Neither the `stand-up comic' with no content expertise nor the `cold-fish expert' with only content expertise receives the highest evaluations consistently".

Do grades students receive or expect to receive affect their evaluations of the course and instructor?

The "expected grades/grading leniency" concern is perhaps the most controversial and, according to Arreola (1995), the most researched, of the potential biases to student evaluations. Murray (1996:18), however, points out that, to the degree that higher grades reflect greater learning, a positive relationship between grades and evaluations is appropriate:

> ...the average correlation of 0.28 found between grades and evaluations may reflect a tendency for highly evaluated teachers to foster high levels of learning in their students, which in turn results in justifiably higher student grades. In other words, the positive correlation between grades and evaluations may be a valid reflection of differential teacher effectiveness rather than an impetus for grade inflation.

Marsh and Roche (1997:1192) also point out that research on the grading leniency effect indicates that the effect is both "weak" and "the size of such an effect is likely to be unsubstantial". Similar to (Murray 1996:1194), they also note that:

> Class-average grades are correlated with class-average students' evaluations of teaching, but the interpretation depends on whether higher grades represent grading leniency, superior learning, or preexisting differences.

Greenwald and Gillmore (1997:1211) posit five hypotheses intended to explain the grades-evaluations correlation. These hypotheses are: (1) teaching effectiveness influences both grades and evaluations; (2) students' general academic motivation influences both grades and evaluations; (3) students' course-specific motivation influences both grades and evaluations; (4) students infer course quality and own

ability from received grades; and (5) students give high evaluations in appreciation for lenient grading .The first and fifth of these possibilities are the most directly contradictory. The first hypothesis holds that in courses taught by good instructors, students learn a lot, deserve high grades, and as a result of their learning, give appropriately high evaluations to their instructors. Therefore, "instructional quality" adequately explains the grades-evaluations correlation (Greenwald and Gillmore, 1997: 1211). However, these researchers argue for the fifth hypothesis that undeserved grades produce undeserved high evaluations. These researchers point out that this was supported by critics of student evaluations in the 1970s. However, support for the "leniency" hypothesis dropped sharply due to "correlational construct-validity research conducted in the late 1970s and early 1980s".

Studies examining construct validity attempt to answer the question "do student evaluations measure the construct (i.e. teaching effectiveness) they are supposed to measure?" Construct validity then, is a measure of whether, and the extent to which, a given survey (or other measure) captures the concept it was designed to assess.

Greenwald and Gillmore (1997:1209) conclude that the results of their study showed that grading leniency does influence student evaluations to a degree sufficient to warrant a statistical correction in order to "remove the unwanted inflation of evaluations produced by lenient grading".

This study, however, was criticized by various scholars in the field. According to Brown (1998), the major criticisms of Greenwald and Gillmore's study are two-fold. One suggests that there may be other possible explanations than the five theories they discuss and they have not dismissed these other explanations. The other criticism has been that what these authors have studied may not be lenient grading at all, but rather, just high marks. Brown (1998:5) suggests two points to consider:

> [I]t's possible that [Greenwald and Gillmore's] "lenient graders" are really just more effective

> teachers who deserve the higher evaluations and whose students earn higher grades. The other suggestion has been that it's not clear that even lenient grading falls outside the circle of teaching effectiveness: to the extent that getting higher grades is motivating to students, a tendency to assign them may in fact be relevant to teaching effectiveness.

McKeachie (1997) also finds Greenwald's and Gillmore's argument "flawed" on a number of counts. He agrees with Greenwald and Gillmore that giving higher than deserved grades may result in receiving higher than deserved evaluations, but only if the students are led to believe that they are learning more than "is typical." But (McKeachie 1997:1220) argues that "students are not so likely to be positively affected if an ineffective teacher seems to be trying to buy good evaluations with easy grades" and cites evidence that this tactic may, in fact, "boomerang."

Marsh and Roche (1997:1192) summarize their review of the literature on the "expected grades/grading leniency" concern and conclude that: "whereas a grading-leniency effect may produce some bias in students' evaluations of teaching, support for this suggestion is weak, and the size of such an effect is likely to be unsubstantial".

