The Canadian Journal of Higher Education
La revue canadienne d'enseignement supérieur
Volume XXVIII, No. 1, 1998 pages 1–20

Do Students Report that Easy Professors Are Excellent Teachers?

DELBERT A. BRODIE
St. Thomas University

Abstract

The present study was conducted to determine if students report that professors are excellent teachers when little studying is required to receive high grades. Altogether 1,939 student evaluations were obtained from 75 first-year university classes representing 15 disciplines. Mean expected grade in a class correlated negatively with how long students reported studying for that class. Across all classes the relationship between student evaluations and different expected grades was underestimated because most grades were concentrated around the mean of the distribution. When grades varied markedly across sections of the same course, the professor assigning highest grades with least studying received highest evaluation, including paradoxically teaching the most intellectually challenging course. Previous correlational studies have underestimated the biasing effect of grading leniency.

Résumé

Cette étude a été menée afin de déterminer si les étudiants estiment que les professeurs sont excellents quand ceux-ci n'exigent d'eux que peu de travail pour atteindre des notes élevées. Un ensemble de 1'949

Some of the results in the manuscript were presented at the Eastern Psychological Association Annual Meeting in Boston, Massachusetts on March 31, 1996.

évaluations d'étudiants ont été obtenues dans 75 classes de première année universitaire dans lesquelles 15 disciplines étaient représentées. La note moyenne attendue d'une classe est en corrélation négative avec la durée du temps d'étude rapporté par les étudiants. Dans l'ensemble des classes, l'effet des notes attendues est sous-estimé car la plupart des résultats se concentrent sur la moyenne de la distribution. Quand les notes varient significativement entre les sections d'un même cours, le professeur qui donne les notes les plus élevées avec une durée d'étude moindre reçoit l'évaluation la meilleure, ceci incluant paradoxalement l'enseignement de cours intellectuellement difficiles. Les études corrélationnelles antérieures ont ainsi sous-estimé les biais liés à l'indulgence du professeur lors de l'évaluation de sa prestation par les étudiants.

Teaching effectiveness refers to the degree to which a teacher facilitates students to achieve educational goals (McKeachie, 1986). In colleges and universities the primary focus of education is on achieving cognitive objectives such as the enhancement of students' knowledge, comprehension, application, analytical reasoning, synthetic reasoning, and evaluation (Bloom, 1956).

According to Aleamoni (1987, p. 111), ". . . a careful scrutiny of actual working systems of instructional evaluation reveals that student ratings of instructor and instruction is still the *only* [italics added] component that is regularly obtained and used." Furthermore, even when other data are available, student evaluations are assumed to be a better measure of teaching effectiveness because only students observe the professor throughout a course (Howard, Conway, & Maxwell, 1985).

Several hypotheses predict that student evaluations are biased measures of teaching effectiveness. For example, social psychologists have long maintained that people interpret events to protect and enhance their images of themselves (cf. Allport, 1937; Sherif & Cantril, 1947). An enormous amount of research has supported self-enhancement hypotheses. For instance, numerous investigators have found that when success and failure were manipulated, people attributed their own success more to internal factors than a stranger's success, and their own failure more to external factors than a stranger's failure (e.g., Sicoly & Ross, 1977; Snyder, Stephan, & Rosenfield, 1976). Besides maintaining and enhancing their self image, people seek consistency among related cognitions (Festinger, 1957). Therefore, if students are motivated to maintain

consistent positive self images, then students will have uniform self assessments of their academic ability, and discrepancies in performance across courses would be attributed to situational factors that do not undermine students' self images. For example, attributing discrepancies in performance to differences in teacher effectiveness would not undermine students' self images, whereas attributing discrepancies to differences in cognitive skills required (e.g., knowledge versus application) would reduce students' image of their cognitive ability.

Like self-enhancement hypotheses, saliency hypotheses predict that student evaluations are biased. More specifically, stimuli that are objectively unusual in their context should alter student evaluations. Typically, students are enrolled simultaneously in several courses, and courses may differ on many dimensions. For instance, in one class the grading standard may be unusual, whereas in another class level of humour may be unusual. Previous research has found that salient stimuli are seen as the most causally powerful, produce the most extreme judgments, and increase consistency of judgments (Taylor, Crocker, Fiske, Sprinzen, & Winkler, 1979; Taylor & Fiske, 1978). Whether salient stimuli create positive or negative judgments depends on which judgment maintains students' positive self images, whether salient stimuli arouse pleasant or unpleasant emotions, etc.

