Grading Leniency Is a Removable Contaminant of Student Ratings

Anthony G. Greenwald and Gerald M. Gillmore
University of Washington

It is well established that students' evaluative ratings of instruction correlate positively with expected course grades. The authors identify 4 additional data patterns that, collectively, discriminate among 5 theories of the grades–ratings correlation. The presence of all 4 of these markers in student ratings data (obtained at University of Washington) was most consistent with the theory that the grades–ratings correlation is due to an unwanted influence of instructors' grading leniency on ratings. This conclusion justifies use of a statistical correction—illustrated here with actual ratings data—to remove the unwanted inflation of ratings produced by lenient grading. Additional research can profitably seek other inappropriate influences on ratings to identify more opportunities for validity-enhancing adjustments.

Imagine that you have just taught a course for the first time and have received low ratings from students. You are about to teach the course again and are convinced that you need to change something. Consider two options. One option is to blame yourself, deciding that you did not explain the material clearly enough; you can correct that by spending more time on basic material, trying to ensure that students will master at least that basic material. The other option is, in effect, to blame the students, deciding that they didn't work hard enough; you can oblige them to work harder by giving weekly paper assignments or quizzes.

Both strategies are likely to raise students' grades and may improve ratings (see Powell, 1977). The retreat-to-basics approach will increase grades at least partly by reducing coverage of course material, so that less work is needed to achieve whatever percentage level of mastery is required for a given grade. The more-frequent-evaluation approach will not make it easier to earn a given grade, but it should get students to achieve more by prodding them to spend more time on the course.

Even though both approaches may prove to be successful, the retreat-to-basics alternative may be favored because of two likely influences. The first influence is from students' written elaborations of their low ratings; these often include complaints that tests covered material that was never clearly explained.[1] The second likely influence is from colleagues who provide advice based on experience with student ratings. As an example of such advice, consider the following:

> Students who think they are getting As tend to think more highly of their professor than students who believe they are getting Cs. So for a professor to maximize evaluations, the best bet is to give out a softball midterm, so that everyone thinks they're getting a great grade. However, if a professor really wants students to learn, the ideal method is to give a hard midterm, and scare the students into studying. Thus, the goals of pedagogy and high instructor evaluation are in direct opposition. If you give out lots of Cs and students think you are a great professor, you're probably excellent. If you give out all A and A minuses, and students think you're just OK, you probably suck.[2]

Are Ratings Influenced by Grades?

Any consideration of strategies to increase ratings is likely to focus quickly on the very simplest strategy that is suggested by academic folklore—just give higher grades. The strategy of giving high grades is so very tempting if only because it is so very simple. One need make no change beyond recalibrating the course's grade scale. To judge from anecdotes available on the academic grape-

Anthony G. Greenwald and Gerald M. Gillmore, Department of Psychology, University of Washington.

Part of the material in this article was presented as an address for the Donald T. Campbell Award from the Society of Personality and Social Psychology at the 103rd Annual Convention of the American Psychological Association, August 1995, New York.

Support for the research reported in this article was provided by Grant SBR-9422242 from the National Science Foundation and Grant MH 41328 from the National Institute of Mental Health.

For comments on various drafts of this article or material preliminary to it, the authors thank Robert D. Abbott, Philip C. Abrami, Kenneth A. Feldman, Joe Horn, George S. Howard, Herbert W. Marsh, Scott E. Maxwell, Jeremy D. Mayer, Robert S. Owen, Lloyd K. Stires, and John E. Stone.

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525. Electronic mail may be sent via Internet to agg@u.washington.edu.

[1] Such complaints are possibly valid but also can represent a very normal tendency to project blame externally following poor performance.

[2] Abridged and quoted with permission from an Internet bulletin board message circulated by Jeremy D. Mayer, Department of Government, Georgetown University, July 11, 1995.

Vol. 52, No. 11, 1209–1217

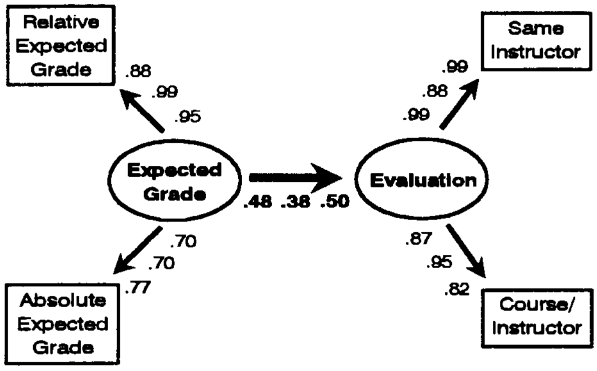
vine, the faith that the grade-increasing strategy works appears to have some basis in real experiences of teachers. Nevertheless, it would be very desirable to have a methodologically sound research answer to the question, If I give higher grades, will I get higher ratings?