Do extraneous variables bias student evaluations?

A number of variables not directly relevant to academic performance have been suggested to affect student evaluations of instruction. They include: size of the class, gender of the instructor and student, level of course, rank of the instructor, student workload, and the value-system or ideology of the instructor. Let’s see what evidence exists for the importance of these variables.

a. Class Size

Many faculty believe that instructors who teach smaller classes are evaluated more highly than instructors who teach larger classes since

smaller classes allow for more instructor-student contact. Aleamoni's (1987) review of the research (see Aleamoni and Hexner, 1980), however, did not yield significant relationships between class size and student evaluations. Aleamoni and Hexner did cite older studies that showed a correlation between evaluations and class size, but they also cited several studies that gave the opposite conclusions. Arreola (1995) describes the findings of some studies which reported a curvilinear relationship between student evaluations and class size. That is, small (approximately under 30 students) and very large classes (approximately 120 students or more), are evaluated more favourably than those classes in the mid-range. (e.g. Kohlan, 1973; Linsky & Straus, 1974; Marsh, Overall, & Kesler, 1979; Pohlmann, 1975 all cited in Arreola, 1995).

In Marsh's (1987) comprehensive review of the research pertaining to student evaluations, in addition to his own study, he concludes that class size is not a bias to student evaluations. Rather, class size has a "moderate" effect on particular aspects of "effective teaching (primarily Group Interaction and Individual Rapport) and these effects are accurately reflected in the student evaluations" (p. 314). Marsh points out that the class size discussion serves to emphasize the multidimensionality of student evaluations; student evaluations cannot be comprehended fully without understanding their multidimensional nature (Marsh, 1987). In Marsh and Roche's 1997 overview of the relationships between a number of extraneous variables and student evaluations, they state that there are "mixed findings" in relation to class size, "but most studies show smaller classes are evaluated somewhat more favorably, although some find curvilinear relationships where large classes also are evaluated favorably" (p. 1194). McKeachie's (1997:1220) comments on the class size issue are also worthy of note:

> The concern about class size seems to me to be valid only if a personnel committee makes the mistake of using evaluations to compare teachers rather than as a measure of teaching effectiveness. There is ample

> evidence that most teachers teach better in small classes. Teachers of small classes require more papers, encourage more discussion, and are more likely to use essay questions on examinations--all of which are likely to contribute to student learning and thinking. Thus, on average, small classes should be evaluated higher than large classes.

b. Gender

According to Arreola (1995), results in the literature regarding gender and evaluations are inconsistent. Aleamoni and Hexner (1980) found no significant relationship between evaluations and gender (of the instructor or student). Other researchers (Doyle and Whitely, 1974; Isaacson, McKeachie et al., 1964 cited in Arreola, 1995) support this conclusion. In Costin et al.'s (1971) review of the research, they also cite seven studies that confirm the absence of significant differences between the evaluations made by male or female students, and the evaluations received by male and female instructors.

In contrast, both Costin and associates and Aleamoni and Hexner also cite a study by Bendig as cited in Costin et al., (1971) which show that female students tended to be slightly more critical of their male instructors than were their fellow male students. And another study by Walker as cited in Costin et al., (1971:520) found that female students evaluated female instructors "significantly higher" than they evaluated male instructors. Furthermore, investigations in the 1970s (Ashton, 1975; Kohlan, 1973; McKeachie et al., 1971; Pohlmann, 1975 all cited in Aleamoni and Hexner, 1980) indicated that female students evaluated instructors more highly in various areas, than male students in the same class. Arreola (1995) observes that there is no consistent view regarding the relationship between gender and student evaluations of instruction and Marsh and Roche (1997:1194) conclude that the gender issue has "mixed findings but little or no effect".

