Journal of Educational Psychology
1997, Vol. 89, No. 4, 743-751

0022-0663/97/$3.00

No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction

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

Samples of about 200 undergraduate courses were investigated in each of 3 consecutive academic terms. Course survey forms assessed evaluative ratings, expected grades, and course workloads. A covariance structure model was developed in exploratory fashion for the 1st term's data, and then successfully cross-validated in each of the next 2 terms. The 2 major features of the successful model were that (a) courses that gave higher grades were better liked (a positive path from expected grades to evaluative ratings), and (b) courses that gave higher grades had lighter workloads (a negative relation between expected grades and workload). These findings support the conclusion that instructors' grading leniency influences ratings. This effect of grading leniency also importantly qualifies the standard interpretation that student ratings are relatively pure indicators of instructional quality.

Student ratings have been both praised as being valid and efficient and criticized as being insensitive and misleading.[1] The present research proceeds from an intermediate view—that student ratings may be imperfect but are nevertheless useful and are also improvable through research. The specific aim of the present research was to construct and confirm a covariance structure model that could identify sources of desired or undesired influences on student ratings.

The most ambitious previous effort to describe and confirm a covariance structure model of student ratings has been Marsh's (1991) hierarchical confirmatory factor analysis of data provided by 35 items of the SEEQ (Student Evaluations of Educational Quality) inventory. Marsh reported substantial confirmatory support for a nine-factor first-order structure overlaid with a four-factor structure in which the higher order factors represented similarity relations among the nine first-order factors.

In contrast with Marsh's (1991) aim of analyzing the dimensional structure of student ratings, the present research sought to evaluate theories of causal influences operating on student ratings. Alternative theories of the influences that affect student ratings imply different patterns of relationships among three categories of measures: (a) evaluative ratings, (b) expected grades, and (c) course workloads.

Thought Experiments and Structural Models

Imagine collecting data in a large set of courses that share the same well-defined educational goal—say, increasing students' foreign language vocabulary. Course performance will be graded, and students receive midterm exams that allow them to develop expectations about what their final grades will be. Student rating surveys are administered, as is typical, before the final exam is given and, therefore, before the final grade can be known. The rating surveys are assumed to produce three types of measures: (a) evaluative ratings of the course, (b) estimates of expected final grade, and (c) self-reports of amount of work done for the course.

In each of four replications of this thought experiment, only a single exogenous (causal) variable is assumed to operate. The four variables are (a) quality of instruction, (b) student ability, (c) student motivation, and (d) grading leniency.[2]

Anthony G. Greenwald, Department of Psychology, University of Washington; Gerald M. Gillmore, Office of Educational Assessment, University of Washington.

The research reported in this article was greatly facilitated by the Office of Educational Assessment at the University of Washington. Additional support was provided by Grant SBR-9422242 from the National Science Foundation, and Grant MH 41328 from the National Institute of Mental Health. For suggestions concerning statistical analyses, we thank Robert D. Abbott, Robert C. MacCallum, and Steven J. Breckler.

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, Guthrie Hall, Box 351525, University of Washington, Seattle, Washington 98195-1525. Electronic mail may be sent via Internet to Anthony G. Greenwald at agg@u.washington.edu or to Gerald M. Gillmore at oea@u.washington.edu.

[1] Reviews and empirical studies concluding in favor of validity of student ratings as measures of quality of instruction can be found, for example, in Cashin (1995), Cohen (1981), Feldman (1997), Howard, Conway, and Maxwell (1985), Howard and Maxwell (1980, 1982), Marsh (1980, 1982, 1984), Marsh and Dunkin (1992), and McKeachie (1979). Critical reviews and empirical critiques of the validity of ratings can be found, for example, in Chacko (1983), Dowell and Neal (1982), Holmes (1972), Powell (1977), Snyder and Clair (1976), Vasta and Sarmiento (1979), and Worthington and Wong (1979). Intermediate positions, suggesting cautious support for validity of ratings while also expressing concerns about the adequacy of that support, have appeared more occasionally (e.g., Abrami, Dickens, Perry, & Leventhal, 1980).