The Grades–Ratings Correlation

This investigation starts from the widely observed phenomenon that course grades are positively correlated with course evaluative ratings (Stumpf & Freedman, 1979). Figure 1 shows the grades–ratings relationship in the form of a structural model that relates two measures of expected grades to two measures of course and instructor evaluations.

The data presented in Figure 1 were obtained in a series of three studies at University of Washington during the 1993–1994 academic year. These studies used a new rating form (Form X; see Gillmore & Greenwald, 1994) that added several measures to forms previously in use at University of Washington. Data were obtained from 200 or more courses in each of several academic terms. Although these were university-wide samples of courses that were diverse in subject matter, class size, and academic level, the courses were also self-selected by virtue of instructors having volunteered to use the new rating form. Results from undergraduate courses for which at least 10 students provided ratings responses are summarized in Figure 1. The positive grades–ratings correlation is measured by the standardized path coefficient (averaging .45 for the three samples) linking the two latent variables of Expected Grade and Evaluation.

Figure 1
Structural Model Including Two Measures of Expected Grade and Two Measures of Evaluative Ratings of Course and Instructor



Data Set	*N*	χ^2	*df*	*p*	rmsea	*P*(close fit)
Autumn '93	205	1.51	2	.47	.00	.62
Winter '94	205	1.37	2	.50	.00	.65
Spring '94	184	0.47	1	.49	.00	.58

Note. The three coefficients on each path are standardized values (i.e., on the same −1 to 1 scale as correlation coefficients) shown in left-to-right order for the three data sets. Statistics report major tests of fit for this structural model. Nonsignificant ($p > .05$) chi-square values indicate satisfactory fit. Chi-square values have an extra degree of freedom when the computational routine added a constraint to avoid a negative variance estimate. Rmsea is the root-mean-square error of approximation index of fit that has been described by Browne and Cudeck (1993) and by MacCallum, Browne, and Sugawara (1996). These authors characterized an rmsea less than .05 as indicating "close" fit, .05–.08 as "close to fair" fit, .08–.10 as "mediocre" fit, and an rmsea greater than .10 as "poor" fit. *P*(close fit) values greater than .05 indicate satisfactory fit.

Five Theories of the Grades–Ratings Correlation

The positive relationship between grades and ratings shown in Figure 1 is typical of many previous studies (see Stumpf & Freedman, 1979, for an overview). Of course, the existence of this grades–ratings correlation prompts a suspicion that ratings can be increased by the strategy of increasing grades, but by no means does it demand that conclusion. Each of the first three of the following five theories explains the grades–ratings correlation by assuming that a third variable influences both grades and ratings. By appealing to third variables, these theories avoid the assumption of a causal influence of grades on ratings. The remaining two theories do assume that grades have a causal influence on ratings.

1. *Teaching effectiveness influences both grades and ratings.* This is the one theory that is fully based on the presumed construct validity of student ratings (see McKeachie, 1979, pp. 390–391). The central principle of the teaching-effectiveness theory is that strong instructors teach courses in which students both (a) learn much (therefore, they earn and deserve high grades) and (b) give appropriately high ratings to the course and to the instructor. Thus, instructional quality is a third variable that explains the grades–ratings correlation in a way that raises no concern about grades having improper influences on ratings.

2. *Students' general academic motivation influences both grades and ratings.* Compared with unmotivated students, students with strong academic motivation should do better in their course work and should more fully appreciate the efforts of the instructor, possibly even inspiring the instructor to superior performance. Courses that attract highly motivated students should give higher grades (because the students work harder) and should get higher ratings (because the motivated students appreciate both course and instructor). Student motivation has been suggested as the operative third variable in several research investigations of student ratings (e.g., Howard & Maxwell, 1980; Marsh, 1984).

3. *Students' course-specific motivation influences both grades and ratings.* This theory differs from the preceding one by supposing that a student's motivation can vary from course to course rather than being a fixed characteristic of the student. Because the two motivation theories credit the relationship between grades and ratings to a characteristic of students, they may appear not to support a teaching-effectiveness interpretation of ratings. However, if student motivation is itself credited to the

instructor—for example, the instructor either attracts highly motivated students or motivates them once they are in the course—these theories retain the interpretation that ratings measure teaching effectiveness.