Both laboratory experiments and experiments with actual classes support predictions from self-enhancement and saliency hypotheses that students' ratings are biased measures of instructional effectiveness. For instance, researchers have found consistently that grades cause students to change their evaluations of professors. That is, in laboratory experiments Perkins, Guerin, and Schleh (1990) and Snyder and Clair (1976) found that students who were randomly assigned higher grades rated the professor higher than students who were assigned lower grades. Similarly, Powell (1977) found with actual classes that as the percentage of correct answers required to receive passing grades increased across different sections of a course taught by the same professor, the number of correct answers on a final exam increased whereas students' grades declined along with evaluations of both the professor and the course. Finally, in actual classes with the same professor Zelby (1974) demonstrated that teaching lower cognitive objectives increased students' ratings of the professor's effectiveness from the bottom 50% of faculty to the upper 25%.

Supporters of student evaluations have dismissed findings from experiments because experiments (a) may be unrepresentative of actual differences among professors and classes, (b) may fail to represent many characteristics that may alter the relationship among the variables examined, may represent experimenter effects, etc. (cf. Abrami, d'Apollonia & Cohen, 1990). Consequently, some investigators have claimed that the best way to investigate the validity of student evaluations as a measure of instructional effectiveness is with multi-section (e.g., Abrami et al., 1990) or multi-method designs (e.g., Howard et al., 1985; Marsh, 1987).

In multi-section designs evaluations are collected from different sections of a single course, whereas in multi-method designs evaluations are obtained from a wide range of courses. Like experiments, most multi-section (e.g., Feldman, 1989; Marsh & Dunkin, 1992) and multi-method studies (e.g., Howard & Maxwell, 1980; Marsh, 1983) have found positive correlations between grades and student evaluations.

In all multi-method studies there are no common exams. Therefore, in multi-method studies higher grades could reflect easier grading standards and less learning (cf. Powell, 1977). Since most professors believe that grading leniency increases student evaluations (Aleamoni, 1987; Felder, 1992; Marsh & Overall, 1979), in classes with no common exams some professors may attempt to obtain high evaluations by assigning high grades. In multi-section courses with common exams, inflating grades is usually harder, but not always impossible. Marsden, McIntosh, and Adolph (1993) reported that after receiving numerous complaints from students in sections of a multi-section course who did not receive inflated grades, the chairperson collected all final exams and gave them to an outside consultant. The consultant supported the students' complaints by reporting that one professor had marked wrong answers correct for his students and incorrectly added his students' scores so they received an additional 12 to 20 percent.

If the correlation between grades and student evaluations in multi-method and multi-section studies is interpreted as grades biasing student evaluations, then typically the correlations are dismissed as having little or no practical usefulness because the correlations are small (e.g., Marsh, 1984). However, if most professors use similar grading standards, then (a) correlations across all classes between grades and student evaluations could be *small,* and (b) for professors that use different grading standards, grades could have *large* effects on student evaluations.

Customarily, correlation is measured using the Pearson product-moment correlation coefficient. Conceptually, the Pearson coefficient is the mean of the cross products of z scores (Minium, King, & Bear, 1993). Scores close to the mean of a distribution have z scores close to zero. Therefore, even if changes in one variable produce large changes in another variable, the mean of the cross products of z scores decreases as the percentage of scores near the mean of the distribution increases.

Unlike multi-method studies, grading leniency is usually not a major concern in multi-section courses with common exams. A major concern in multi-section courses is the infrequent measurement of higher level thinking. Normally, the common exam is the final exam, and ". . . final examinations typically weigh knowledge much more heavily than application, problem solving, or other cognitive objectives" (McKeachie, 1986, p. 275). Since students prefer lower cognitive objectives (Zelby, 1974), the observed correlations in multi-section studies could be due to final exam performance and student evaluations declining as professors stress higher cognitive objectives.