[2] Quality of instruction is judged to be the major influence on ratings in the reviews that have concluded in favor of validity of ratings (see Footnote 1). Student characteristics, including ability and motivation, were described as important additional influences

The thought experiments are constructed so that intercorrelations among their four variables can be caused only by direct or indirect effects of each experiment's single causal variable. The expected relationships shown in Figure 1 were generated from assumptions about direct effects of each experiment's causal variable, together with a few assumptions about the manner in which that variable would interact with students' grade aspirations. *Grade aspirations*—that is, students' expectations of the grade appropriate for their work in the course—were assumed to have two effects: (a) *work regulation*—students are assumed to adjust their work level if needed to achieve the aspired grade (e.g., they will work harder if they perceive themselves heading toward a below-aspiration grade), and (b) *grade satisfaction*—students' satisfaction with a course (expressed in their evaluative ratings) should reflect their performance relative to aspiration; ratings should be higher for students who expect to exceed aspiration than for those who expect to fall short.[3]

Quality of instruction differences between courses can lead to higher ratings by any or all of three routes. First, the instructor may succeed in getting students to work harder and thereby achieve more, leading to higher grades and higher evaluative ratings (through grade satisfaction). Second, the instructor may teach more efficiently, such that students achieve more, leading to higher grades and higher ratings even independently of work level. Third, ratings may be directly responsive to quality of instruction (as they are intended to be). Assuming the presence of all three of these causal effects, the quality-of-instruction thought experiment should yield positive correlations among expected grade, reported workload, and evaluative rating measures. The correlations of workload with both expected grades and ratings for this thought experiment are indicated with double-ended curved arrows in the upper left panel of Figure 1 because they are expected to come about as indirect consequences of other (causal) relations.

Student motivation differences should lead to differences in work and (therefore) achievement and expected grade. However, these motivation-caused higher expected grades will not necessarily lead to higher ratings (through grade satisfaction), because highly motivated students may have correspondingly high grade aspirations. Nevertheless, higher motivation may be associated with higher ratings if, as seems reasonable, highly motivated students also have a favorable attitude toward instruction. This experiment leads to the same expected positive intercorrelations among measures of expected grade, workload, and ratings as for the quality-of-instruction experiment, even though there is a difference in the underlying patterns of causation.

Student ability differences between courses should lead to higher grades by virtue of ability-related achievement differences. As was the case for the motivation experiment, these expected-grade differences do not lead to higher ratings, because higher ability students should have correspondingly higher grade aspirations. With no basis for predicting differences among courses in ratings or work, this experiment yields the expectation that evaluative ratings, expected grades, and measures of workload will be uncorrelated.

Grading leniency differences between courses lead directly to differences in expected grades, because lenient grading on midterm tests should create an expectation of higher final grades. Further, because leniency-caused high expected grades should exceed students' grade aspirations, grade satisfaction effects should result in a positive correlation between expected grades and evaluative ratings. Still further, work-regulation effects should lead students whose expected grades exceed aspirations to reduce work investment. Note that the work-regulation effect indirectly results in depressed student achievement associated with lenient grading, indicated with a negative-signed double-ended arrow in the lower right panel of Figure 1.

Diagnostic Value of Workload Measures

Many student rating surveys include measures of both evaluative ratings and expected grades. In the context of Figure 1's analysis, such surveys improve at least slightly over ones with evaluative ratings alone, because the observed correlation between ratings and expected grades can potentially distinguish situations in which observed grade differences between courses are due to preexisting ability differences from ones in which they are due to student motivation, quality of instruction, or grading leniency. The availability of a workload measure adds the capability of distinguishing situations in which observed grade differences are due to variations in grading leniency. When expected grade differences are due to varying leniency—and only in this case—there should be a negative correlation between expected grades and course workloads. The correlation of workloads with expected grades is therefore of substantial value in distinguishing among sources of influence on evaluative ratings.

on ratings in, among others, the articles by Howard and Maxwell (1980) and Marsh (1984). Grading leniency has been suggested as a major influence in published critiques of ratings validity (see Footnote 1), and has been described as at least a minor influence in most of the reviews that have been favorable to ratings validity.

[3] In published discussions of student ratings, grade satisfaction has most often been treated without specifying whether it was assumed that students derive satisfaction from absolute levels of performance or from their performance relative to aspiration. In generating expected results for the thought experiments, grade satisfaction was interpreted in relative-to-aspiration form for three reasons: (a) social psychological theories of social comparison (Festinger, 1954), level of aspiration (Lewin, Dembo, Festinger, & Sears, 1944), and social exchange (Thibaut & Kelley, 1959) all use an assumption that outcomes are evaluated by comparison with some expectation or standard; (b) it seems intuitively obvious that relation of expected grades to aspiration is essential to satisfaction—for example, two students, one with a B− average and the other with an A− average, should not be equally satisfied with a B+ grade; and (c) data from smaller studies preliminary to the present ones indicated that expected grades had stronger correlations with rating measures when expected grades were assessed in relative-to-aspiration form rather than in absolute form (this finding was also observed in the present research).