4. *Students infer course quality and own ability from received grades.* Social psychological *attribution* theories describe how people make inferences both about their own traits and about the properties of situations in which they act by observing the outcomes of their actions. Research in the attribution-theory tradition shows that favorable outcomes for one's own behavior typically lead to inferences that one has desirable traits, whereas unfavorable outcomes may lead one to perceive situational obstacles to success. A simple summary of these attributional principles is that people tend to accept credit for desired outcomes while denying responsibility for undesired outcomes (Greenwald, 1980). Applying this principle to the academic context, one would expect that high grades will be self-attributed to intelligence or diligence and that low grades will be attributed to poor instruction. Social psychological attribution theory matured after the peak of research activity on student ratings, perhaps explaining why this interpretation has been infrequently mentioned in research on student ratings. Some recent discussions of attribution interpretations appear in articles by Gigliotti and Buchtel (1990) and Theall, Franklin, and Ludlow (1990); see also the overview by Feldman (1997).

5. *Students give high ratings in appreciation for lenient grading.* The idea that praise induces liking for the praiser (especially if the praise is greater than expected) is familiar in social psychology (Aronson & Linder, 1965). The translation of this familiar principle into the ratings context is that the instructor, in effect, praises the student by means of a high grade, and the student's return liking is expressed by providing high ratings. This *leniency* or *grade-satisfaction* theory has been a focus of much controversy in past research on validity of student ratings. The leniency interpretation was advocated by researchers who were critical of ratings validity in the 1970s, including those who published demonstrations in natural classroom settings that grade manipulations affected student ratings (Chacko, 1983; Holmes, 1972; Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979). However, support for the leniency theory declined sharply in the wake of correlational construct-validity research conducted in the late 1970s and early 1980s. Mentions of leniency or grade-satisfaction theories in post-1980 publications appear mostly in the context of asserting that leniency may account for only minor and ignorable influences on student ratings (see quotations of such conclusions in Greenwald, 1997, this issue).

Four Theory-Diagnostic Patterns in Correlational Student Ratings Data

Table 1 presents four data patterns that can collectively discriminate among the five theoretical interpretations of the grades–ratings correlation. For completeness, the

Table 1
Success of Five Theories in Explaining Five Patterns in Student Ratings Data

	Established pattern	Diagnostic pattern			
Type of explanation and hypothesis	Positive between-classes grades–ratings correlation	Positive within-classes grades–ratings correlation	Greater correlation for relative grade than absolute grade	Grade correlation radiates to peripheral items (halo)[a]	Negative between-classes grades–workload correlation
Third variable affects both grades and ratings					
Third variable is instructor's teaching effectiveness	✓	x	x	x	x
Third variable is student's general academic motivation	✓	✓	x	x	x
Third variable is student's course-specific motivation	✓	✓	✓	x	x
Grades influence ratings					
Attribution: Grades provide information about course quality and student ability	✓	✓	✓	✓	x
Leniency: Students reward/punish instructors who give high/low grades	✓	✓	✓	✓	✓

Note. ✓ = hypothesis predicts result; x = hypothesis predicts either a null or opposite-direction result.
[a] This halo effect is a positive grades–ratings correlation (across students, within courses) for items that, rationally, should be evaluated in the same way by all students in the same class (i.e., independently of their grades).

grades–ratings correlation also appears (as the first listed pattern) in Table 1.

With the exception of one finding that was tested only during a single academic term (the grade-related halo effect listed below), the following four findings have been corroborated in separate data collections over three or more academic terms in university-wide samples of courses at University of Washington. As each finding is described, its use to evaluate the five theories is explained.

1. Positive grades–ratings relationships within classes. In addition to between-classes grades–ratings correlations as described in Figure 1, grades–ratings correlations are also routinely obtained within classes (Stumpf & Freedman, 1979). In the University of Washington data, the within-classes relationship has been observed very reliably. Because, in the teaching-effectiveness theory, the variable that influences both grades and ratings is a constant (the instructor) within any classroom, that theory does not explain within-classes covariations of grades and ratings. By contrast, the two third-variable theories that allow student differences within a classroom to be related to ratings are able to explain the within-classes grades–ratings correlation. Also, of course, the attribution and leniency theories very directly explain why students who get higher grades in any class should evaluate that course more positively than others.

