An Investigation into the Validity of Student Evaluations of Teaching in Accounting Education

Penelope J. Yunker
James A. Yunker

Western Illinois University

Abstract

Although there exists considerable dissent, the majority view among higher education professionals is that student evaluations of teaching are valid on the basis of a considerable body of research showing a positive correlation between student evaluations of faculty and objective measures of student achievement. The single most serious caveat concerning this evidence is that this correlation may be an associative relationship (rather than a causative relationship) generated by underlying variables such as student ability and interest in the subject matter. It seems plausible that capable students with a high level of interest in the subject matter will both do better on examinations, and take a more favorable view of the efforts of their instructors. Much of the cited research showing a positive correlation between student evaluations and student achievement either fails to control for student ability/interest at all, or controls for it by means which may be inadequate (such as random assignment of students to classes). The present research examines the relationship between student achievement (as measured by the grade which the student earned in Intermediate Accounting I) and student evaluations (as measured by the mean class evaluation of the student's instructor in Introductory Accounting II), controlling for ability with three variables: grade which the student earned in Introductory Accounting II, student's Grade Point Average, and student's ACT score. Controlling for ability, a statistically significant *negative* relationship is found between student evaluations and student achievement. This research therefore strengthens concerns of some faculty that the widespread and significant application of student evaluations of teaching in faculty performance evaluation may be partially responsible for the interrelated phenomena of grade inflation and content debasement.

Background

Student evaluations of teaching (SEs) have been institutionally entrenched within the U.S. higher education system for a period of decades. In this respect, accounting higher education is typical of higher education as a whole. On the basis of a 1986 survey of 241 heads of accounting departments, Yunker and Sterner (1988) reported that 90 percent of departments administer "formal student evaluations" for purposes of administrative evaluation of faculty teaching effectiveness. The mean weight attributed by the respondents to this type of evidence (on a scale from 1 to 5) was 4.170. For comparison, the next highest mean weight (on a list of eleven possible types of evidence) was that of "informal student evaluations": 2.402. A 1993 survey of 84 accounting department heads by Calderon and Green (1997) found 95 percent administering student

evaluations of teaching. Some 82 percent of respondents reported using additional types of evidence for purposes of evaluating teaching effectiveness, but while the survey listed nine other types of evidence, the mean number of additional types utilized was only three. It would appear that not only are SEs almost universally utilized in accounting higher education, they are in fact the dominant source of information on faculty teaching effectiveness.

Although student evaluations of teaching have long been a fact of life at almost all institutions of higher education, research on the reliability, validity and utility of SEs continues steadily. The ongoing research affects the weight and practical importance attached to SEs in the real-world performance evaluation process. Research which supports the reliability, validity and utility of SEs tends to increase their practical significance in faculty performance evaluation, while research which does not tends to have the opposite effect.[1] The three issues of reliability, validity and utility of student evaluations of teaching, although conceptually distinct, are highly interrelated in practice. Reliability refers to the degree to which student evaluations of a specific teacher are stable over time. Validity refers to the degree to which student evaluations are related to objective measures of student achievement. Utility refers to the degree to which student evaluations are useful in faculty performance evaluation. Clearly, to the extent that student evaluations are more reliable and more valid, they are more useful in performance evaluation. Also, validity is clearly the critical issue. Student evaluations of teaching could be extremely reliable, but if they lack significant validity, their utility for purposes of faculty performance evaluation would be minimal.

Literature Review

Justifiably, therefore, a very large quantity of empirical research addresses the validity of SEs. The best-known and most highly influential survey of this research is an article by Peter A. Cohen, based on a 1980 Ph.D. dissertation at the University of Michigan, published in the ***Review of Educational Research*** in fall 1981.[2] Based on a "meta-analysis" of 41 independent validity studies covering 68 different experiments meeting certain specified criteria (research based on actual courses, unit of analysis the section or instructor rather than the individual student, achievement measured with a common testing instrument), Cohen estimated the overall simple correlation coefficient between student evaluations of teaching and student achievement to be +.43. On this basis, he concluded that SEs possess a substantial degree of validity, and thereby a substantial degree of utility in faculty performance evaluation.

Although this conclusion could probably be described as the consensus viewpoint within the contemporary U.S. higher education system, a substantial amount of skepticism remains.[3] The single most serious concern is that an observed positive correlation between student evaluations and student achievement may be associational in

[1] An indication of the volume of literature on SEs is the fact that under the subheading "Student Evaluations of Teacher Performance" in the OVID search engine of the ERIC database for 1984-1999, restricting to English language and journal publications, there are 659 references.

[2] To this day, the Cohen study remains the largest single survey devoted solely and specifically to the basic validity question of whether overall student ratings of teachers are positively related to objective measures of student achievement. A smaller-scale but more recent survey of validity studies in that of Abrami (1990). The validity issue is of course a major sub-topic in several large-scale general surveys of the SE literature both before and after Cohen: Costin et al (1971), Marsh (1987), Feldman (1989), and Cashin (1995). In the most recent large-scale survey of the empirical accounting literature by Rebele et al. (1998b), some accounting literature on SEs is cited (pp. 191-193), but none pertaining to validity. However, Rebele et al. do describe at some length Cashin's positive conclusions, based on the general education literature, concerning the probable reliability, validity and utility of SEs.