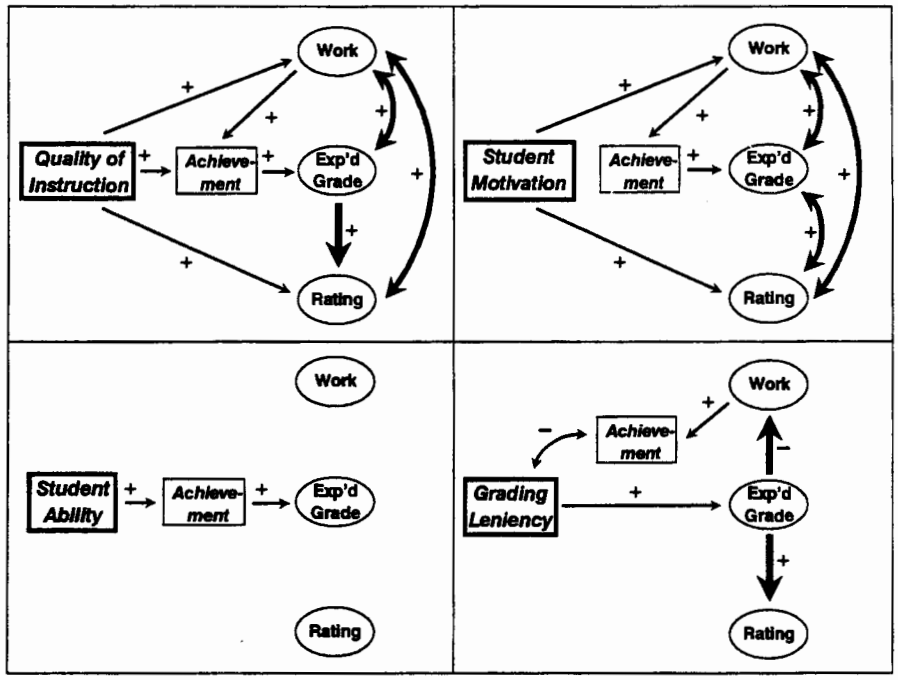

Figure 1. Structural models of thought experiments. The exogenous (causal) variable of each experiment, named in italics in the rectangle at the left of each panel, is assumed to be the only factor that can produce differences among experiments in correlations among evaluative ratings, reported workload, and expected grade. Thick arrows indicate observed relationships. Assumed (but unobserved) relationships involving the exogenous variable and (unmeasured) achievement are indicated by thin arrows. Single-ended arrows represent assumed causal links between variables; double-ended curved arrows indicate expected correlations that should result indirectly from other relationships. Exp'd = expected.

It is unlikely that, in any actual student ratings data set, only one of Figure 1's four causal variables would be active. Despite this likely multivariate causal complexity of actual ratings data, the analysis developed in Figure 1 is still useful. To the extent that one observes positive correlations between measures of workloads and expected grades, one may conclude that quality of instruction or student motivation differences are operating. To the extent that these same correlations are negative, one may conclude that grading leniency differences between courses are operating.

Method

The research reported in this article used student ratings data obtained at University of Washington to assess the sources of influence diagramed in Figure 1. This required adding workload measures to student ratings surveys previously in use at University of Washington.

Recent History of Student Ratings at University of Washington

Between 1974 and 1995, the University of Washington used a family of five (or more, at times) 22-item survey rating forms. A set of 11 evaluative items was common to all of these forms. The remaining 11 items on each form were tailored to the specific instructional setting in which each form was used, such as lecture, seminar, recitation, or laboratory. Factor analyses of these 22-item surveys repeatedly revealed them to be dominated by a single evaluative factor.[4] Beginning in 1992, the present authors sought to determine the extent to which this evaluative factor was influenced by instructor grading policies. The first study was a pilot investigation, conducted with the cooperation of a small sample of Psychology Department instructors who agreed to add some items to their regular course evaluation surveys.

Results of the initial pilot study (reported by Greenwald, 1992) led the authors to develop the rating form that was used in the university-wide research reported here. This form, identified as Form X, was introduced at University of Washington in Autumn 1993 as an "experimental" form that was made available to all faculty as an alternative to the existing student rating forms. Although University of Washington does not require faculty to obtain student ratings, nevertheless faculty do have some incentive to use the available student rating forms. Perhaps most significantly, all faculty are required, by legislation of the university's Faculty Senate, to report at least one formal evaluation of instruction during an academic year in order to be considered either for promotion or for merit raise in salary.

Procedures

University of Washington uses an academic calendar with four 3-month terms (quarters) per year. Similar procedures were used in

[4] These unpublished findings are available from University of Washington's Office of Educational Assessment.

the three consecutive terms (Autumn 1993, Winter 1994, and Spring 1994) during which the present data were collected. Early in each quarter, all instructors responsible for university courses at any level were offered the opportunity to request use of student ratings, and were given a choice that included Form X along with the family of other forms that were previously in use. In the first quarter, all instructors also received a memorandum that described the intent of using data obtained with Form X as a basis for improving the student rating system at University of Washington.

Analyses used individual courses as the units of analysis, limiting eligibility for inclusion to courses that used numerical grading, had undergraduate enrollments, were offered for three, four, or five credits, and had provided Form X data for at least 10 students. Table 1 provides some descriptive information for samples of courses obtained for each of the three quarters.