Despite intense disagreement over whether student evaluations measure instructional effectiveness (Gaski, 1987), student evaluations are widely used by professors to change their teaching, by researchers to operationally define effective teaching, by students to select professors, and by administrators to evaluate faculty (Abrami et al., 1990). The implicit assumption made frequently by professors, researchers, students, and administrators is that since student evaluations are positively correlated with measures of student learning in some situations, then student evaluations will be positively correlated with student learning in most or all situations. However, student evaluations may be positively correlated with student learning only when departments stress uniform standards and professors know that they will be evaluated based on their students' performance on common exams.

The purpose of the present study was to test three hypotheses when (a) there were no common exams in any courses, (b) each professor determined course material (e.g., textbooks, selection of chapters), (c) professors used diverse grading standards, (d) there were mandatory standardized student evaluations, and (e) student evaluations affected professors' income (e.g., whether rehired, promoted, etc.). First, it was predicted that when there are no common exams, some professors will use procedures that increase student evaluations while decreasing student

learning (Zelby, 1977), such as a lenient grading standard (Powell, 1977) and/or teaching lower cognitive objectives (Zelby, 1974). The first prediction is based on the findings that (a) most professors believe that grading leniency increases student evaluations (Aleamoni, 1987; Felder, 1992; Marsh & Overall, 1979), (b) when students evaluate, some professors make changes to produce higher evaluations (Keutzer, 1993; Marsh & Roche, 1993), and (c) increasing student evaluations is easier than increasing learning (Abrami, Perry, & Leventhal, 1982; Abrami & Mizener, 1985). Second, it was predicted that previous correlational studies have underestimated the relationship between different grades and student evaluations because, unlike experiments, in correlational studies most grades are close to the mean. Finally, based on previous experimental studies and hypotheses of self-enhancement and saliency, it was predicted that in correlational studies students will evaluate easy professors more favorably than hard professors.

A professor was defined as easy if most students in a class expected to receive high grades with little studying. It was assumed that unlike easiness, varying teacher effectiveness would alter grades within narrow limits. That is, it is easy to structure grading standards so most students in a class either fail or receive excellent grades (e.g., base grades on difficult class tests versus easy homework assignments). On the other hand, teaching most students to exhibit excellent knowledge, comprehension, analytical reasoning, etc. over all course material is arduous. Furthermore, based on both laboratory and field investigations, Abrami et al. (1990, p. 221) concluded that "instructors may have genuinely small effects on what students learn." Finally, even if different professors produce various amounts of learning in the classroom, learning may still be equal across classes because ". . . students are likely to compensate for poor instruction by studying harder in order to achieve the grades to which they aspire . . ." (McKeachie, 1987, p. 344).

Method

Participants

Anonymous evaluations of teachers/courses were obtained for all first year courses evaluated at the end of winter term 1994. There were 1,939 evaluations of teachers/courses from 75 classes representing 15 disciplines at St. Thomas University. According to *Maclean's* ("Primarily

undergraduate," 1993), the academic ability of St. Thomas students was average based on comparisons of the mean high school grades of first year students at primarily undergraduate universities in Canada (i.e., ranked 12 out of 23).

There were 7 to 70 student evaluations per class with only three classes above 43. The median number of evaluations was 24.

Materials

The evaluation form contained 22 statements on a teacher/course and 7 biographical questions about the student. When rating a teacher/course, students selected a response on a 5-point scale or had the option of indicating that the statement does not apply. On one statement students were asked to rate from poor to excellent "Overall, how would you evaluate this course?". On the other 21 statements, students indicated the extent to which they disagreed or agreed with an assertion (e.g., "The professor communicates effectively"). A condensed version of the 22 statements is presented along with results in Table 1.

When evaluating teachers/courses, higher scores always denoted more positive evaluations. Since students answer questions about a course differently depending on which professor is teaching the course, all statements on teachers/courses were assumed to measure students' perception of teacher effectiveness.

Data were analyzed from five biographical questions. Three questions asked about a specific course (i.e., expected grade, number of hours studied per week, and number of classes missed). The other two questions asked for the student's year in university and cumulative average grade in all classes. Expected grade in a course was measured on a 5-point scale from 0 for failure to 4 for A or excellent. Hours studied per week for a course was recorded as 0 to 2, 3 to 5, 6 to 8, or more than 8. Number of classes missed in a course was represented by the following 6 categories: 0, 1–3, 4–6, 7–9, 10–12, more than 12. Year in university was represented by 5 years starting at year 1. A student's cumulative average grade was measured on an 8-point scale from below 1.00 to more than 4.00 (i.e., A+ = 4.30) with the categories between the two extremes increasing by half a grade point (e.g., 1.00–1.50, 1.51–2.00).