[3] A few illustrative references to the recent anti-SE literature include Carey (1993), Goldman (1993), Kishor (1995), Fram (1996), Trout (1997), Williams and Ceci (1997), Wilson (1998), and Martin (1998).

nature rather than causative. The presumption of SE validity depends on two conditions: (1) there is a causative relation between high quality teaching and high student achievement; (2) there is a causative relation between high quality teaching and high student evaluations of teaching. But it has long been recognized that an observed positive relationship between student evaluations and student achievement may be the result of an underlying unobserved variable—other than actual high quality teaching - which influences both student achievement and student evaluations in the same direction. One obvious candidate for this variable is the ability and/or attitude of the individual student. It seems plausible that more capable and enthusiastic students would both do better on examinations *and* take a more positive view of the efforts of their instructors. Self-selection might amplify this tendency over time, as better students gravitate toward instructors reputed to be superior on the basis of past evaluations. While student ability/attitude is the most obvious factor that might create an "artificial" positive relationship between achievement and evaluations, other factors might have the same effect, such as personal characteristics of students aside from ability/attitude, whether the course is required or elective, the difficulty of course material, class size, the hour of the class, and so on.

The desirability in principle of controlling for student ability is taken for granted by researchers, and is the main reason for the current consensus that class means are the appropriate units of measure, rather than individual students, for purposes of SE validity studies. Obviously there will be less variance in student ability across classes than across individual students. Both the Cohen survey cited above, and a contemporaneous survey by David A. Dowell and James A. Neal (***Journal of Higher Education***, 1982), discarded all validity studies based on individual student observations on this methodological ground. But a certain amount of subjective judgment is involved in specifying whether or not a given study "properly controls" for student ability. For example, one researcher might accept that random assignment of students to classes properly controls for differences in student ability (and other possibly relevant student characteristics); while another might not, since even after random assignment, there will be some variation in student ability as measured by class mean. Cohen deemed that 25 of the 68 experiments included in his meta-analysis (approximately 37 percent) properly controlled for mean student ability. Dowell and Neal, in their meta-analysis based on essentially the same literature, found only six studies which they deemed to have met this criterion. Dowell and Neal were much more skeptical on the question of SE validity than Cohen. They included in their six studies one by Rodin and Rodin (1972) which found a statistically significant negative relation between student evaluations and student achievement, and stressed the fact that even among the five studies finding a positive relation between evaluations and achievement, the partial correlation between the two, controlling for student ability, was extremely small.[4]

A survey article on SEs in the ***Journal of Accounting Education*** (Penny Wright et al., 1984) found three empirical studies in the general education literature which were deemed to have properly controlled for student ability and which showed a positive

[4] The number of original validity studies examining the relationship between student evaluations and student achievement seems to be thinning out in the more recent literature, perhaps owing to the general consensus that it has been sufficiently well established that the relationship is positive and statistically significant. A relatively rare recent example of the type of study typical of those cited in Cohen's 1981 survey is Koon and Murray (1995). This study finds in favor of the validity of SEs on the basis of a statistically significant +.30 simple correlation coefficient between mean overall teacher effectiveness rating and mean final exam performance. The sample consisted of 36 faculty teaching a freshman level psychology course, none of the faculty being graduate students. The control exercised for prior student ability and interest was random assignment of students to classes. But no data were shown to demonstrate that prior student ability and interest were indeed approximately constant over classes.

relation between student performance on final examinations or standardized tests and student evaluations of faculty. In view of the low partial correlations found, this was not deemed a sufficient weight of evidence to justify uncritical use of SEs. Wright et al. concluded: "The validity of student ratings, as evidenced by their association with learning criteria, has generally been found to be weak. Student ratings appear to measure a host of factors, such as the instructor's personality and expressiveness, which may or may not relate to students' learning." The questionable validity of student evaluations of teaching was also invoked by James D. Newton (1988) in urging that considerable caution and discretion be exercised by accounting program administrators in drawing inferences about actual teaching ability from student evaluation (SE) ratings.

These authors, therefore, emphasized the issue of the numerical size of the relationship between student evaluations and student achievement. It is apparent both from casual empiricism and the research literature that a great many factors affect student evaluations of teaching apart from actual teaching effectiveness: these other factors include student characteristics (attitude and/or ability, and so on), personal faculty characteristics which relate indirectly or not at all to teaching effectiveness (age, gender, race, personality, and so on), and environmental conditions (hour of class, and so on).[5] Some of these other factors may affect student achievement, others may not. Even taking Cohen's simple correlation coefficient of +.43 at face value, it implies an R-squared of approximately .185, indicating that only 18.5 percent of the variance of student achievement scores is statistically explained by variance in student evaluation ratings. Dowell and Neal argued that the partial correlation coefficient, controlling properly for other variables affecting student achievement scores, might well be much lower than this.