Measures

In reaction to the domination of the university's previous student rating surveys by a single evaluative factor, Form X was designed with items in separate sections that differed in response format. All items were scored as course medians, computed using a standard algorithm that treats tied scores as if they were uniformly distributed across an interval ranging from the tied value's boundary with the next lower value to its boundary with the next higher value (Guilford, 1965, pp. 49–53).

Course/Instructor. The first 11 items on Form X requested characterization of the frequency of 11 desirable characteristics of the course and instructor, such as "The instructor gave very clear explanations" and "Extra help was readily available." Each characteristic was rated in response to the question "How frequently was each of the following a true description of the course?" on a 7-point scale that was anchored by *never* (0), *about half* (3), and *always* (6). On the basis of correlations among these items, 7 of the 11 were retained and averaged to compose the Course/Instructor subscale. Correlations among the 7 retained items ranged from .60 to .94.

Self/Progress. Student learning outcomes were judged on 7 items in response to the stem, "Relative to other college courses you have taken, how would you describe your progress in this course with regard to. . . ." The rated outcomes included "learning the conceptual and factual knowledge of the course" and "developing an ability to express yourself in writing or orally in this field." Each outcome was rated on a 7-point scale that was anchored by *none* (0), *average* (3), and *great* (6). The correlations among these items ranged from .66 to .87. All seven items were averaged into the Self/Progress scale.

Same Instructor. A measure of appreciation for the instructor was obtained from responses to "If you had it to do over again and this course was optional for your program, would you enroll in it if the same instructor taught it?" Responses to this single-item measure were obtained on a 7-point scale that was anchored by *certainly not* (0), *neutral* (3), and *certainly* (6).

Absolute Expected Grade. For numerically graded courses at University of Washington, assigned grades range from 0.0 (E or failing) to 4.0 (A), with all tenth-of-a-point values possible except those between 0.0 and 0.7 (D−). One item on Form X asked students to indicate their expected grade by choosing among 11 options that covered mutually exclusive subranges of the possible grade range.

Relative Expected Grade. A second expected-grade measure was developed in small-sample data collections in the two academic years preceding the introduction of Form X. These pilot data repeatedly suggested the usefulness of items that asked students to characterize expected grade in relation to their average of previous work. In response to the stem, "Relative to other college courses you have taken do you expect your grade in this course to be . . . ", students responded on a 7-point scale that was anchored by *much less* (0), *average* (3), and *much greater* (6).

Workload. The major workload item asked students to estimate the number of hours per week that they "spent on this class, including attending classes, doing readings, reviewing notes, writing papers and any other course related work." The 12 response options, which ranged from "under 2" to "22 or more," were converted to an Hours Worked per Credit index by dividing number of hours worked by the number of credit hours associated with the course. Form X included three other items intended to assess various aspects of students' work investment in their courses. These requested 7-point relative-to-other-college-courses-you-have-taken ratings for "the intellectual challenge this course presented," "the amount of effort to succeed in this course," and "your involvement in this class (doing assignments, attending classes, etc.)."

Table 1
Characteristics of the Three Data Sets

Characteristics	Autumn 1993	Winter 1994	Spring 1994
No. of courses using student ratings	2,799	2,713	2,541
No. of courses using Form X	337	297	264
Mean (SD) enrollment of courses using Form X	43.4 (60.4)	41.2 (51.9)	34.5 (36.6)
Mean (SD) no. of respondents in courses using Form X	28.2 (35.7)	27.4 (33.0)	20.2 (17.6)
No. of Form X courses meeting inclusion requirements[a]	205	205	184

[a]Use of numerical grading; undergraduate enrollments; three, four, or five credits; at least 10 complete Form X responses.

Results

Exploratory Analyses of Autumn 1993 Data

Table 2 provides the matrix of correlations among the three evaluative measures (Course/Instructor, Self/Progress, and Same Instructor), the two measures of expected grade (Absolute Expected Grade and Relative Expected Grade), and the four measures related to workload (rated Challenge, Effort, and Involvement and estimated Hours Worked per Credit). The three evaluative measures were highly intercorrelated (all $rs \geq .80$), as were the two expected grade measures ($r = .67$). Intercorrelations among the four workload measures were all positive and statistically significant (all two-tailed $ps \leq .005$) but were variable in magnitude, ranging from $r = .21$ to .77. It can be seen in Table 2 that expected grade measures were (a) positively correlated with Evaluative rating measures at moderate levels (ranging from

Table 2
Autumn 1993 Intercorrelations Among Major Measures

Measure	Evaluation			Expected grade		Course workload			
	1	2	3	4	5	6	7	8	9
1. Course and Instructor	—								
2. Self/Progress	.88	—							
3. Same Instructor	.87	.80	—						
4. Absolute Expected Grade	.33	.40	.37	—					
5. Relative Expected Grade	.34	.42	.42	.67	—				
6. Challenge	.31	.33	.24	−.16	−.29	—			
7. Effort	.04	.11	−.03	−.35	−.46	.77	—		
8. Involvement	.24	.34	.25	.05	.17	.37	.40	—	
9. Hours Worked per Credit	−.22	−.21	−.26	−.23	−.38	.33	.48	.21	—

Note. $N = 205$ courses. For $r \geq .20$, two-tailed $p \leq .005$.