c. Level of the Course

More studies are consistent with the belief that the level of the course exerts some effect on student evaluations than not. Aleamoni and Hexner (1980) mention 8 researchers who found no meaningful relationship between the level of the course and student evaluations. Conversely, they cite 18 other investigators who concluded that higher level students (e.g. graduate students, 4th year students) tend to give higher evaluations to instructors than more junior level students (e.g. 1st year, 2nd year) (see Aleamoni and Hexner, 1980). Marsh (1977:1194) states that "graduate-level courses are evaluated somewhat more favorably [and that] weak, inconsistent findings suggest upper division courses are evaluated higher than lower division courses". This is probably not surprising in that students in smaller higher level courses are likely to be more dedicated and knowledgeable about the area of instruction and to receive more personal interactive forms of instruction. Aleamoni (1987) concludes that the level of the course should be considered when reviewing student evaluations.

d. Rank of the Instructor

Rank of the instructor appears to have little consistent effect on student evaluations. Arreola (1995) cites 5 studies that show that higher ranked instructors received higher evaluations and 5 studies that report no meaningful correlation between rank of the instructor and student evaluations (Arreola, 1995). Similarly, Aleamoni (1987) comments that there are some studies that report correlations between instructor rank and student evaluations, but both researchers agree that no consistent pattern has appeared in the literature. Again, Marsh and Roche (1997:1194) state that there have been "mixed findings but little or no effect".

e. Instructor Ideology and Values

There is little direct evidence regarding this issue. Two studies examined teacher's social-political attitudes and ideologies (Bausell & Magoon as cited in Feldman, 1987; Wilson et al., 1975) and found no relationship between these characteristics and evaluations of teaching effectiveness.

In addition to these specific studies, numerous studies have examined professorial personality and attitudes and how these relate to student evaluations. In a comprehensive review of the relation between professor personality and attitudes, Feldman (1986) found that professors' perceptions of their own personality were not related to student evaluations, but student perceptions of professorial personality were. Interestingly, colleague evaluations of personality were more strongly related to student evaluations of effectiveness than to self-reported personality. Erdle, Murray and Rushton (1985) provided preliminary evidence that the relationship between personality and student evaluations may be mediated by classroom behaviors.

Related to the personality research are studies examining attitude similarity. In a study comparing course evaluations with differences between perceived professor characteristics, and current and ideal self, Thomas, Ribich and Freie (1982) found that, as predicted, students whose current and ideal selves were closer to their perceptions of the professor also evaluated the course and professor more highly. Relationships between the ideal self and professor were stronger than those between current self and professor. In a similar study, Abrami and Mizener (1985:701) had students evaluate their own attitudes and perceived professor attitudes on a variety of topics. They found that although there was a significant relationship between perceived similarity and both evaluations of effectiveness and course grade, these relationships all but disappeared when professor effects were controlled for. In other words, perceived similarity, course grades and student evaluations were all predicted most efficiently by

who the instructor was. The authors conclude that "the validity of student evaluations is not substantially affected by student/instructor attitude similarity". Other research using similar methodology also supports this conclusion (Tollefson, Chen & Kleinsasser, 1989). Concerning attitudes, Feldman (1987) reports that professors who are perceived as more committed to undergraduate teaching and more student-centered in their approach tend to receive higher evaluations.

In conclusion then, research from a variety of sources and examining a variety of attitudes suggests that personality, attitudes (whether political, social, or regarding teaching) play a negligible role in determining student evaluations. If there is an effect it is mediated through either classroom behavior which is related to teaching effectiveness, or perceived similarity, which means there is not a consistent effect for all students.

f. Student Workload/Course Difficulty

Some faculty believe that the workload and the difficulty of the courses they teach have a significant effect on the evaluations they receive. Contrary to popular opinion, easy professors do not necessarily receive high student evaluations. Some research shows that students see demanding professors as being better (more effective) than easy professors, hence the higher evaluations. The research on this particular variable, however, has produced some surprising results. Marsh (1987:316) found that "higher levels of workload/difficulty were positively correlated with student evaluations" (p. 316) and therefore did not constitute a bias. In his 1977 overview, Marsh Marsh (1997:1194) concludes that "harder, more difficult courses requiring more effort and time are evaluated somewhat more favorably".

g. Other Extraneous Variables

Researchers have also studied the effects of other potential biases such as required versus elective courses and academic discipline.