2. Stronger grades–ratings relationships with relative (rather than absolute) measures of expected grade. The structural model shown in Figure 1 includes two measures of expected grades: absolute and relative expected grades. The absolute measure used class medians on the 0.0 (*E* or fail) to 4.0 *(A)* grading system in use at University of Washington. The relative measure used class medians on a measure that asked each student to report the relationship of the grade expected in the rated course to the student's average grade in other courses. The stronger weight of the relative measure on the Expected-Grade latent variable (see Figure 1) reflects the finding that the grades–ratings relationship was stronger for the relative-grade measure than for the absolute-grade measure. In regression analyses that predicted ratings simultaneously from both of the expected-grade measures, the relative-grade measure yielded a substantial gain in the percentage of ratings variance explained, over and above that explained by the absolute expected-grade measure. By contrast, the absolute-grade measure accounted for very little variance beyond what was explained by the relative-grade measure. The superiority of the relative-grade measure was evident in both between-courses and within-courses analyses. The comparison of relative- and absolute-grade measures was a novel feature of the University of Washington research. Consequently, this finding—that the grades–ratings correlation is stronger for the relative-grade measure—is previously unreported in the research literature on student ratings.

The teaching-effectiveness interpretation does not explain any within-classes grades–ratings correlation, let alone the greater strength of this correlation for the relative-grade measure than the absolute-grade measure. The general academic motivation theory, which ties ratings to the student's assumed stable level of motivation, also has trouble explaining the superiority of the relative-grade measure, unless it is (implausibly) assumed that highly motivated students always report that expected grades are above their average. By contrast, the course-specific motivation theory and the attribution and leniency theories readily explain why ratings associated with a specific course are higher when the grade in that course is relatively high for the student.

3. Grade-related halo effect in judging course characteristics. In the winter quarter of 1994, approximately 100 instructors at University of Washington agreed to add a small set of items to their regular rating forms. The added items included three judgments that, a priori, were unlikely to be more than weakly related to quality of instruction. These three items sought students' judgments of (a) legibility of the instructor's handwriting, (b) audibility of the instructor's voice, and (c) quality of classroom facilities to aid instruction (such as an overhead projector). Figure 2 shows the magnitudes of grades–ratings correlations for these three items both between and within courses. There was no evidence of a grades–ratings relationship in the between-courses analyses, consistent with the assumption that these items are peripheral to instructional quality. However, the within-courses analyses showed clear positive relationships. Although these within-courses relationships were smaller than within-courses relationships observed for global course-rating items, they were extremely stable statistically. Because all students in the same classroom saw the same instructor's handwriting, heard the same instructor's voice, and had the same classroom teaching aids, the observation of these within-sections relationships is remarkable. The content of items on which these grade–halo effects occurred—especially their noncentrality to most conceptions of instructional quality—suggests the potency of grade influences on students' ratings.[3]

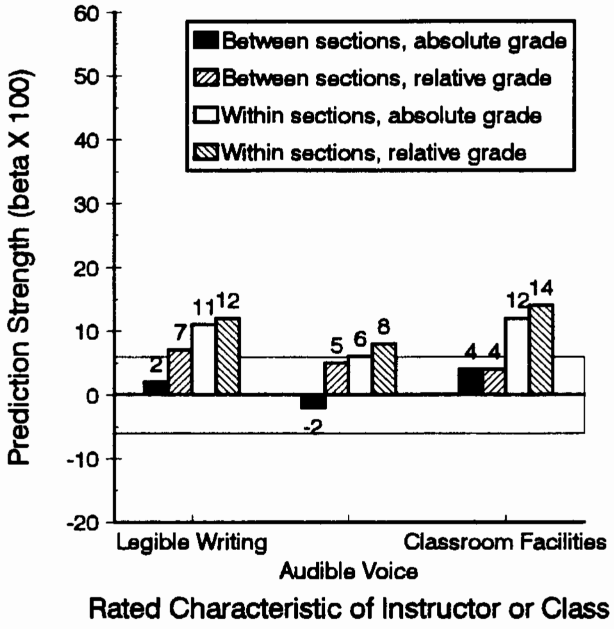
All three of the third-variable theories should expect data patterns at odds with the halo effects shown in Figure 2. For the teaching-effectiveness theory, if there are any grade effects on the legibility, audibility, and class facilities items, those effects should appear in between-classes analyses (but they do not), and they should not appear in within-classes analyses (but they do). The two student-motivation third-variable theories are strained in attempts to account for the pattern of grade-related effects on these

[3] Previous findings that front-of-class seating is associated with higher grades (e.g., Knowles, 1982) provide the basis for a possible student-motivation interpretation of the within-courses relationships of expected grades to ratings of the audibility of the instructor's voice and the legibility of the instructor's handwriting, although not the relationship to ratings of classroom facilities. The authors thank Lloyd K. Stires (personal communication, October 26, 1995) for noting the relevance of the classroom-seating variable to these data.

three items. To spell this out, one might suppose that highly motivated students are more likely to easily read the instructor's handwriting, to clearly hear the instructor, and perhaps even to notice the classroom facilities. Given either student-motivation interpretation, however, these effects should have appeared in between-courses analyses as well as in within-courses analyses. The two social psychological theories that credit grade influences on ratings to irrational, motivated judgment processes are quite consistent with radiation of the halo effect to peripheral judgments.