$r = .33$ to .42) but (b) negatively related with three of the four workload measures (Involvement was the exception). Correlations between Workload and Evaluative ratings measures were not consistently positive or negative.

Covariance structure modeling of the Autumn 1993 data was conducted in an exploratory fashion, using the CALIS module of the SAS statistical package, with individual course data as input and maximum likelihood estimation. The exploratory strategy was first to associate the three evaluative ratings measures, the two expected grade measures, and the four workload measures with three corresponding latent variables or factors, respectively labeled *Evaulation, Expected Grade,* and *Workload.* Second, relations among these factors were modeled with the restriction of using at least two measures for each latent variable. In general, (a) better fits were obtained when latent variables were represented by two, rather than by three or four measures; (b) good fits were achieved when the Evaluation factor was represented by any two of its three measures; and (c) the three rating measures of Workload (Challenge, Effort, and Involvement) could be included successfully in a structural model only by associating them with both the Evaluation and Workload factors, rather than with the Workload factor alone.[5]

The model shown in Figure 2 is the best fitting of those for which each latent variable was represented by two measured variables. Good fit of this model is indicated by both the nonsignificant chi-square value and by the low root-mean-square error of approximation (*rmsea*) index.[6] Figure 2's model treats Expected Grade as an exogenous factor that influences both Evaluation and Workload. For any structural model (that is, any model that specifies paths linking the three latent variables), there were nine related measurement models in which each latent variable was represented by two measures. These nine models were constructed by using one of the three possible pairs of measures from the set of three measures of Evaluation, in combination with one of the three pairs of Workload measures that could be formed by pairing one of the three rating measures of workload with the Hours Worked per Credit measure. Fits that were close to fair or better (rmsea < .08) were obtained for obtained for five of the other eight similar models.[7]

Additional exploratory analyses examined alternative possibilities for identifying relationship structures involving the three latent variables. All but one alternative structure fit very poorly. The one less-than-terrible alternative model was one that reversed the directions of the two structural paths of Figure 2's model. That is, both Evaluation and Workload were in the role of exogenous factors, each having a structural path to Expected Grade. For the version of this reversed model that used the same six measures as in Figure 2, $\chi^2(6, N = 205) = 18.9$, $p = .004$, rmsea = .103 (indicating poor fit). The values of path coefficients for this reversed model were similar to those shown in Figure 2 (but, of course, the path directions were reversed). The possibility of developing a plausible interpretation for this alternative structural model is considered in the *Discussion* section.

Cross-Validation: Winter and Spring 1994 Data Sets

Figure 2's model was arrived at in partly exploratory fashion, and had one notable ad hoc feature—its link of the

[5] At the suggestion of a reviewer of an earlier draft, exploratory factor analyses were conducted to assess the possibility of identifying a measurement model in which Challenge, Effort, or Involvement could be associated with one latent variable, rather than two. However, regardless of rotation strategy and selective elimination of other variables, each of these measures invariably had significant loadings on two factors, one defined primarily by evaluative ratings measures included in the analysis and the other by the Hours Worked per Credit measure.

[6] *rmsea* is the root-mean-square error of approximation fit index that has been described by Browne and Cudeck (1993) and MacCallum, Browne, and Sugawara (1996). These authors characterize rmsea < .05 as indicating close fit, .05–.08 as close to fair fit, .08–.10 as mediocre fit, and rmsea > .10 as poor fit.

[7] In the *Discussion* section, it will be noted that there are several structural models that are statistically indistinguishable from the one shown in Figure 2. The bearing of these alternatives on interpretation will be considered in the *Discussion* section.