The literature supports the belief that elective courses are evaluated more highly than required courses (Arreola, 1995; Marsh and Roche, 1997). Feldman (1978) found a small positive relationship between class evaluations and the students' average intrinsic interest (prior subject interest) in the subject area. Thus, required courses may receive lower evaluations simply because students are less interested in them. For this reason, it may be a good idea for faculty to include an item that assesses student interest in the course.

In addition, according to Marsh and Roche, courses in the sciences appear to be evaluated lower than courses in the humanities, but they describe this as a "weak tendency" and suggest that there have not been enough studies done to draw any firm conclusions. In summary, Marsh and Roche (1997) agree with McKeachie (1990:195) who, in turn, points out that although there are a number of variables that could potentially bias student evaluations of instruction, these variables have little effect. McKeachie (1990:195) says:

> Potentially contaminating variables such as...class size, or required versus elective classes, make a difference, but not a large enough difference to cause researchers to misclassify a good teacher as "poor." Although one should also get evidence from other sources if a teaching evaluation is to lead to an important personnel decision, student evaluations are the best validated of all the practical sources of relevant data.

Can student evaluations be used to improve instruction?

To determine the effects of student evaluations on the quality of teaching, Murray (1996) reviewed research evidence from three different sources: faculty surveys, field studies, and longitudinal comparisons. Let us now examine the evidence from these three sources more closely.

a. Faculty Surveys

Although the impact of student evaluations on instructional quality is not assessed directly by faculty surveys, they do provide a useful index of instructor beliefs regarding the issue. Murray (1996:5) reviewed the results from eight published surveys of faculty opinion from across the United States and Canada which included either one or both of the following questions: "Do student evaluations provide useful feedback for improvement of teaching?" and "Have student evaluations led to improved teaching?".

Although the findings differed somewhat between studies, generally, faculty participants agreed that student evaluations do lead to improvement in teaching (Murray 1996:5). In fact, "across all surveys reviewed...and with differential weighting according to sample size, 73.4% of respondents said that student evaluations provided useful feedback, and 68.8% said that student evaluations have led to improved teaching".

b. Field Studies

A study conducted by McKeachie et al. as cited in Murray, (1996) compared different groups of teachers who, half-way through the semester, received either a) a computer printout of student evaluations; b) a printout of student evaluations plus individual consultation with a faculty development "expert" who provided support and explicit suggestions for improvement or c) no student evaluations feedback (Murray, 1996:7). These conditions produced significant differences in their evaluations at the end of the semester. The "feedback-plus-consultation" group received the highest evaluations, the feedback-only group received the next highest evaluations and the no-feedback control group received the lowest evaluations. These findings led the investigators to conclude that (Murry, 1966:7) "student feedback alone led to modest improvement in perceived quality of teaching, whereas student feedback supplemented by expert consultation produced much larger gains in teaching".

Murray also cites meta-analyses of field experiments carried out by Cohen (1980) and Menges and Brinko (1986) that reached similar conclusions. Based on these findings, Murray has concluded that field experiments suggest that student evaluation (1966:9) feedback alone "leads to a modest improvement in faculty teaching performance," and student evaluation feedback "supplemented either by expert consultation or by clarification of specific teaching behaviors leads to more substantial gains in quality of teaching".

c. Longitudinal Comparisons

Comparisons of mean student evaluation scores longitudinally over a number of years after student evaluations have been used in a particular department or faculty have also been used to assess the long-term effects of evaluation feedback on teaching effectiveness. Murray (1996:11) notes that this approach is based on the assumption that if student evaluations do contribute to the improvement of teaching, then the effect should be reflected "in a gradual increase across years in the average teacher evaluation score of participating faculty members". The published research on longitudinal studies has produced mixed results. Some studies find a longitudinal improvement in mean student evaluations for the department or faculty as a whole and some do not. Murray concludes that the mixed results are due to the fact that most of the studies have not fulfilled all the methodological conditions necessary to provide meaningful results (e.g. the mean evaluations should be "compared across a minimum of 10 years or 10 semesters," and that the same student evaluation form should be employed for the duration of the study.