Table 1
Principal Components Factor Analysis of Student Evaluations

Statement	Factor Loadings
• course requirements were clearly communicated	.68
• course activities concur with objectives	.77
• professor did not arbitrarily cancel classes	.44
• class time was generally useful	.76
• professor communicates effectively	.80
• topics covered formed a coherent course	.79
• parts of course were effectively coordinated	.79
• projects and assignments aided understanding	.71
• reading materials in the course were valuable	.67
• professor showed genuine concern for my progress	.76
• professor was available for consultation	.65
• methods of evaluations were fair	.74
• sufficient feedback was provided	.77
• I received helpful comments on my work	.77
• course helped me to grasp difficult concepts	.74
• course helped me to think for myself	.69
• course was intellectually challenging	.68
• I was encouraged to express my own views	.65
• experiences and questions were effectively used	.76
• professor displayed interest and enthusiasm	.71
• I would recommend course to others	.83
• overall evaluation of course	.86

Procedure

According to university regulations, all classes had to be evaluated at the start of a class period during the last two weeks of classes when the professor was not present. Printed on the evaluation form and as part of the standardized instructions, students were told that ". . . evaluations are carried out for the purposes of improving instruction and of providing evidence of effective teaching." When students finished rating the professor/course, evaluation forms were submitted to the registrar's office.

After course grades were submitted to the registrar, the university gave professors their mean scores along with both the department and university means for each of the 22 statements that evaluated teacher effectiveness. No means were provided for any of the biographical questions.

Results

When a response represented a range of values (e.g., 3–5 hours), data were analyzed using the midpoint of the range of values. On two scales (i.e., study hours & classes missed), the highest response represented an unspecified upper limit (e.g., more than 8 hours). These extreme responses were selected on less than 3% of the evaluations. To arrive at a value, it was assumed that the unbounded upper interval, like lower intervals, contained the next three adjacent integers.

Multi-method biographical data

Analyses of variance indicated that classes differed significantly ($p < .05$) on expected grade, $F(74,1737) = 3.74$, average grade, $F(74,1120) = 1.36$, study hours, $F(74,1745) = 5.05$, number of classes missed, $F(74,1750) = 4.24$, and year in university, $F(74,1755) = 5.19$. The mean, standard deviation, and range for the five biographical variables are presented in Table 2. Inspection of the range reveals large differences among classes on each biographical question.

Pearson product-moment correlations among the five biographical variables across 75 class means revealed a significant correlation between study hours and expected grade, $r = -.38$, $p < .001$. The correlation between expected grades and average grades, $r = .11$, $p = .34$, and all other correlations were nonsignificant ($p > .05$).

As shown in Figure 1, expected grades tended to decrease as study hours for a given course increased, and most scores were concentrated

Table 2
Means, Standard Deviation of Class Means and Range of Class Means for Five Biographical Variables

Biographical Variable	M	SD	Range
Grade point average	3.04	.19	2.63 to 3.60
Grade point expected	3.04	.32	2.00 to 3.92
Study hours	3.22	.88	1.20 to 6.00
Classes missed	4.36	1.32	2.04 to 8.47
Year in university	1.39	.34	1.00 to 3.00

near the university's mean. To test the null hypothesis that the distributions of scores were evenly distributed, ten equal width intervals were created for each biographical variable. Chi square tests indicated that class means were not equally distributed across the 10 intervals for any biographical variable, X^2 (9) $\geq$ 22.32, $p < .01$. For each variable the three adjacent intervals closest to the mean had the largest frequencies and represented at least 50% of the scores, whereas the three intervals at the extremes with the lowest frequencies represented 8% or less of the scores. The hypothesis of a normal distribution of class means was retained for expected grade, average grade, and classes missed, Lilliefors(75) $\leq$ .086, $p > .20$, but rejected for study hours and year in university, Lilliefors(75) $\geq$ .109, $p < .05$.