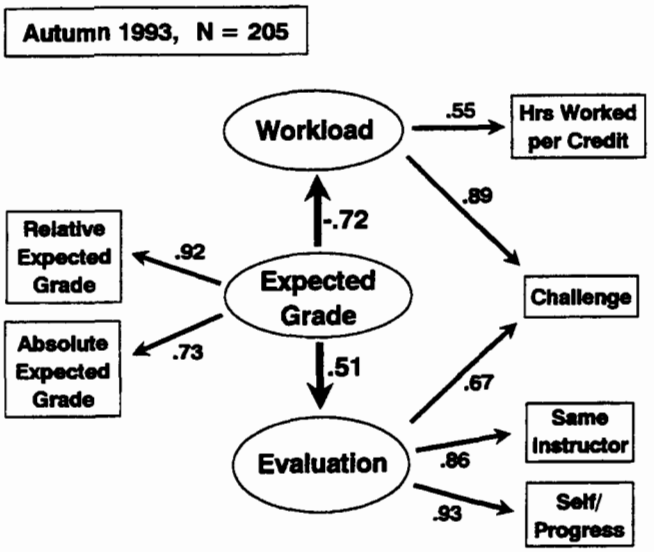

Figure 2. Structural model for Form X data of Autumn 1993. The model shown is the best fitting of nine models having the same structural paths among latent variables, but different pairs of measures used to represent the Evaluation and Workload factors. $\chi^2(6) = 8.19, p = .22$, rmsea = .042.

Challenge measure to two latent variables, Evaluation and Workload. Consequently, acceptance of Figure 2's model should depend on successful cross-validation. The next step in research was therefore to test Figure 2's model in confirmatory fashion on the data for the Winter and Spring terms of 1993–1994. Table 3 shows the standardized structural path coefficients and fit characteristics obtained from these confirmatory analyses, together (for comparison) with those of the original model for Autumn quarter and one additional analysis that combined data for three quarters.

Table 3 indicates that fits for the confirmatory analyses ranged from close to mediocre (see Footnote 6). The higher chi-square and rmsea values for the two confirmatory analyses (relative to the initial exploratory analysis) are not surprising—the close fit of the exploratory analysis plausibly involved some degree of capitalizing on chance. It is reassuring, however, that the entire class of nine measurement variants on Figure 2's structural model fit reasonably (and equally) well in all three data sets. For Autumn, Winter, and Spring, respectively, two, three, and two of the nine measurement variants showed close fit by the rmsea index (rmsea < .05), and three of the nine showed close fit in the combined analysis of the three data sets.

Figure 3 presents the model that, overall, showed best fit in the collection of three data sets. This model, which replaced Figure 2's Self/Progress measure (for the Evaluation latent variable) with the Course/Instructor measure, achieved the rmsea < .05 criterion of close fit in the Winter and Spring data sets and also in the combined data set shown in Figure 3. The variant in which Effort replaced Challenge in Figure 3's model also achieved close fit (rmsea < .05) in the Winter, Spring, and combined data sets. Table 4 provides the correlation matrix, including the three measures not included in Figure 3's model, for the combined three-term data set.

Discussion

Support for Grading Leniency Model

The availability of a successful structural model for relations among evaluative ratings, expected grades, and course workloads provides a basis for tentative conclusions about influences on student ratings. The successful model (Figures 2 and 3) was similar to only one of the four thought-experiment patterns of Figure 1. Only the Grading Leniency model (lower right panel of Figure 1) contains the

Table 3
Confirmatory Analyses of Grading Leniency Model

Characteristic	Exploratory analysis: Autumn 1993[c]	Confirmatory analyses: Winter 1994	Confirmatory analyses: Spring 1994	Combined analysis: 1993–1994[d]
No. of courses meeting inclusion requirements[a]	205	205	184	594
Standardized coefficient for path: Expected Grade to Evaluation	.51	.44	.61	.44
Standardized coefficient for path: Expected Grade to Workload	−.72	−.48	−.46	−.49
χ^2 (df)[b]	8.19 (6)	15.7 (6)	16.2 (7)	7.99 (6)
p value	.22	.02	.02	.22
rmsea index of fit	.042	.089	.085	.024

[a]Use of numerical grading; undergraduate enrollments; 3, 4, or 5 credits; and at least 10 completed Form X responses. [b]The additional degree of freedom for the Spring 1994 data set is due to a constraint added in the computational routine in order to keep estimated error variances nonnegative. [c]This analysis is diagramed in Figure 2. [d]This data set is diagramed in Figure 3. However, Figure 3's model has replaced Figure 2's Self/Progress measure of Evaluation with the Course/Instructor measure.

Note. rmsea = Root mean square error of approximation (see Footnote 6 for interpretation).