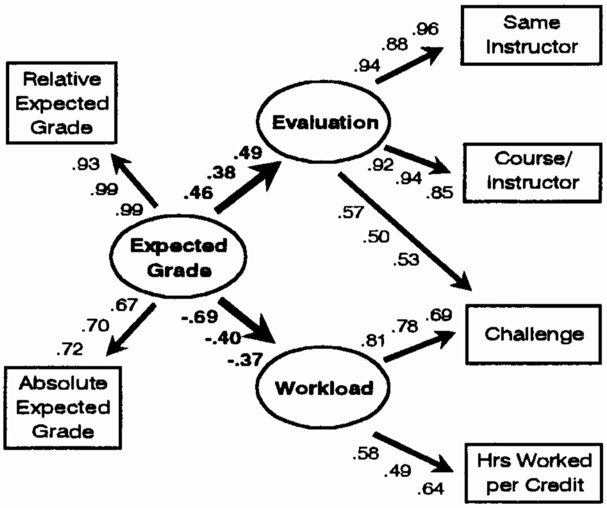
4. Negative grades–workload relationship between classes. It seems reasonable that students should work harder in courses in which they receive high grades than in ones in which they receive low grades. The reasonableness of this expectation rests on two assumptions: (a) that grades awarded in a course provide an indicator of students' achievement or learning in the course and (b) that students work harder in courses in which they learn much than in courses in which they learn little. From these two assumptions, it follows that students should tend to work harder in courses that give high grades than in courses that give low grades. However, in data obtained repeatedly at University of Washington, this expected positive relationship between grades and course workload was not found. To the contrary, the data repeatedly revealed a substantial negative relationship between expected course grades and workload—students reported doing more work in courses that had low expected grades than in courses that had high expected grades. This relationship, based on data obtained at University of Washington in three terms of the 1993–1994 academic year, is shown in the structural equation model depicted in Figure 3. Tests of the expected grades–workload relationship have rarely been reported in previous research. However, other studies have indeed observed the same surprising negative relationship between expected grades and workload in between-courses analyses (e.g., Marsh, 1980, pp. 234–235).

Figure 2
Effect of Grades on Items That Appear Peripheral to the Construct of Quality of Instruction



Note. Results are reported as beta coefficients, which provide an effect-size measure that is standardized relative to individual subjects' variability. Data are from 66 courses (those that had data from more than 10 respondents) at University of Washington in the winter of 1994. Total sample sizes ranged from 1,558 to 1,610 for the various analyses. Within-courses coefficients were estimated after fitting section means of the dependent variable to the data. The horizontal rectangle includes beta values that were not different from zero by a conservative statistical criterion (α = .005, two-tailed).

Figure 3
Structural Equation Model Replicated on Three Data Sets From the 1993–1994 Academic Year



Data Set	*N*	χ^2	*df*	*p*	rmsea	*P*(close fit)
Autumn '93	205	11.50	6	.07	.07	.27
Winter '94	205	6.40	7	.49	.00	.77
Spring '94	184	5.73	7	.57	.00	.80

Note. The *challenge* and *hrs worked per credit* measures are based, respectively, on Items 20 and 26 of University of Washington's Form X (see Gillmore & Greenwald, 1994). The negative between-courses relationship between Expected Grade and Workload is measured by the standardized coefficients (*M* = −.45) for the path linking their latent variables. See Figure 1's caption for notes on interpretation of goodness-of-fit statistics. Hrs = hours.

All three third-variable theories imply nonnegative relationships between expected grades and workload. This is most readily seen for the two motivational theories. If students earn high grades by virtue of high motivation (i.e., by working hard), then a positive relationship between expected grades and workload is clearly anticipated. For the teaching-effectiveness theory, it might be assumed that effective teachers manage to get their students to do more work, and, thus, if high grades are explained by effective teaching, a positive relationship between expected grades and workload is anticipated. If, however, it is assumed that effective teachers are just more efficient in imparting knowledge to students, then expected grades should be unrelated (but not negatively related) to workload. The attribution theory appears to be irrelevant to the grade–workload relationship because it is equally possible to explain a high grade as being due to hard work as it is to explain hard work as being due to a low expected grade. Therefore, the attributional link between judged workload and expected grade is quite uncertain. Only the leniency theory readily explains the observed negative relationship. The explanation is that strict-grading instructors induce students to work hard in order to avoid very low grades. These interpretations of the relationship between expected grades and workload have been described in more detail by Greenwald and Gillmore (in press).