Other problems exist with studies cited by Cohen showing a positive relation between SEs and student achievement measures apart from possible misleading conclusions owing to inadequate controls for student ability. The typical study included in Cohen's meta-analysis involved graduate assistants teaching sections of introductory courses, all of which took a common final examination prepared by the investigator, normally a senior-level faculty member. No doubt a competitive condition existed among the instructor "subjects," all of whom would likely expend an unusual amount of time and effort to impart knowledge so that their students would do well on the examinations and to maintain as friendly relations as possible with their students to improve their student evaluations. In addition, there are obvious practical constraints on the generosity graduate assistants may exercise in awarding grades. This rather artificial experimental situation might not provide a satisfactory approximation of the way student evaluations are normally implemented in actual practice. It is well established that, all other things being equal, students expecting higher grades in a course will tend to rate their teachers more highly. In light of this, it stretches credulity to propose that it was coincidental that a very substantial grade inflation occurred in U.S. higher education at about the same time that SEs came into wide application for purposes of administrative decision-making concerning faculty retention, tenure, promotion, and salary adjustments.[6] The observed

[5] Some illustrative research on factors affecting student evaluations of teaching aside from teacher effectiveness is as follows. On the effect of gender: Feldman (1992, 1993), Andersen and Miller (1997), Fernandez and Mateo (1997). On the effect of age and/or personality: Feldman (1983, 1986), Renaud and Murray (1996). On the effect of class level: Sailor et al (1997). On the effect of controversial course content: Ludwig and Meacham (1997). A considerable proportion of the relatively limited amount of empirical research on SEs in accounting education journals concerns determinants of overall ratings of teachers: Hooper and Page (1986), Kreuze and Newell (1987), Mulford and Schneider (1988), Rahman et al. (1988), Deberg and Wilson (1990), and Briscoe et al. (1996). The last two contributions look at type of course (e.g., judgment-based versus standards-based), and find that all other things being equal, courses requiring more judgment and less memorization tend to display lower student evaluations of teaching.

[6] There is a large literature on the relationship between grades and student evaluations of teaching, and the possibly

phenomenon of grade inflation raises concerns about a possibly unobserved phenomenon of content debasement. To a faculty member interested in securing higher student evaluations, grade inflation (giving higher grades for a given amount of knowledge) and content debasement (reducing the amount of knowledge necessary to achieve a given grade) are close substitutes. Unlike the instructor subjects of the SE validity experiments, most real-world faculty members are highly autonomous with respect to what they teach, what they expect students to know for examinations, and what criteria they use to convert numerical examination scores into letter grades.[7] These considerations suggest that in some non-experimental situations there might be a negative, rather than a positive, relationship between student evaluations of teaching and student achievement, controlling for student ability. This research does in fact point in this direction.

To the authors' knowledge, this study is the first to look at the SE validity question specifically in the context of accounting education. The quantity of research on SEs in accounting education has apparently been fairly modest. Between the three large-scale literature survey articles by Rebele et al. (1991, 1998a, 1998b), cumulating to 184 pages and covering empirical research published in the major accounting education journals over the 1985-1997 interval, only nine journal articles pertaining to student evaluations of teaching are cited, and none of them address the validity question. It is legitimate for accounting educators to give special consideration to SE validity research in their own area because of the possibility that the empirical relationship between student evaluations and student achievement might vary systematically across disciplines.[8] Yunker and Marlin (1984) suggested that there is more likely to be a positive relationship between evaluations and achievement in disciplines which students consider to be important to their future success. For example, computer programming might be considered by most business students to be more important than economics, which could explain the presence of a positive relationship between evaluations and achievement in courses in computer programming, and the absence of such a relationship in courses in economics.[9] The subjects in the present experiment were accounting majors enrolled in accounting courses, so it would seem that they would have to deem the subject matter to be important. This aspect of the research would seem to favor the finding of a positive relationship between student evaluations and student achievement. However, as noted, a

related phenomena of content debasement and grade inflation. See, for example: Feldman (1975), Bejar and Blew (1981), Nelson and Lynch (1984), Goldman (1985), Rustagi (1997), Greenwald and Gillmore (1997), Hardy (1997), and Beaver (1997). Interestingly for accounting educators, a study by Cluskey (1997) finds that accounting has been relatively immune from grade inflation. However, another study by Addy and Herring (1996) pertaining to accounting education, indicates that following the administrative implementation of a minimum grade point average for graduation, faculty raised the grades of the less capable students to enable them to graduate. Although this study does not examine the effect of administrative application of SEs on grades, it does suggest that accounting education is perhaps not altogether immune from tendencies toward grade inflation. Mixed results have been obtained by accounting education researchers on the relation between expected grades and student evaluations of teaching: Deberg and Wilson (1990) and Mulford and Schneider (1988) find a positive relationship, while Kreuze and Newell (1987) do not.

[7] One possibly indicative study by O'Connell and Dickinson (1993) finds that while the correlation between objective amount learned and student evaluations is very low, the correlation between perceived amount learned and student evaluations is very high. Another study by Brodie (1998) indicated that professors assigning the highest grades for the least studying received the highest student evaluations.

[8] According to Wright et al. (1984): "Few research studies addressing these issues [reliability and validity] have been conducted in accounting program SEtings... Accounting programs may be made up of courses and students which are more homogeneous than the SEtings in which most validity research has been conducted. This justifies the replication of these studies using accounting courses..."