Murray (1996:21-22) has summarized his findings related to the effects of student evaluations on the improvement of teaching into four general conclusions:

1. Converging evidence from three independent sources, namely faculty surveys, field experiments, and longitudinal

comparisons, supports the view that student evaluation of teaching has contributed significantly to improvement of certain aspects of college and university teaching.

2. The contribution of student evaluation to improvement of teaching is greatly enhanced by expert consultation with instructional development specialists. This finding provides support for the positive impact of instructional development offices and programs in improving teaching. More research is needed to decide the most effective ways of combining student evaluation with expert consultation.
3. There is no clear evidence that student evaluation of teaching has led to negative side effects commonly attributed to it, such as grade inflation and entrenchment of traditional methods of teaching.
4. Evidence that student evaluation leads to significant improvement of teaching, in combination with research demonstrating the reliability and validity of student evaluation forms, provides strong justification for the use of student evaluation of college and university teaching, both as diagnostic feedback to faculty members and as one of several sources of information considered in decisions on faculty hiring, retention, salary, and promotion. However, since students are capable of assessing only some aspects of teaching, student evaluation should never be the only source of data on teaching in faculty personnel decisions.

Other studies have also provided evidence that student evaluations contribute to the improvement of teaching. Wilson as cited in Weimer & Lenze, (1997) conducted a study in which award-winning teachers were asked to characterize their teaching behaviors. Student evaluations were then carried out with a group of "teacher-clients." He consulted with his clients regarding their teaching evaluations and made specific concrete suggestions for improvement, including the teaching behaviors cited by the award-winning teachers. A second evaluation was conducted after an intervening semester. No difference was seen in the evaluations received by the comparison

group who received only student evaluation feedback but no consultation (Weimer & Lenze, 1997:209). For the teacher-clients who received such input, however, there was a "statistically important change in overall teaching effectiveness evaluations for 52 percent of the faculty clients". Furthermore, the Weimer & Lenze (1997:209) suggested that the "items on which the greatest number of faculty showed statistically important change were those for which the suggestions were most concrete, specific and behavioral" (ibid). Stevens and Aleamoni as cited in Weimer & Lenze (1997: 303) similarly reported that "provision of consultation in addition to student evaluations feedback resulted in an increase in student evaluations that was maintained over time" Weimer & Lenze 1997:209). Weimer & Lenze (1997: 303) suggest that more longitudinal research is needed in this area, and recommend that student evaluations feedback "must be integrated with a system of instructor training and available instructional support services".

Conclusions and Recommendations

Leading scholars in faculty evaluation research have commented on a number of important factors with respect to student evaluations. For example, McKeachie (1997:1223) suggested that a variety of student evaluation forms are necessary in order to account for the differences between the various modes of teaching prevailing today (e.g. the increasing use of technology, virtual universities, and cooperative learning). He also pointed out that researchers "need to study what teachers can do to help students become more sophisticated evaluators". Most importantly, McKeachie (1997:1223) argued for more research "on how to train members of personnel committees to be better evaluators, and research is needed on ways of communicating the results of student evaluations to improve the quality of their use". As noted numerous times throughout this paper, the literature clearly demonstrates that student evaluation forms that are psychometrically sound, are reliable, valid, relatively free from bias, and useful in improving teaching. Scriven's as cited in d'Apollonia and Abrami (1997b:19) made general conclusion that

"student evaluations are not only a valid, but often the only valid way to get much of the information needed for most evaluations". Marsh and Dunkin (1997: 311-312) conclude that despite "ill-founded fears" on the part of the faculty, and claims based on research "fraught with methodological weaknesses... the bulk of the research, however, has supported the continued use of student evaluations of instruction as well as advocating further scrutiny".