Summary Evaluation of the Five Theories

Each theory predicts a different subset of the four diagnostic data patterns presented in Table 1, ranging from the teaching-effectiveness theory predicting none of them to the leniency theory predicting all of them (see Table 1). Each of the three third-variable theories fails to explain at least two of the four findings. The two direct-cause (grades influence ratings) theories fare best as a class, and of these two, the leniency theory is favored by virtue of being the only theory to explain the negative relationship between grades and workload.

Conclusions

Yes, I Can Get Higher Ratings by Giving Higher Grades

Recall that this conclusion has been previously supported by experimental studies in which grading policies were manipulated in natural classroom settings (see also Greenwald, 1997, this issue). Figures 1 and 3 suggest that the magnitude of this effect corresponds to a standardized path coefficient as high as .50. In the context of the grading-leniency interpretation, this .50 figure means that in the population of courses included in the University of Washington data sets, changing from giving grades one standard deviation below the university mean to one standard deviation above should produce a one standard-deviation change in one's percentile rank in the university's student ratings. A standard-deviation change from, say, half a standard deviation below the university mean rating to half a standard deviation above would be a change from the university's 31st percentile of instructors to the 69th percentile. Giving high grades, by itself, might not be sufficient to ensure high ratings. Nevertheless, if an instructor varied nothing between two course offerings other than grading policy, higher ratings would be expected in the more leniently graded course.

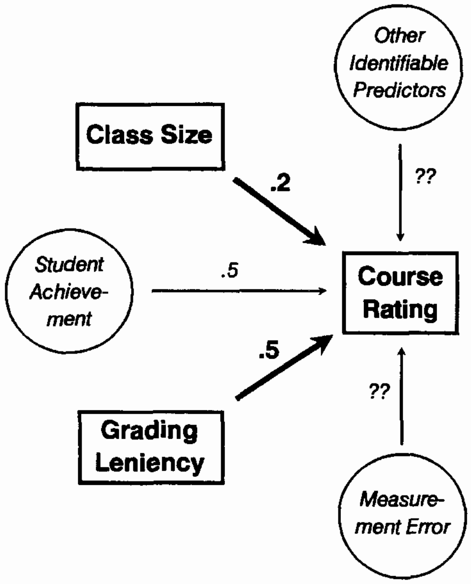
Discriminant Invalidity Plus Convergent Validity

The apparent effect of grading leniency on ratings, as documented in the University of Washington studies, indicates that student ratings have a failing of discriminant validity. At the same time, student ratings have repeatedly been shown to have moderate convergent validity. In other words, at the same time that student ratings provide a distorted measure of instructional quality, they also have a valid correlation with instructional quality. Figure 4 presents a theoretical model consistent with student ratings having both discriminant invalidity and convergent validity.

Making Student Ratings More Useful

Figure 5 presents actual ratings data to which adjustments have been applied on the basis of the model in Figure 4. These adjustments can be seen to have shifted the relative standing of courses up or down by more than three deciles for about 10% of the sample of courses. Note, for example, that courses very near the median before adjustment are distributed from the highest to the lowest decile after adjustment.

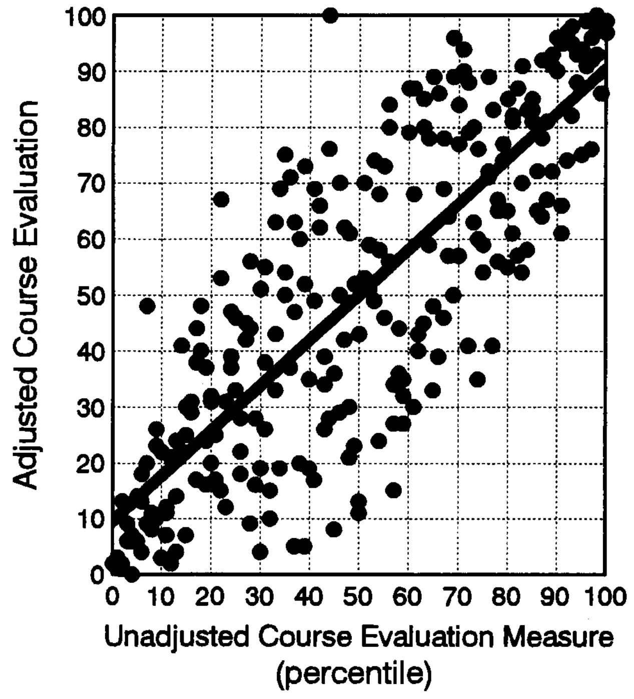
Figure 4
Discriminant Invalidity With Convergent Validity



Note. A model of influences on student ratings, including both a valid influence of teaching effectiveness (correlation with student achievement—convergent validity) and unwanted influences of grading leniency and class size (discriminant invalidity). Rectangles indicate measured variables. Circles indicate unmeasured variables.