To learn the extent that the concentration of scores near the mean reduced the correlation, correlations between expected grades and study hours were recalculated by dividing the variable on the abscissa into 75 equal intervals with the same range as the obtained data. Ten sets of values on the ordinate were computed using the regression equation obtained from the data in Figure 1 plus randomly selected values of residual (i.e., difference between predicted and observed values). For each set of hypothetical data, values of the residual were selected without replacement. Hence, hypothetical data altered the distribution of scores while having little or no effect on the accuracy of prediction. The median correlation with an equal distribution of values on the abscissa was substantially higher than the correlation obtained with scores concentrated near the mean (i.e., –.57 vs. –.38).

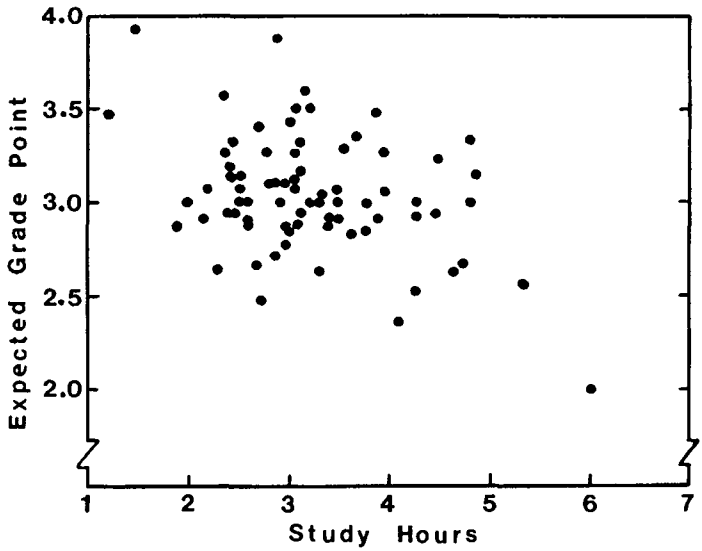

Figure 1

Circles represent class means as a function of students' reported study hours per week for a given class and their expected grade point in that class. Even if students did not study, the minimum length of time that students could report studying was one hour per week per class.

Multi-method evaluation

Analyses of variance indicated that the classes differed significantly on each of the 22 statements evaluating teacher effectiveness, median $F(74,1831) = 6.46$, $p < .001$, with median ratings across questions varying from 2.00 to 4.44.

A principal component analysis of the 22 statements identified only one factor with an eigenvalue greater than chance values using the parallel analysis criterion developed by Longman, Cota, Holden, and Fekken (1989). Inspection of the correlations in Table 1 between the principal component factor and each statement on the student evaluation form indicates all items correlated positively with the factor (i.e., .44 to .86, with only one correlation below .65 for 1,410 students who rated each

statement). The correlation between the factor and the simple unweighted average of all 22 statements was .99. The factor accounted for 54% of the variance on student evaluations. The reliability coefficient for 22 statements using Cronbach's alpha was .96.

Relationship between multi-method biographical data and evaluations

Using means from 75 classes, Pearson product-moment correlations between the principal component factor and each of the five biographical variables revealed a significant ($p < .05$) correlation only with expected grades ($r = .23$). As shown in Figure 2, like biographical data, factor scores were concentrated near the university's mean. The hypothesis of an equal distribution of factor scores across ten equal width intervals was rejected, X^2 (9) = 39.00, $p < .01$, whereas the hypothesis of a normal distribution was retained, Lilliefors(75) = .089, $p > .20$.

Inspection of Figure 2 also indicates that when class grades were within one standard deviation of the university's mean, factor scores varied widely and were not associated with changes in expected grades. In contrast, when class grades were more than one standard deviation away from the university's mean, factor scores were almost always positive if grades were above the mean, and negative if grades were below the mean.

Using the same procedure described earlier for biographical data, the correlation was increased substantially when grades were divided into 75 equal intervals and hypothetical factor scores were calculated using the original regression equation plus randomly selected values of residual. Specifically, the correlation increased from .23 to a median correlation of .36. Finally, since in experiments different levels of an independent variable represent markedly different values, a correlation also was calculated using the obtained data when grades were deleted within one standard deviation of the university's mean. As expected, this procedure increased the correlation substantially (i.e., from .23 to .48).