There are a number of points that should be taken into account when using students' evaluations either for teaching improvement or making personnel decisions. Sorcinelli (1999) enumerates these points as follows.

a) *Use multiple sources:* For whatever purpose results may be used, student evaluations represent only one source of information about teaching. Student evaluations should be supplemented by peer evaluations, alumni evaluations, self-evaluations, and portfolios containing descriptions of course materials, teaching methods, innovations, and students' pre- and post-test scores, and other evidence of teaching effectiveness.

b) *Obtain a sufficient number of evaluators*: At least 8 to 10 evaluators, preferably 15 or more, is the recommended number. The proportion of a class that evaluates an instructor is important. If a fifth or more of class members are absent or choose not to respond, then the results might not be representative.

c) *Use multiple sets of evaluations:* Evaluations from only one course or one term might not represent a teacher's performance (for course improvement, evaluations from a single course can be helpful). For personnel decisions, use five or more sets of evaluations taught over more than one semester.

d) *Take into account course characteristics:* Small classes (less than 15 students) often receive slightly more favorable

evaluations. Courses required by the university that are not a part of a student's major or minor tend to receive somewhat lower evaluations. Evaluation also may differ because of the nature of the course (e.g., humanities Vs. social or natural sciences). For each characteristic, the differences are slight, but together they might be significant.

e) *Rely more on summary items (e.g., how effective the course was overall) than on other items for personnel decisions:* Overall evaluations of the teacher or course tend to correlate higher with student learning than do specific diagnostic items. Therefore, decisions initially should focus on the overall evaluation items.

f) *For teaching improvement, use diagnostic items and written comments:* Summary items provide limited feedback; diagnostic items and students' written comments in response to open ended questions can help point to teachers' strengths and weaknesses. Although studies have shown that some teachers can improve after receiving evaluation results, change is more likely if a knowledgeable colleague or teaching improvement consultant can help interpret scores, provide encouragement, and suggest teaching improvement strategies. Centra (1994) also proposed the NVHM model, which states that at least four conditions must be fulfilled for student evaluations to lead to improvement in instruction: (1) **N--**instructors must learn something **n**ew from them; (2) **V--**instructors must **v**alue the new information; (3) **H--**they must understand **h**ow to make improvements; and (4) **M--**instructors must be **m**otivated to make the changes and improvements. He also pointed out that the knowledge gained from evaluations is most effective when there is a gap between how students evaluate the instructor and how the instructor evaluates him/herself.

g) *Use comparative data, but with caution:* Student evaluations tend to be favorable; comparative data (preferably local norms) provide a context within which faculty and administrators can

interpret individual reports. It is important not to over interpret; differences of less than 10 percentage points on any item or factor generally are not critical(e.g., 4.74 Vs. 4.79 on a 5-point scale).

h) *Employ standard procedures for administering forms in each class:* When results will be used in personnel decisions, it is critical. Someone other than the teacher (e.g., student, staff member) should distribute, collect, and return questionnaires to a central office. The teacher should not be present during the process. Ratings should be administered in the final week or two of the class, preferably not after or during a final exam. Evaluation results should not be returned to instructors until after they have reported grades for the course.

i) *Do not over use forms:* "Evaluation fatigue" may occur if ratings are too long or are required in every course in every term. For personnel decisions, a short form (4 to 6 summary items) should suffice. For improvement, a medium form (16 to 20 items) or a long form (30 to 36 items) is appropriate. A random or representative selection of courses is recommended, particularly for tenured professors.

References

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

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. **Review of Educational Research,** 52, 446-464.

Abrami, P. C., & Mizener, D. A. (1985). Student/Instructor attitude similarity, student evaluations and course performance. **Journal of Educational Psychology**, 77(6), 693-702.

Aleamoni, L. M. (1987). **Typical Faculty concerns about student evaluation of teaching. In Techniques for evaluation and improving instruction**. **New**

Directions for teaching and learning. L. M. Aleamoni (ed.). No. 31. San Francisco: Jossey-Bass.

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, 67-84.

Arreola, R. A. (1995). **Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators.** Bolton, MA: Anker Publishing Co.

Braskamp, L. A., and Ory, J. C. (1994). **Assessing Faculty Work**. San Francisco: Jossey-Bass Publishers.

Brown, J. (1998). 10 ways to get better student evaluations: 2 that may actually work. **Core Issues**, 8, 1-7. (York University newsletter).

