

Predicting Student Evaluation of Instructor and Course: Revisiting the Relationship Among Course Grades, Improving Teaching, and Summative Evaluation

C. S. Gaede, M. D. Svinicki, and D. M. Zimmaro

Historically, students' ratings of instruction have been used for three purposes: making personnel or administrative decisions, improving teaching, and guiding students' selections of courses. In practice, one instrument is often used for all three purposes. This paper presents an examination of the relationship between items on ratings instruments and the information they provide for each of these three purposes, as well as for overall ratings of teaching effectiveness. It will also present some hypotheses about other variables that might be affecting students' ratings.

At least as far back as 1979, McKeachie (1979) pointed out that different purposes require different scales or items. McKeachie said that general items are likely to be sufficient for personnel decisions, but more diagnostic items are necessary for improving teaching. Global, summative items, such as overall instructor effectiveness and the overall course satisfaction, are frequently used for an "administrative decision making" cluster. Expanding on McKeachie's work, Cohen (1981) identified six dimensions and their item characteristics for what might be thought of as an "improving teaching" cluster:

1. *Skill*. The Skill dimension encompasses the overriding qualities to which students respond when rating instructors. Typical items addressing this dimension concern competence: "The instructor has a good command of the subject matter"; "The instructor gives clear explanations"; "The instructor teaches near the class level."

2. *Rapport*. The Rapport dimension is addressed by items about a teacher's empathy, friendliness, approachability, and accessibility: "The instructor is friendly"; "The instructor is permissive and flexible"; "The instructor is available to talk with students outside of class."

3. *Structure*. The Structure dimension, concerning how well an instructor planned and organized the course, is typically addressed by items about administrative skills: "The instructor has everything going according to schedule"; "The instructor uses class time well"; "The instructor explains course requirements."

4. *Difficulty*. The Difficulty dimension, concerning the amount and difficulty of the work a teacher expects of students, is typically addressed by items about workload and timelines: "The instructor assigned difficult reading"; "The instructor asked for more than students could get done"; "This course required more work than others of comparable credit hours."

5. *Interaction*. The Interaction dimension, concerning the degree to which students are encouraged to share ideas and be active in class, is typically addressed by items about atmosphere: "The instructor encourages students to express various points of view"; "The instructor encourages students to volunteer their own opinions"; "The instructor facilitates classroom discussion."

6. *Feedback*. The Feedback dimension, concerning an instructor's concern with the quality of students' work, is typically addressed by items about communication: "The instructor tells students when they have done a particularly good job"; "The instructor checks to see if students have learned well before going on to new material"; "The instructor keeps students informed of their

progress.” (pp. 293-294)

However, the literature is silent with respect to developing a cluster of items for the goal of guiding students in course selection or educational experiences, a silence that may be explained by a general belief that students use responses to the “improving teaching” cluster and the global, summative items for this purpose.

Student ratings and teaching effectiveness

It is widely thought that student ratings of instruction are measures of teaching effectiveness. McKeachie’s (1979) definition of teaching effectiveness, “the degree to which one has facilitated student achievement of educational goals” (p. 385), has become widely accepted. Student achievement is normally thought of in terms of cognitive outcomes (McKeachie, 1979), which are operationalized in the form of final examination scores or final course grades. Yet, as measures of student achievement, final examination scores and final course grades have proven to have weak relationships to student ratings. With respect to course grades, McKeachie commented in 1979:

I doubt that student ratings will ever account for the majority of the variance between classes in student cognitive achievement. Most student rating forms ask students to evaluate *teaching*, not their own *learning*. (p. 386)

Student ratings and grading

Furthermore, many faculty members believe that they can manipulate student ratings to achieve positive outcomes by being lenient in assigning course grades. The practice of grading leniently to produce more positive student ratings is said to be a function of faculty rank, with assistant and associate professors, who are pursuing tenure and promotion, the most likely to engage in the practice. Full professors, who have achieved tenure and promotion goals, are

thought to be the least susceptible to the temptation of manipulating student ratings via grading leniency. Consequently, continuing to confound the interpretation of students' instructor ratings is the tension between the expectation that student achievement is reflected in final examination scores and final course grades and the counter possibility that final grades may be manipulated in order to obtain more positive student ratings.