Multi-section data

Most departments had four or fewer sections of the same course, no department had five or six sections, and four departments had seven or more sections. Since differences in the course material could potentially account for differences in each of the previous analyses, final analyses were restricted to departments where there were seven or more sections

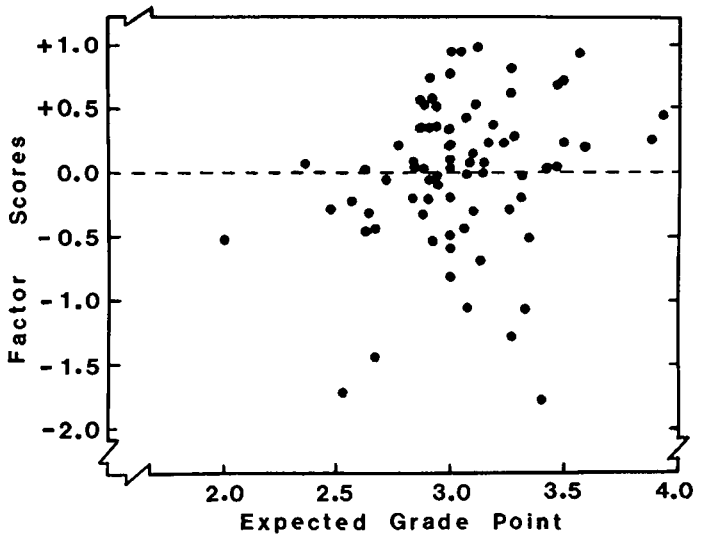

Figure 2

Circles represent class means as a function of students' expected grade point in a given class and principal component factor score for that class.

of the same course taught by different professors, and sections did not differ significantly in students' year in university. Only two departments satisfied these criteria. Since the course evaluations that were available to the author were anonymous, it was impossible to determine the precise identity of these departments. Nevertheless, based on the number of sections, the departments were either philosophy, psychology, or sociology.

In department X sections differed significantly ($p < .05$) on three biographical variables: expected grade, $F(6,130) = 4.58$, study hours, $F(6,128) = 2.20$, and classes missed, $F(6,131) = 2.94$. Sections did not differ significantly on students' cumulative average grades, $F(6,74) = 1.58$, $p = .17$, and expected grades in a class did not increase as cumulative average grades increased. That is, the two sections with the highest reported cumulative average grades were the sections with the lowest expected grades in department X.

Section means for the three biographical variables with significant differences are presented in the first three columns of Table 3. Inspection of these means indicates that students in section G had the highest expected grade, almost the highest number of classes missed, and the shortest study time. Compared to all 75 classes (cf. Table 2 & Figure 1), students in section G missed an average number of classes, but they were above the ninety-fifth percentile in expected grade, and below the eleventh percentile in study time. Like students in section G, students in section F expected high grades. However, rather than reporting short study times, students in section F reported the highest study time in the department. Nevertheless, length of study time in section F was less than one standard deviation above the university mean.

In department X sections differed significantly on each of the 22 statements on teacher effectiveness, median $\underline{F}(1,139) = 5.18$, $\underline{p} < .001$ with difference between the lowest and highest section means ranging from .68 to 2.61. Compared to other sections of the same course, section G received the highest evaluation on 17 statements (e.g., "overall how would you evaluate this course"; "I found this course intellectually challenging"), whereas section F received the highest evaluation on only 3

Table 3
Biographical and Factor Means for Each Section in Department X

Section	Expected grade	Study hours	Classes Missed	Factor
A	2.67_{a}	2.67_{ab}	3.00_{ab}	$-.43_{a}$
B	2.83_{a}	3.63_{ab}	2.04_{a}	$-.20_{ac}$
C	2.91_{ab}	3.42_{ab}	3.58_{ab}	$-.53_{a}$
D	2.94_{ab}	2.41_{ab}	3.82_{ab}	$.52_{bc}$
E	3.08_{ab}	3.05_{ab}	4.25_{b}	$.08_{ad}$
F	3.48_{b}	3.86_{b}	3.78_{ab}	$.69_{bd}$
G	3.57_{b}	2.34_{a}	4.24_{b}	$.94_{b}$

Note: Means in the same column that do not share subscripts with the same letter differ at $\underline{p} < .05$ in Tukey honestly significant difference comparison.

statements, sign test $p < .01$. Whenever section G did not receive the highest rating, the difference in ratings between the section with the highest rating and section G was always less than two tenths of a point on a 5-point scale. Tukey honestly significant difference comparison indicated that on 21 statements section G received significantly higher ratings than other sections, whereas Section F received significantly higher ratings than other sections on only 12 statements. Differences in evaluations of teacher effectiveness between sections F and G were significant on only two statements. Specifically, compared to students in section F, students in section G agreed significantly more strongly with statements that course requirements were clearly communicated and it was clear how the topics covered formed a coherent course of study.