Centra, J. A. (1993). **Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness**. San Francisco: Jossey-Bass.

Centra, J. A. (1994). **Current Issues in Evaluating and Improving College Teaching.** Paper presented at the annual meeting of the American Educational Research Association (AERA) meeting in Atlanta, April.

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

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student evaluations of college teaching: Reliability, validity, and usefulness. **Review of Educational Research**, 41, 511-535.

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

d'Apollonia, S., & Abrami, P. C. (1997b). In Response. **Change,** September/October, 18-19.

Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior and student evaluations of college teaching effectiveness: A path analysis. **Journal of Educational Psychology**, 77(4), 394-407.

Feldman, K. A. (1978). Course Characteristics and College Students' Ratings of their Own Teachers: *What We Know and What We Don't*. **Research in Higher Education**, 9, 199-242.

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(2), 139-213.

Feldman, K. A. (1989). The association between student evaluations of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multi-section validity studies. **Research in Higher Education**, 30, 137-194.

Frey, P. W. (1978). A two-dimensional analysis of student evaluations of instruction. **Research in Higher Education**, 9, 69-91.

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

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

Grush, J. E., & Costin, F. (1975). The student as consumer of the teaching process. **American Educational Research Journal**, 12, 55-66.

Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluation of instructors independently nominated as best and worst teachers by graduating seniors. **American Educational Journal**, 14, 441-447.

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., and Dunkin, M. J. (1997). “Students' evaluations of university teaching”. In R. Perry and J. Smart (eds.), **Effective Teaching in Higher Education: Research and Practice**. New York: Agathon Press.

Marsh, H. W., & Overall, J. U. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. **Journal of Educational Psychology**, 71, 856-865.

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

Marsh, H. W., & Ware, J. E. (1982*).* Effects of expressiveness, content coverage, and incentive on multidimensional student evaluation scales: New interpretations of the Dr. Fox effect. **Journal of Educational Psychology**, 74, 126-134.

McKeachie, W. J. (1990). Research on College Teaching: The Historical Background*.* **Journal of Educational Psychology**, 82, 189-200.

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

McKeachie, W. J., Lin, Y., & Mendelson, C. N. (1978). A small study assessing teacher effectiveness: Does reaming last? **Contemporary Educational Psychology**, 3, 352-357.

Murray, H. G. (1996**). "Does Evaluation of Teaching Lead to Improvement of Teaching?"** Submitted to International Journal of Academic Development,

Murray, H. G. (1983). Low inference classroom teaching behaviors and student evaluations of college teaching effectiveness. **Journal of Educational Psychology**, 71, 856-865.

Murray, H. G. (1984). *The impact of formative and summative evaluation of leaching in North American universities*. **Assessment and Evaluation in Higher Education**. 9, 117-131.

Phillips, P. (1998). Student views of student evaluations of teaching: Core Issues, York University newsletter, 8, 9-11.

Sorcinelli, M. D. (1999). The Evaluation of Teaching: The 40–Year Debate about Student, Colleague, and Self-Evaluations. In Pescosolido,B.A. and R. Aminzade,eds.,**The Social Worlds of Higher Education: Handbook for Teaching in a New Century.** California: Pine Forge Press

Thomas, D., Ribich, F., & Freie, J. (1982). The relationship between psychological identification with instructors and student evaluations of college courses. **Instructional Science** 11(2), 139-154.

Tollefson, N., Chen, J. S., & Kleinsasser, A. (1989). The relationship of students' attitudes about effective teaching to students' evaluations of effective teaching. **Educational and Psychology,** 49(3). 529-536.

Weimer, M. & Lenze, L. F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In R. Perry and J. Smart (eds.), **Effective Teaching in Higher Education: Research and Practice**. New York: Agathon Press.

Williams, W. M., & Ceci, S. J. (1997). *"How'm I doing?"* Problems with student evaluations of instructors and courses. **Change**, September/October, 13-23.

Wilson, R. C., Gaff, J. G., Dienst, E.R, Wood, L., Bavry, J.L. (1975). **College Professors and their Impact on Students**. New York: Wiley.