[9] As explained by Yunker and Marlin (1984), this outcome has to do with the possibility that knowledge in a specific discipline (such as economics), deemed by the student to be unimportant to his/her future success, is thereby a "non-superior good." Encountering a highly effective instructor in economics, the student will not utilize the high teaching effectiveness of the instructor to learn more about economics, but will rather learn the same limited amount needed to earn the desired grade, and will take the benefit in spending more time on other subjects (such as computer programming) deemed by that student to be more important to future success. Thus the absence of a positive evaluation-achievement relationship in economics courses.

negative relationship was found (controlling for student ability).

Research Design

The material coverage in a standard accounting program is highly cumulative, meaning that later courses build directly on concepts and procedures developed in earlier courses. Therefore the level of achievement of students in later courses is directly dependent on the level of their achievement in earlier courses. Two courses which are especially closely related in the accounting program at the authors' university (a regional Midwestern school) are Introductory Accounting II and Intermediate Accounting I. The great majority of students who take Intermediate Accounting I are accounting majors who have taken Introductory Accounting II the previous semester. As chair of the Department of Accountancy at this university, the principal investigator in this research has convenient access to both student and faculty records. Student records maintained by the registrar's office cover course grades and measures of overall academic achievement. Faculty records maintained by the department of accountancy cover student evaluation ratings for all courses given. The strategy of the research is to look at the relationship between achievement of students in Intermediate Accounting I and the following independent variables: (1) mean instructor rating in the Introductory Accounting II course taken the previous semester, (2) various measures of student ability/achievement. Most educators would probably regard it as a strength of the present research that the achievement variable utilized represents "long-term learning" rather than "short-term learning."

By comparing course grade sheets, a sample was obtained of 283 students who had taken Intermediate Accounting I sometime between the fall semester of 1991 and the fall semester of 1998, and who could be traced to a specific section of Introductory Accounting II in the previous semester. Some 46 sections of Introductory Accounting II were involved, taught by 12 different faculty members: 6 full professors, 1 assistant professor, and 5 instructors. The breakdown of the 46 sections by rank of teacher was 22 by full professors, 10 by the assistant professor, and 14 by instructors. For each student, the grades were recorded in Intermediate Accounting I and Introductory Accounting II along with the mean overall rating given by the class to that student's teacher in Introductory Accounting II. Additional information obtained on each student included the latest cumulative Grade Point Average and the ACT score reported on the admission application. For students taking Intermediate Accounting I earlier in the 1990s, the GPA was the final GPA recorded upon graduation. For more recent students who had not yet graduated, the GPA was that recorded as of fall semester 1998. The ACT score was not available for a substantial proportion of the sample: Of the 283 total students in the sample, ACT scores were obtained for 183. Table 1 provides salient information on the five variables utilized in this research.

A word is required on the use of individual students as the unit of observation in this research. As mentioned above, the consensus among education researchers is that class mean observations are preferable to individual student observations for purposes of validity studies of student evaluations of teaching. Of course, this consensus violates a fundamental statistical precept that the investigator should not discard potentially useful information. Clearly, averaging variables over classes discards a great deal of individual student data. Possibly, therefore, the consensus is in error. The reason for the consensus in this particular context is that it is regarded as highly plausible that high-ability students will tend to rate the same teacher more highly than will low-ability students. Therefore, if a positive association is found between student achievement and student evaluations

using students as observations, this association could be spurious, i.e., an artifact resulting from the fact that high-ability students tend simultaneously to attain high achievement and to rate their teachers highly. Obviously there will be less variation in student ability over classes than over individual students; hence (according to the consensus) use of class means effectively controls for this potential bias.

The problem with this logic is that if data exists on individual students, a more efficient procedure for purposes of controlling for the effect of student ability would be to utilize a multivariate statistical technique such as multiple regression or partial correlation. The question of interest is: "Do students *of a given ability level* who rate their instructors more highly tend to do better on achievement tests?" The way to answer this question is to regress student achievement scores on student evaluations of teaching *and* measures of student ability, or to obtain a partial correlation coefficient between achievement and evaluation, holding ability constant.

It should be emphasized, however, that the present study does *not* in fact utilize individual student data on the independent variable of interest (RATING). The "rating" referred to is not the rating applied by the student himself or herself to the teacher in Introductory Accounting II, but rather the mean rating applied by the entire class in which the student was enrolled. In fact, we do not have data on the rating applied by the individual student to his/her teacher in Introductory Accounting II; otherwise individual student data would have been used for this variable. But the positive side of not having individual student data on the RATING variable is that this research does not have to confront the above-described consensus among education researchers, at least as far as this particular variable is concerned. For purposes of comparison, however, the principal regression equation of the research, reported in Table 3 below, is estimated on the basis of both individual student observations and class mean observations.

Findings

Table 2 shows a simple correlation coefficient matrix for the five variables involved in the study. The correlations between ACT score and other variables are based on the 183 cases for which data was available on ACT score. The other correlations are based on all 283 cases in the sample. As expected, the correlations among the two grade variables GRINTER1 and GRINTRO2 and the two general academic achievement variables GPA and ACT are positive and fairly large. The simple correlation coefficient between the achievement variable of interest (grade in Intermediate Accounting I = GRINTER1) and the faculty evaluation variable RATING is positive, but very small and statistically insignificant.