Figure 5

Example of Adjustment of Actual Ratings Data to Reduce Unwanted Influences of Grading Policy and Class Size



Note. Data are from University of Washington, winter quarter of 1994 (*N* = 254 courses).

Workload Measures Are Useful

The consistent finding of a negative relationship between course grades and workload (illustrated in Figure 3) is disturbing. This relationship has never become a focus of research attention, probably because workload measures are not included in many course rating forms. The inclusion of workload estimates in course evaluation forms, as advocated by Greenwald and Gillmore (in press), can ensure that this important aspect of differences among courses does not continue to escape attention.

The Baby and the Bathwater

The results reported in this article might be regarded as sufficient reason to abandon the entire enterprise of collecting and reporting student ratings. However, there are three good reasons to conclude just the opposite—that student ratings measures deserve increased attention.

First, in many cases there is no readily available alternative method of evaluating instruction. Although expert appraisals and standardized achievement tests might provide more valid assessments, regrettably both of those alternatives greatly exceed student ratings in cost. The present limited use of such alternatives may indicate their relative impracticality.

Second, although the influence of grading leniency means that student ratings have a deficiency in discriminant validity, the evidence for convergent validity of student ratings cannot and should not be dismissed. As illustrated in Figure 5, theory-based statistical adjustments can increase the usefulness of that information.

Third, student ratings almost certainly contain useful information that is independent of their correlation with student achievement. That is, student ratings provide information about how well students like a course. This assessment of liking or attitude can be very useful, in the same way that an assessment of bedside manner is useful in evaluating a physician. The assessment of bedside manner may not describe the physician's success in preventing or curing illness, but it should predict patients' willingness to adhere to prescribed treatments and to return for future checkups. Similarly, knowledge of how much a teacher is liked should provide information that can predict a student's willingness to do assigned work and to register for further course work from that teacher.

In summary, there is an instructional-quality baby (convergent validity) in with the bathwater (discriminant invalidity) of grades–ratings correlations and other possible contaminants. It seems much wiser to give that baby a bath and make it presentable than to throw the baby out with the bathwater.

REFERENCES

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

Aronson, E., & Linder, D. E. (1965). Gain and loss of esteem as determinants of interpersonal attractiveness. *Journal of Experimental Social Psychology, 1,* 156–171.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chacko, T. I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly, 8*(2), 19–25.

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

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon Press.

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

Gillmore, G. M., & Greenwald, A. G. (1994, April). *The effects of course demands and grading leniency on student ratings of instruction.* Paper presented at meetings of the American Educational Research Association, New Orleans, LA.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35,* 603–618.

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

Greenwald, A. G., & Gillmore, G. M. (in press). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology.*

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology, 63,* 130–133.

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.

Knowles, E. S. (1982). A comment on the study of classroom ecology: A lament for the good old days. *Personality and Social Psychology Bulletin, 8,* 357–361.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17,* 219–237.

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

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

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218–1225.

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

Stumpf, S. A., & Freedman, R. D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology, 71,* 293–302.

Theall, M., Franklin, J., & Ludlow, L. (1990). Attributions and retributions: Student ratings and the perceived causes of performance. *Instructional Evaluation, 11,* 12–17.

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

Worthington, A. G., & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71,* 764–775.

Postscript

Comment by Greenwald

For the past two decades, the dominant view among researchers of student ratings has been that ratings provide valid and substantially bias-free measures of teaching effectiveness. The wide acceptance of this view is indicated by the prevalence of research efforts directed at establishing convergent validity, relative to ones directed at pursuing discriminant-validity criticisms.