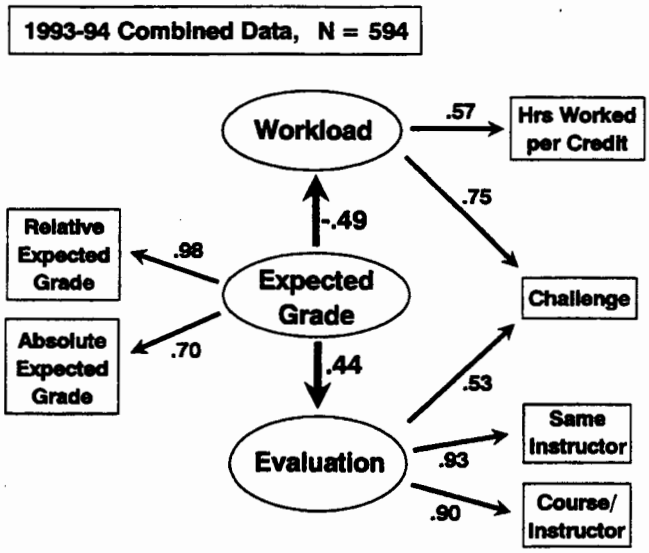

Figure 3. Structural model for Form X data combined over the three data sets (Autumn 1993, Winter 1994, and Spring 1994). The model shown is the best fitting of nine similar models that had the same structural paths among latent variables. It differs in just one measure from Figure 2's model (the Course/Instructor summary rating measure replaces the Self/Progress measure). $\chi^2(6) = 7.99$, $p = .24$, rmsea = .024.

negative path between Expected Grade and Workload that was so strongly evident in models that successfully fit the present data.

Alternative (Reversed-Path) Model

As noted previously, a structural model that reversed the directions of the two paths of the Grading Leniency model provided an approximately satisfactory fit for the Autumn 1993 data. Although statistical properties of this reversed-path model warrant giving it some consideration, a severe disadvantage of the model is the seeming impossibility of constructing a coherent interpretation of its structural paths. The positive (reversed) path from Evaluation to Expected Grade is, by itself, easily enough explainable—it can be understood by assuming that students who like a course will perform better in it, thereby earning better grades. However, a negative path from Workload to Expected Grade defies interpretation—that is, deciding to work hard in a course should lead to expecting a high grade, not a low grade.

Statistically Equivalent Structural Models

Figure 4 displays, alongside the successful model of Figures 2 and 3, four statistically indistinguishable alternative latent variable structures. The existence of such statistically equivalent models is routine in covariance structure modeling (see Breckler, 1990). Choices among such models can only be made on the basis of nonstatistical criteria such as plausibility. In the case of Figure 4's five statistically equivalent models, plausible interpretations can be constructed for four. Only the one that treats evaluative ratings as the sole exogenous variable seems patently implausible, chiefly because establishment of course workload and grading policy temporally precede course evaluation. The remaining four models share a directed path from Expected Grade to Evaluation but allow the causal direction of the link between Expected Grade and Workload to be expressed in either direction. It is, indeed, difficult to choose on plausibility criteria among these remaining four models.

Models (not shown in Figure 4) in which Workload has a direct connection (in either direction) to Evaluation fit much less well with the data than did the models in Figure 4. The collection of successful models shown in Figure 4 share the structural feature that the Expected Grade latent variable (i.e., instructor grading policy) occupies a central position. That is, all of these models have paths connecting Expected Grade to both Workload and Evaluation, and they have no path connecting Workload to Evaluation. This central position of the Expected Grade latent variable justifies maintaining the designation of Grading Leniency as a collective label for all of Figure 4's plausible models.

Table 4
Intercorrelations Among Major Measures (Combined Data for Entire 1993–1994 Academic Year)

	Evaluation			Expected grade		Course workload			
Measure	1	2	3	4	5	6	7	8	9
1. Course and Instructor	—								
2. Self/Progress	.82	—							
3. Same Instructor	.84	.79	—						
4. Absolute Expected Grade	.27	.36	.30	—					
5. Relative Expected Grade	.36	.47	.41	.68	—				
6. Challenge	.35	.38	.33	−.10	−.14	—			
7. Effort	.07	.14	.04	−.27	−.35	.74	—		
8. Involvement	.25	.39	.26	.06	.12	.49	.61	—	
9. Hours Worked Per Credit	−.12	−.05	−.11	−.15	−.28	.36	.57	.42	—

Note. $N = 594$ courses. For $r \geq .12$, two-tailed $p \leq .005$.

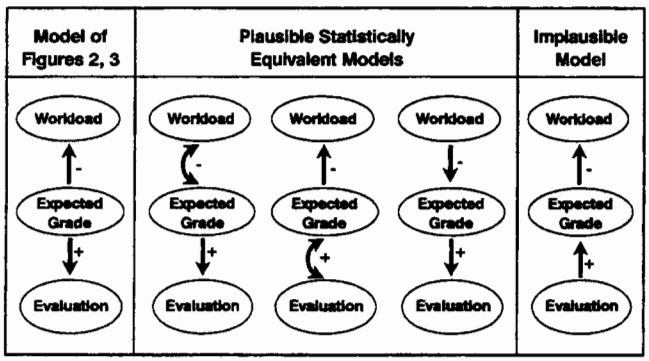

Figure 4. Statistically indistinguishable structural models of relationships among Expected Grade, Evaluation, and Workload.