In this particular study, the grade achieved in Introductory Accounting II (GRINTRO2) is considered a control variable rather than the achievement variable of interest. However, in view of the fact that at the authors' university the major examinations in this course are common to all sections, and grades are mostly dependent on performance on these objective examinations, the grade in Introductory Accounting II is fairly close to a standard achievement variable in a typical SE validity study. It is interesting to note, therefore, that the simple correlation coefficient between RATING and GRINTRO2 is positive and sufficiently large to be significant at a high level of confidence. It is also true that using class mean observations on all variables rather than individual student observations (which reduces the sample size to the 46 sections involved in the study), the simple correlation coefficient is considerably larger (.3945 rather than .1856) and continues to be statistically significant at a high level of confidence despite the reduction in sample size. This evidence is therefore comparable to

evidence obtained by many of the SE validity studies cited by Cohen as supportive of the validity of SEs.

Considering GRINTRO2 as a grade variable, as opposed to an achievement variable, this evidence is consistent with the large body of research showing a positive relationship between student grades and instructor ratings. The observed positive grade-rating correlation is interpreted differently by supporters and skeptics of SEs. Supporters maintain that it is reasonable that high-quality teaching produces students who earn high grades. Skeptics see the causation running in the opposite direction and suggest that by means of generous grading a faculty member produces students who rate him/her more highly.

Table 3 presents results from a multiple regression analysis of the data. Columns (1) through (4) enter in succession: first the principal independent variable of interest RATING, and then successively the three ability-achievement control independent variables GRINTRO2, GPA, ACT. The order of the control variables reflects their "closeness" to achievement in Intermediate Accounting II: first achievement in Introductory Accounting II, then overall achievement at the college level (GPA), and finally pre-college ability-achievement (ACT). Column (5) shows the re-estimation of column (4) using class mean observations rather than individual student observations. Column (6) shows the application of partial correlation to check the multiple regression result shown in column (4).

Column (1) is consistent with the simple correlation result between GRINTER1 and RATING shown above in Table 2: not controlling for ability-achievement, there is a very small and statistically insignificant positive relationship between the two variables. As the ability-achievement control variables are entered in columns (2) through (4), the relationship between GRINTER1 and RATING becomes negative and statistically significant at the 95 percent confidence level in columns (2) and (4). In column (3), containing GRINTRO2 and GPA as the control variables, the t-value is −1.66, which is significant at the 90 percent confidence level which, while below the customary 95 percent confidence level, is still quite high. Column (6) shows the partial correlation coefficients and corresponding t-values for each of the independent variables in column (4), holding constant the other independent variables. As expected, the t-values for the partial correlation coefficients duplicate the t-values for the multiple regression coefficients.

Column (5) estimates the same multiple regression equation shown in column (4), but does so not on the basis of 183 individual student observations on all variables except for RATING (which is a class mean), but on the basis of 46 class mean observations on all variables. Although the signs of the estimated regression coefficients are the same as in column (4), the t-values have all been reduced substantially, and only GRINTRO2 and GPA remain statistically significant at the 90 percent level of confidence. It was argued above, that for purposes of controlling for prior ability-achievement in SE validity studies, it is methodologically superior to utilize individual student observations (when student data is available) than to average all student data into class means. A comparison of columns (4) and (5) in Table 3 serves as support for this argument. Presumably all three of the ability-achievement control variables must be positively related to the dependent variable of achievement in the Intermediate Accounting I course. Column (4) shows this to be the case, while column (5) does not, owing to the much smaller number of observations used in the estimation of the latter regression equation.

Discussion

The fundamental finding of this research is that, controlling for ability-achievement at the individual student level, students in Intermediate Accounting I who have been in Introductory Accounting II courses in which the teacher has been rated more highly tend to do worse than students who have been in Introductory Accounting II courses in which the teacher has been rated less highly. The negative effect of student evaluation rating in the Introductory II course on student achievement in the Intermediate I course is not large in a numerical sense, but it is highly significant in a statistical sense. Assuming the correctness of the statistical analysis, what might account for such a result?

Frequently mentioned in the education literature is "teaching style bias" as a possible confounding factor in validity studies. Some teachers in a study might orient their teaching style to the less capable students: these students will rate the instructor more highly, yet because of their academic deficiencies do poorly on achievement tests, leading to a negative relation between evaluations and achievement. Other teachers might orient their teaching style to the more capable students: these students will rate the instructor more highly, and because of their academic abilities will do well on achievement tests, leading to a positive relation between evaluations and achievement. If a given study contains both types of teachers in approximately equal proportions, the two effects will cancel out, leading to the invalid conclusion that there is no relationship between evaluations and achievement. Since the indication here is a negative relationship between student evaluations and student achievement (controlling for student ability), an implication might be that at the authors' university, teachers in the Introductory Accounting II course tend to orient their teaching styles toward the less capable students.

Of course, one means of orienting one's teaching style to the less capable students would simply be to reduce the amount of material covered in the course: to use an uncharitable term, "content debasement." Discussions of "teaching style bias" in the education literature tend to assume that the material is the same whatever the teaching style. But in the academic real world, the amount of material, not just the manner in which it is presented, is controlled by the teacher. The highly capable students would not necessarily recognize that they are being short-changed as far as amount of material is concerned and thereupon rate the teacher lower because their intellects are not being adequately challenged. How often does one hear students, even highly capable students, complaining about how easy a particular course is? Quite possibly, both more capable and less capable students would react to content debasement in the same way: by being pleased that they are apparently learning a considerable amount with modest effort, attributing the situation to the high teaching effectiveness of the teacher and thereupon rating the teacher more highly. Evidence demonstrating a negative relationship between student evaluations and student achievement, controlling for student ability, as obtained here, is compatible with the potential efficacy of content debasement as a means of improving student evaluations.