It is all too easy for someone who had no role in constructing the dominant view to register dissatisfaction with it. With that advance confession of perspective, I confess further to some amazement at the respect accorded to this dominant view in the state-of-the-art articles by Marsh and Roche (1997, this issue), d'Apollonia and Abrami (1997, this issue), and McKeachie (1997, this issue) in this *Current Issues* section. These scholars, like many others, appear not to be disconcerted by convergent-validity findings that typically report correlations of .40 or less with nonratings indicators of teaching effectiveness (e.g., ratings typically explain only about 15% of the variance in achievement measures). I am similarly puzzled by the wariness with which these scholars and others treat research findings that, to my reading, indicate that ratings measures can be unfair. I have in mind not only evidence concerning grading leniency (with which this article is primarily concerned) but also evidence that students (in making ratings judgments) mistake enthusiastic teaching style for effective teaching (Abrami, Leventhal, & Perry, 1982) and that being assigned to teach large classes lowers one's ratings (we have very reliably found this effect at University of Washington). Our program of research at University of Washington definitely aims to upset the dominant view.

Comment by Gillmore

This article describes a means of adjusting student ratings for variations in grading leniency and possibly other variables, such as class size. D'Apollonia and Abrami (1997) explicitly recommend against this practice, worrying that "poor instructors may be doubly rewarded both by students and by evaluators" (p. 1205). Marsh and Roche (1997) make it clear that they oppose such adjustments as unnecessary and wrongheaded. McKeachie (1997) lists situations in which such adjustments would be inappropriate (p. 1222). McKeachie also acknowledges the existence of "cheating," referring to teachers who inflate grades to get high ratings. Although such cheaters appear clearly to deserve downward ratings adjustments, it would be a mistake to conceive our suggested adjustments as a moral response to this kind of abuse. Our purpose is different. Take two courses that differ only in that students in one expect higher grades than those in the other. Although (by assumption) students in both courses learn the same amount and receive the same quality of instruction, our reading of the research evidence indicates that the former course will receive generally higher evaluative ratings. Even excellent teachers whose outstanding pedagogy leads to high student achievement will receive elevated ratings if their students expect very high grades rather than just high grades. We advocate ratings adjustments for such situations.

Student instructional ratings provide data for three distinct functions: personnel decisions, instructional improvement, and information to students. It is to achieve fairness in personnel decisions that adjustments for grades, class size, and perhaps other variables are potentially most useful and justifiable. We do not think that teaching careers should be injured when faculty take on the difficult task of teaching large sections or when they uphold strict grading standards. The question is not whether adjustments will turn an imperfect measure into a perfect one but rather whether adjustments can improve decisions that must make use of a necessarily imperfect measure.

Comment by Greenwald and Gillmore

Because our research is previously unpublished and disagrees with the views of other authors in this *Current Issues* section, it is perhaps unsurprising that the other authors find our article problematic, such that their postscripts focus on this article. We respond here to those comments.

The postscripts by d'Apollonia and Abrami (1997) and by Marsh and Roche (1997) object on several grounds to our use of correlational evidence to discriminate among five theories of causal influences on student ratings. Both d'Apollonia and Abrami and Marsh and Roche object that our correlational data were equivocal—that is, not up to the task of choosing among the theories. Both d'Apollonia and Abrami and Marsh and

Roche also invite readers to persuade themselves of the ambiguity of our analyses by constructing explanations for all of the data patterns in our Table 1 from the hypothesis that "true student learning [causes] . . . both high ratings and high grades" (p. 1208). We wish readers good luck in this effort. Our own attempts to do just this were decidedly unsuccessful.

Marsh and Roche (1997) also observe that within-courses correlations of expected grades with ratings should be irrelevant to theories about between-courses differences in grading leniency. Our view is that the leniency theory differs from the other theories in its implications for the three within-courses correlation patterns shown in the middle three columns of Table 1. Furthermore, this article sought to show that progress in comparatively evaluating theories can be made when these differential implications are considered.

Referring to this article, Marsh and Roche (1997) observe that "Greenwald and Gillmore's (1997) critical variable should be grading leniency (not expected grades)" (p. 1197). Because we doubt that grading leniency can be measured directly in student ratings, we believe that the best way to study the effect of grading leniency, separated from the possible correlation of grades with teaching effectiveness, is by means of experiments with grading-leniency manipulations. As is noted in the postscript in Greenwald's (1997) article, Marsh and Roche (and also d'Apollonia and Abrami, 1997) regard the multiple existing natural classroom experiments that found effects of grading leniency on student ratings to be methodologically flawed.

Finally, Marsh and Roche (1997) say that we "inappropriately impl[ied] causation from correlation" (p. 1197) in concluding that instructors can increase their student ratings by grading more leniently. We believe, to the contrary, that we appropriately concluded that a causal effect of grading leniency is the best interpretation of the five grades–ratings data patterns that we reviewed. At the same time, we acknowledge that the statistical adjustment used in Figure 5 may not be the most satisfactory possible adjustment. Toward the goal of producing superior adjustments, we urge further research to identify additional biases in student ratings.