Concerns About Generality of Findings

Authoritative reviews of research on student ratings (see references in Footnote 1) have concluded that instructor grading policy constitutes, at most, a minor influence on evaluative ratings. In conflict with those views, the Grading Leniency model of the present research implicates a strong influence of grading policy on student ratings. The Grading Leniency model also fits with past demonstrations that instructors can manipulate student ratings favorably or unfavorably by adopting strict or lenient (respectively) grading standards (see references in Footnote 1).

Although there is no reason to hesitate in concluding that the Grading Leniency model provides a good account for the three data sets investigated in the present research, the authors hesitate to generalize broadly on the basis of these data. The Grading Leniency model may be valid only for the specific instructional setting at University of Washington, and perhaps only for the subset of University of Washington courses that used Form X. Fortunately, concerns about generality of the present findings are empirically resolvable. Analyses resembling the present ones can readily be done at other institutions in which student rating surveys contain measures of course workloads and expected grades.

Disagreement With Prior Conclusions

Why does the present research disagree with previous reviewers' conclusions that grading leniency has no more than minor perturbing effects on ratings? An important part of the answer is that the present research was able to make central use of course workload measures. Figure 1 makes clear why course workload data can play so important a role in assessing the presence of grading leniency effects. In particular, finding a negative path between Expected Grade and Workload is a critical indicator of the causal effect of grading leniency. Some previous studies have reported negative correlations between expected grades and workload (e.g., Marsh, 1980, pp. 234–235). However, the full import of that negative correlation can become clear only when it is examined in conjunction with evaluative ratings data. Another plausible explanation for disagreement with prior conclusions is that the psychological properties of student ratings might have changed in the approximate two decades elapsed since the analysis and reporting of the data on which prior reviewers have based their conclusions. This interesting possibility might be evaluated by locating and reanalyzing older data sets that contain workload and expected grade measures along with student ratings. Unfortunately, the present authors did not have access to any such older data sets.

Refining Student Ratings

The present conclusions indicate a partial failure of discriminant validity for student ratings. That is, student ratings were found to be sensitive to something (grading leniency) that they are not intended to measure. To observe that ratings measures are thus contaminated does not mean that the ratings fail to measure what they are intended to measure. They may just be measuring more than they are intended to measure—in which case it can be well worth trying to purify or refine their measurement properties.

Two methods of refining student ratings measures can be pursued. The more obvious is to calculate and apply an adjustment for the contaminating effect of grading leniency. In Figure 3's model, Grading Leniency explained 20% of the variance of the Evaluation factor. Much of this unwanted influence of grading policy on ratings can be removed statistically by using expected grade measures as the basis for a covariance adjustment. The second and less obvious possibility follows from 80% of the variance in Evaluation being unexplained by the best-fitting structural model. Some fraction (perhaps large) of that unexplained variance can represent desirable outcomes of instruction.

Previous convergent validation studies (reviewed by Abrami, Cohen, & d'Apollonia, 1988) have found correlations averaging approximately $r = .40$ between evaluative ratings and measures of achievement in multisection validity designs. In these studies multiple sections of the same course receive grades based on the same or similar examinations, thereby controlling grading criteria. The $r = .40$ convergent validity figure may be seen either as an underestimate, because error of measurement and restriction of range of measures can attenuate validity estimates (Cohen, 1981, p. 301), or as an overestimate, because of uncontrolled third variables that might inflate validity estimates (Marsh & Dunkin, 1992, p. 170). Using a perhaps optimistic estimate of 20% of variance in ratings explained by the desirable correlation of student ratings with achievement differences, together with an approximate 20% explained by grading leniency, one still is left with an unexplained 60% of variance in student ratings. Even if this remaining 60% of ratings variance is uncorrelated with achievement, it may still be correlated with desirable attitudinal outcomes of instruction, such as liking for the course's subject and interest in further study.

References

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58,* 151–179.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evalua-

tions of instruction? *Journal of Educational Psychology, 72,* 107–118.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

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.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

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

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

Dowell, D. A., & Neal, J. A. (1982). A selective view of the validity of student ratings of teaching. *Journal of Higher Education, 53,* 51–62.

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

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

Greenwald, A. G. (1992). *Using student ratings to assess instructional quality.* Unpublished manuscript, University of Washington.

Greenwald, A. G. (1996). *Applying social psychology to reveal a major (but correctable) flaw in student evaluations of teaching.* Unpublished manuscript, University of Washington.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw Hill.

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., 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.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16,* 175–188.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and the behavior disorders* (pp. 333–378). New York: Ronald Press.

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. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74,* 264–279.

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. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83,* 285–296.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143–233). New York: Agathon Press.

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

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

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.

Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups.* New York: Wiley.

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.

Received September 24, 1996
Revision received April 25, 1997
Accepted April 25, 1997 ■