In the case of this particular research, opportunities for deliberate content debasement by the teachers in Introductory Accounting II at the authors' university are actually fairly limited. This course uses a common textbook as well as common examinations, and while faculty teaching the course design their own syllabi, topics to be covered are specified by the principles sub-committee of the departmental curriculum committee. Teachers cannot simply eliminate topics from their syllabi. However, they can control the level of rigor with which these topics are presented, and the manner in which they are presented (e.g., straightforward expository lecturing to students versus

give-and-take Socratic dialogue with students). Although students take common examinations in Introductory Accounting II, examination scores are not compared across instructors and utilized for purposes of evaluating teaching effectiveness. Faculty teaching the course also determine what proportion of their contact time with students is devoted to serious consideration of accounting issues versus friendly personal interaction. Much of what favorably disposes students toward faculty members and may induce them to higher ratings of these faculty, may not be of much importance to the intellectual progress of the students.

The allegation of organizationally dysfunctional gamesmanship on the part of accounting faculty seeking higher student evaluations has recently been forcefully stated by James R. Martin (1998). Martin's proposal that SEs be scrapped is clearly politically infeasible, and his constructive suggestions on the proper functioning of the higher education system, based on the thinking of the management guru, W. E. Deming, have a definite air of utopianism about them. What probably happens in accounting departments in which skepticism regarding SEs is dominant is that they are collected and reported, but virtually ignored in practical decision-making regarding faculty performance—unless the ratings are either extraordinarily high or extraordinarily low.

Another seemingly more forward-looking possibility would involve increasing the relative weight of other indicators of teaching performance apart from student evaluations (such as teaching materials: syllabi, lecture notes, handouts, and so on; student examination papers; solicited alumni evaluations; peer evaluations based on classroom visitations); and/or changing the fundamental nature of student evaluations of teaching. These approaches have been recommended by Calderon et al. (1996), under the putative authority of the American Accounting Association.[10] With respect to the latter, Calderon et al. propose dropping the subjective overall rating item from the evaluation form (the item worded along the following lines: "Rate this instructor relative to other instructors you have had at this university"), and instead using the mean rating over a series of objective items ("questions to which students can competently respond"), all of which are presumably positively related to teaching effectiveness. Examples of such items would include (paraphrasing Calderon et al.): "Rate the instructor's organization and preparation for class," "Rate the instructor's enthusiasm for teaching and concern for students' progress," "Rate the instructor's fairness in grading practices," "Rate the extent to which examinations were related to the material covered," "Rate the extent to which the instructor increased your level of interest in the subject matter," "Rate the contribution of the instructor to your understanding of the subject matter." In addition, Calderon et al. propose that the questionnaire also collect information on student and course characteristics which plausibly affect student evaluations of teaching (expected grade, required versus elective, gender, perception of workload, etc.) and that the final faculty rating be adjusted for these factors. While the Calderon et al. recommendations on the nature of SEs are unobjectionable in principle, they would obviously be very difficult to implement successfully in practice.[11]

---

[10] The article by Calderon et al. (1996) containing these recommendations is described as the report of the Promoting and Evaluating Teaching Effectiveness Committee of the American Accounting Association Teaching and Curriculum Section (Thomas G. Calderon, Chair).

[11] One might contemplate an estimated regression equation in which the explanatory variables were factors affecting student evaluations which are not obviously and directly related to teaching effectiveness (teacher gender, teacher age, class level, course type, etc.). This equation could be used to produce an estimated rating for a particular course. The adjusted rating of the faculty member would then be his/her actual rating less the estimated rating. The adjusted rating would be far more meaningful as a measure of teaching effectiveness than the actual rating. To our knowledge, no formal adjustment procedures have been applied in practice. Those engaged in faculty performance evaluation make informal

Conclusion

In conclusion, this research points toward potential invalidity of student evaluations of teaching in accounting education, and suggests that they be applied cautiously in faculty performance evaluation. While it is conceded that the majority of past validity studies on SEs have supported their validity, it could be that these studies have not adequately controlled for student ability-achievement (and other possible factors) which may affect the relationship between student evaluations and student achievement. In addition, accounting as an academic discipline might be intrinsically different from disciplines in which a positive relationship has been found between student evaluations and student achievement. Still another possibly relevant aspect of the present experiment is that, unlike the typical SE validity study which looks at the relationship between student evaluations and student achievement in the same course, this study looks at the relationship between student evaluations and student achievement in the following course in a standard accounting sequence ("long-term learning" rather than "short-term learning"). Additional research on the SE validity issue in accounting education would probably be helpful.

References

Abrami, P. C. (1990). Validity of student ratings of instruction: what we know and what we do not know. ***Journal of Educational Psychology*** 82 (June): 219-231.

Addy, N., and C. Herring (1996). Grade inflation effects of administrative policies. ***Issues in Accounting Education*** 11 (Spring): 1-13.