Finally, in department X sections differed significantly ($p < .01$) on the principal component factor, $F(6,103) = 9.27$. As shown in the last column in Table 3, sections F and G received the highest positive loadings on this factor of any section, and the weights were significantly higher than weights for sections A, B, and C. Compared to all 75 classes (cf. Figure 2), the loading for section G was above the ninety-fifth percentile.

Data from department Y differed markedly from department X. In department Y none of the expected class grades was above the university's mean, the section with the highest expected grade did not study the least, and none of the sections differed on the principal component factor, $F(6,170) = 1.46$, $p = .19$.

Discussion

Hypotheses of saliency, but not self-enhancement, predict large variability in student evaluations of courses with the same expected grades, especially if the students expect the same grades across all of their classes. In this context grades are not objectively unusual, and therefore not salient. There are potentially many stimuli other than grades that may be salient. For example, course format (e.g., self-paced vs. lecture) or course material (e.g., mathematics vs. humanities) could be salient. Since the students and professors were anonymous in the present study, testing for other salient stimuli was impossible.

Differences in expected grades across classes in the present study do not appear to be caused by differences in the ability of students. This conclusion is based partly on the absence of any significant correlation

between students' average grades in all classes and students' expected grades in specific classes using multi-method data. The conclusion is also based on the absence of any significant differences in students' cumulative average grades in department X.

College students typically need to spend about two hours studying outside class for every hour spent in class (Michael, 1991). Since in the present study students were in each class three hours per week, the expected study time per week should be six hours. Inspection of Figure 1, however, indicates that in most courses students reported studying less than six hours a week. Low levels of studying and the negative correlation between class grades and study hours are consistent with the finding that most professors believe that grading leniency increases student evaluations (Aleamoni, 1987; Felder, 1992; Marsh & Overall, 1979), and the hypothesis that some professors may be more concerned about student evaluations than student learning.

Like previous multi-method studies (e.g., Gigliotti & Buchtel, 1990; Marsh, 1983), in the present study the correlation between expected grades and student evaluations was small when all class means were included in the analysis. Investigators have interpreted small correlations between grades and student evaluations as indicating that if student evaluations are biased due to professors using different grading standards, then the bias is small. However, inspection of Figure 1 or 2 indicates that most of the expected class grades were close to the university's mean grade. Therefore, small correlations could simply indicate that most professors used similar grading standards. The artifact produced in multi-method and multi-section designs by having most scores close to the mean may have induced Aleamoni (1987) and Felder (1992) to incorrectly label the grading leniency bias as a myth. When most class grades are close to the university's mean, then a small correlation across all classes does not indicate that markedly different grades have little effect on student evaluations. As shown in the present multi-method study, the correlation was increased substantially if grades were either evenly distributed or if the middle grades were deleted. Furthermore, in department X the section with the highest grades and least studying received substantially higher student evaluations than other sections.

Unlike correlational studies, in experiments there are usually an equal number of observations at each level of the independent variable, and different levels of the independent variable represent discontinuous

distributions. Therefore, changes in the distribution of scores from experiments to correlational studies can account for different estimates of the biasing effect of grades on student evaluations.

The results from the present study are consistent with the hypothesis that the biasing effect of grades that has been observed in experiments (e.g., Perkins et al., 1990; Powell, 1977; Snyder & Clair, 1976) generalize to correlational studies. The fact that in some classes most students expected high grades with little studying indicates that some professors taught easy courses. Even though grading leniency decreases learning (Powell, 1977), easy courses received high student evaluations. Hence, if student evaluations measure students' beliefs about excellence of teaching, then students believe that professors who teach easy courses are excellent teachers.

If student evaluations are used to measure teaching effectiveness, recognizing that sometimes students produce paradoxical evaluations is important. For example, in the present study it seems illogical that students in the section with extremely high grades and short study times rated that section to be the most intellectually challenging. A similar non sequitur was observed by Powell (1977) when he compared lenient and stringent grading standards. His students rated their own effort markedly lower if the professor used a lenient grading standard. Nevertheless, Powell (1977) found that students agreed more strongly that the professor stimulated effort and thinking when the professor used a lenient grading standard.

In conclusion, by themselves high student evaluations do *not* indicate that a professor is an effective teacher. Sometimes the professor with the highest student evaluation may generate the least studying and produce the least learning. Student ratings should not be used to evaluate teaching without information on factors which may substantially bias the evaluations (e.g., expected grades, study time) and an independent measure of student learning. Obtaining additional information would both reduce the misuse of student evaluations and provide alternative measures of teaching effectiveness. For instance, in some situations how long students report studying for a course may be a better measure of teaching effectiveness than student evaluations. Rather than trying to measure learning indirectly by using either study time or student evaluations, however, the best measures of teaching effectiveness are standardized tests of students' knowledge, comprehension, application, and other cognitive abilities.

References

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

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

Abrami, P.C., Perry, R.P., Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74,* 111–125.

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

Allport, G.W. (1937). *Personality: A psychological interpretation.* New York: Holt.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain.* New York, NY: D. McKay, Inc.

Felder, R.M. (1992). What do they know, anyway? *Chemical Engineering Education, 26,* 134–135.

Feldman, K.A. (1989). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multi-section validity studies. *Research in Higher Education, 30,* 583–645.

Festinger, L. (1957). *A theory of cognitive dissonance.* Evanston, IL: Row, Peterson.

Gaski, J.F. (1987). On "construct validity of measures of college teaching effectiveness." *Journal of Educational Psychology, 79,* 326–330.

Gigliotti, R.J., & Buchtel, F.S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82,* 341–351.

Howard, G.S., Conway, C.G., & Maxwell, S.E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology, 77,* 187–196.

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

Keutzer, C.S. (1993). Midterm evaluation of teaching provides helpful feedback to instructors. *Teaching of Psychology, 20,* 238–240.

Longman, R.S., Cota, A.A., Holden, R.R., & Fekken, G.C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24,* 59–69.

Marsden, W., McIntosh, A., & Adolph, C. (1993, August 14). The Fabrikant file. *The Gazette,* Montreal, pp. B1–B10.

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

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

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

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

Marsh, H.W., & Overall, J.U. (1979). *Validity of students' evaluations of teaching: A comparison with instructor self evaluations by teaching assistants, undergraduate faculty and graduate faculty.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service, No. ED 177 205)

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

McKeachie, W.J. (1986). *Teaching tips: A guidebook for the beginning college teacher* (8th ed.). Toronto, ON: D.C. Health.

McKeachie, W. J. (1987). Instructional evaluation: Current issues and possible improvements. *Journal of Higher Education, 58,* 344–350.

Michael, J. (1991). A behavioral perspective on college teaching. *Behavior Analyst, 14,* 229–239.

Minium, E.W., King, B.M., & Bear, G. (1993). *Statistical reasoning in psychology and education* (3rd ed.). Toronto, ON: Wiley.

Perkins, D., Guerin, D., & Schleh, J. (1990). Effects of grading standards information, assigned grade, and grade discrepancies on students' evaluations. *Psychological Reports, 66,* 635–642.

Powell, R.W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education, 7,* 193–205.

Primarily undergraduate universities. (1993, November 15). *Maclean's, 106*(46), 34–35.

Sherif, M., & Cantril, H. (1947). *The psychology of ego-involvement, social attitudes and identifications.* New York, NY: Wiley.

Sicoly, F., & Ross, M. (1977). Facilitation of ego-biased attributions by means of self-serving observer feedback. *Journal of Personality and Social Psychology, 35,* 734–741.

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

Snyder, M.L., Stephan, W.G., & Rosenfield, D. (1976). Egotism and attribution. *Journal of Personality and Social Psychology, 33,* 435–441.

Taylor, S.E., Crocker, J., Fiske, S.T., Sprinzen, M., & Winkler, J.D. (1979). The generalizability of salience effects. *Journal of Personality and Social Psychology, 37,* 357–368.

Taylor, S.E., & Fiske, S.T. (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). New York, NY: Academic Press.

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

Zelby, L.W. (1977). Good teaching: A problem in education. *Social Science, 52,* 133–138.