Andersen, K., and E. D. Miller (1997). Gender and student evaluations of teaching. ***Ps: Political Science and Politics*** 30 (June): 216-219.

Beaver, W. (1997). Declining college standards: it's not the courses, it's the grades. ***College Board Review*** 181 (July): 2-7.

Bejar, I. I., and E. O. Blew (1981). Grade inflation and the validity of the Scholastic Aptitude Test. ***American Educational Research Journal*** 18 (Summer): 143-156.

Briscoe, N., G. W. Glezen, and W. C. Letzkus (1996). The association of accounting course content groupings and student evaluations. ***Accounting Educators' Journal*** 8 (Fall): 14-26.

Brodie, D. A. (1998). Do students report that easy professors are excellent teachers? ***Canadian Journal of Higher Education*** 28: 1-20.

Calderon, T. G., A. L. Gabbin, and B. P. Green (1996). Summary of promoting and evaluating effective teaching: American Accounting Association Teaching and Curriculum Section. ***Journal of Accounting Education*** 14 (Fall): 367-383.

Calderon, T. G., and B. P. Green (1997). Use of multiple information sources in assessing accounting faculty teaching performance. ***Journal of Accounting Education*** 15 (Spring): 221-239.

Carey, G. W. (1993). Thoughts on the lesser evil: student evaluations. ***Perspectives on Political Science*** 22 (Winter): 17-20.

Cashin, W. E. (1995). Student ratings of teaching: the research revisited," IDEA Paper No. 32, Center for Faculty Evaluation and Development in Higher Education, Kansas State University (September).

Cluskey, G. R., Jr. (1997). Accounting grade inflation. ***Journal of Education for Business*** 72 (May-June): 273-277.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of

---

adjustments of student evaluations, adjustments which are no doubt subject to a very large margin of error.

multisection validity studies. ***Review of Educational Research*** 51 (Fall): 281-310.

Costin, F., W. T. Greenough, and R. J. Menges (1971). Student ratings of college instructors: reliability, validity and usefulness. ***Review of Educational Research*** 41 (December): 511-535.

Deberg, C. L., and J. R. Wilson (1990). An empirical investigation of the potential confounding variables in student evaluations of teaching. ***Journal of Accounting Education*** 8 (Spring): 37-42.

Dowell, D. A., and J. A. Neal (1982). A selective review of the validity of student ratings of teaching. ***Journal of Higher Education*** 53 (January/February): 51-62.

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. ***Research in Higher Education*** 4: 69-111.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. ***Research in Higher Education*** 18: 3-124.

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: 139-213.

Feldman, K. A. **(**1989). The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. ***Research in Higher Education*** 30 (December): 583-645.

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 (June): 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 (April): 151-211.

Fernandez, J., and M. A. Mateo (1997). Student and faculty gender in ratings of university teaching. ***Sex Roles: A Journal of Research*** 37 (December): 997-1003.

Fram, E. H. (1996). Marketing, higher education, and student responsibility. ***College Board Review*** 179 (November): 2-5.

Goldman, L. (1985). The betrayal of the gatekeepers: grade inflation. ***Journal of General Education*** 37: 97-121.

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 (Spring): 57-64.

Greenwald, A. G., and G. M. Gillmore (1997). No pain, no gain? the importance of measuring course workload in student ratings of instruction. ***Journal of Educational Psychology*** 89 (December): 743-751.

Hardy, L. (1997). Grade inflation. ***American School Board Journal*** 184 (December): 28-30.

Hooper, P., and J. Page (1986). Measuring teaching effectiveness by student evaluation. ***Issues in Accounting Education*** 1 (Spring): 56-64.

Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgment: consequences for the validity of student ratings of instruction. ***Research in Higher Education*** 36 (April): 177-195.

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

Kreuze, J., and G. E. Newell (1987). Student ratings of accounting instructors: a search for important determinants. ***Journal of Accounting Education*** 5 (Spring): 87-98.

Ludwig, J. M., and J. A. Meacham (1997). Teaching controversial courses: student evaluations of instructors and content. ***Educational Research Quarterly*** 21 (September): 27-38.

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

Martin, J. R. (1998). Evaluating faculty based on student opinions: problems, implications, and recommendations from Deming's theory of management perspective. ***Issues in Accounting Education*** 13 (November): 1079-1094.

Mulford, C. W., and A. Schneider (1988). An empirical study of structural and controllable factors affecting faculty evaluations. In Bill N. Schwartz, ed., ***Advances in Accounting***, Volume 6 (Greenwich, Ct.: JAI Press): 205-215.

Nelson, J. P., and K. A. Lynch (1984). Grade inflation, real income, simultaneity, and teaching evaluations. ***Journal of Economic Education*** 15 (Winter): 21-37.

Newton, J. D. (1988). Using student evaluation of teaching in administrative control: the validity problem. ***Journal of Accounting Education*** 6 (Spring): 1-14.

O'Connell, D. Q., and D. J. Dickinson (1993). Student ratings of instruction as a function of testing conditions and perceptions of amount learned. ***Journal of Research and Development in Education*** 27 (Fall): 18-23.

Rahman, M., M. Canlar, and D. Lambert (1988). Factors affecting accounting instructor evaluation: a test of two paradigms. ***Accounting Educators' Journal*** 1 (Fall): 134-146.

Rebele, J. E., D. E. Stout, and J. M. Hassell (1991). A review of empirical research in accounting education: 1985-1991. ***Journal of Accounting Education*** 9 (Fall): 167-231.

Rebele, J. E., B. A. Apostolou, F. A. Buckless, J. M. Hassell, L. R. Paquette, and D. E. Stout (1998a). Accounting education literature review (1991-1997), part I: curriculum and instructional approaches. ***Journal of Accounting Education*** 16 (Winter): 1-52.

Rebele, J. E., B. A. Apostolou, F. A. Buckless, J. M. Hassell, L. R. Paquette, and D. E. Stout (1998b). Accounting education literature review (1991-1997), part II: students, educational technology, assessment and faculty issues. ***Journal of Accounting Education*** 16 (Spring): 179-245.

Renaud, R. D., and H. G. Murray (1996). Aging, personality, and teaching effectiveness in academic psychologists. ***Research in Higher Education*** 37 (June): 323-340.

Rodin, M., and B. Rodin (1972). Student evaluations of teachers. ***Science*** 177 (September 29): 1164-1166.

Rustagi, N. K. (1997). A study of the retention of basic skills. ***Journal of Education for Business*** 73 (November-December): 72-76.

Sailor, P., B. R. Worthen, and E. Shin (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. ***Assessment and Evaluation in Higher Education*** 22 (September): 261-269.

Trout, P. A. (1997). What the numbers mean: providing a context for numerical student evaluations of courses. ***Change*** 29 (September-October): 24-30.

Williams, W. M., and S. J. Ceci (1997). 'How'm I doing?' problems with student ratings of instructors and courses. ***Change*** 29 (September-October): 12-23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. ***Chronicle of Higher Education*** 44 (January 16): A12-A14.

Wright, P., R. Whittenton, and G. E. Whittenberg (1984). Student ratings of teaching effectiveness: what the research reveals. ***Journal of Accounting Education*** 2 (Fall): 5-30.

Yunker, J. A., and J. W. Marlin (1984). Performance evaluation of college and university faculty: an economic perspective. ***Educational Administration Quarterly*** 20 (Winter): 9-37.

Yunker, P. J., and J. Sterner (1988). A survey study of faculty performance evaluation in accounting. ***Accounting Educators' Journal*** 1 (Fall): 63-71.

Table 1: Variable Information

Mnemonic	Type	Description	Mean	Max	Min	Std Dev
GRINTER1	dependent variable	Grade achieved by student in Intermediate Accounting I (A = 4; B = 3; C = 2; D, F, W = 1)	2.17	4.00	1.00	1.014
RATING	independent var. 1 (variable of interest)	Mean instructor rating by section of Introductory Accounting II which student had taken (5-point scale: 5 = best; 1 = worst)	4.01	4.93	2.54	0.588
GRINTRO 2	independent var. 2 (ability-achievement control 1)	Grade achieved by student in Introductory Accounting II (A = 4; B = 3; C = 2; D,F,W = 1)	2.96	4.00	1.00	0.809
GPA	independent var. 3 (ability-achievement control 2)	Student's Cumulative Grade Point Average on a 4 point scale (A = 4; B = 3; C = 2; D = 1)	3.08	4.00	1.76	0.505
ACT	independent var. 4 (ability-achievement control 3)	Student's ACT score prior to admission to the University	22.65	32.00	15.00	3.365

Table 2: Simple Correlation Coefficients

	GRINTER1	RATING	GRINTRO2	GPA	ACT
GRINTER1	1.0000				
RATING	0.0215	1.0000			
GRINTRO2	0.6014	0.1856	1.0000		
GPA	0.6030	0.0559	0.5727	1.0000	
ACT	0.4970	~0.0587	0.4222	0.4617	1.0000

GRINTER1 = Grade student achieved in Intermediate Accounting I
RATING = Class mean rating of instructor in student's Introductory Accounting II course
GRINTRO2 = Grade student achieved in Introductory Accounting II course
GPA = Student's cumulative Grade Point Average
ACT = Student's ACT score

Table 3: Multiple Regression Analysis

	Dependent Variable = GRINTER1					
Independent Variables	**Estimated regression coefficients of independent variables (t-values in parentheses; asterisk: significant at p = 0.05)**					**Partial Correl. (t-values)**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Constant	2.0242	0.5136	˜1.1390	˜1.8404	˜1.2845	—
RATING	0.0370 (0.10)	˜0.1618 ˜(1.95*)	˜0.1279 ˜(1.66)	˜0.2039 ˜(2.15*)	˜0.0946 ˜(0.73)	˜0.1592 ˜(2.15*)
GRINTRO2	—	0.7798 (12.91*)	0.5044 (7.41*)	0.5288 (6.50*)	0.4485 (1.76)	0.4383 (6.50*)
GPA	—	—	0.7562 (7.06*)	0.7297 (5.50*)	0.6606 (1.87)	0.3813 (5.50*)
ACT	—	—	—	0.0430 (2.35*)	0.0214 (0.62)	0.1195 (2.35*)
R-Squared	0.0004	0.3735	0.4686	0.5395	0.4748	—
Sample size	283 students	283 students	283 students	183 students	46 courses	183 students