Student ratings and course workload

In addition to grading leniency, course workload is a factor thought by some to bias student ratings and by others to be an important part of teaching effectiveness. One theory of student ratings of workload predicts a nonlinear relationship whereby ratings increase "as the workload increases to an optimal level, then flatten out or even decline for an excessive workload" (Marsh & Roche, 2000, p. 204). A nonlinear pattern in student ratings similar to the predicted pattern would indicate that students are responding based on their own level of satisfaction rather than on their expectation of a higher grade.

Student ratings and the unit of analysis

Relationships between variables and student ratings of instruction may be influenced by the unit of analysis chosen. It seems intuitive to strive for a student level unit of analysis. McKeachie (1979), on the other hand, said that "teachers may be differentially effective for different students" (p. 390), and "we can expect agreement only if we expect the instructor to be equally effective for all students, an assumption that research shows to be generally untrue" (p. 393). McKeachie makes a plausible point, so the most appropriate unit of analysis may be course-level data: mean ratings given by groups of students. Using course level data also allows researchers and administrators to obtain student ratings under conditions of anonymity, without any appearance of compromise.

Another way to interpret student ratings

Considering that final examination scores and course grades are proven to have weak relationships to student ratings, perhaps it would help to examine the issue from another point of view. Table 1 shows a hypothetical set of item clusters that might be representative of several important variables for assessing student ratings.

Table 1
A hypothetical set of item clusters representing different uses of student ratings

	Purpose of Rating Item		
	Cluster 1: Improving Teaching	Cluster 2: Administrative Decision Making	Cluster 3: Student Achievement
Items Included	<ul style="list-style-type: none">• Communication effectiveness• Value of course• Rapport with students• Course organization• Course difficulty• Effectiveness of assignments	<ul style="list-style-type: none">• Overall instructor rating• Overall course rating	<ul style="list-style-type: none">• Before course overall student GPA• Expected course grade• Actual average course grade

If an “improving teaching” cluster reflects teaching abilities that an instructor can improve over time, then the “administrative decision making” cluster and the “student achievement” cluster could be regarded as outcomes related to those abilities. Whereas the final course grade is an outcome an instructor assigns to a student, the overall ratings of the instructor and of the course are outcomes a student assigns to an instructor, which should be useful for administrative decision-making. Because administrative decisions quite frequently are based on two summative items pertaining to instructor and course, perhaps the question should be, “What is the relationship of items in the ‘improving teaching’ and the ‘student achievement’ clusters to the items in the ‘administrative decision’ cluster?” Asked differently, “Which items or what combination of items in the ‘improving teaching’ and ‘student achievement’ clusters best predicts item responses in the ‘administrative decision’ cluster?” It seems reasonable to inquire

about all factors that may influence ratings from overall instructor and course items, because those ratings appear to be administratively accepted as valid proxies of teaching effectiveness.

As helpful as discussions found in the literature may be, they often do not satisfy constituents in a local setting, who believe that their institution is different and does not fit the larger pattern. In order to obtain information specific to a single university, it was decided to investigate the validity of course-instructor survey items by determining which items from the “improving teaching” and “student achievement” clusters best predicted item responses in the “administrative decision” cluster, using course-instructor survey (CIS) data and student achievement data from a single Research 1 institution for the fall 2003 and spring 2004 semesters. In addition, the study sought to determine the influence of course grades and grading leniency bias on student responses to CIS items, in anticipation of developing an analytic model that could be used in a confirmatory study addressing these same issues, using CIS data and student achievement data from subsequent semesters. It was not the purpose of this study to investigate the relationship between ratings from the CIS instrument and students’ use of such information in course selection, other than to note the presence of items which could be used for that purpose.

Method

Sample Selection

The first step was to obtain data from institutional records for all courses surveyed during the fall 2003 and spring 2004 semesters and then match cases from the course-instructor survey data with student records data. Because responses to the course-instructor survey were anonymous, it was not possible to match cases on a student level; instead, cases were matched on the course level using the university’s five-digit unique number for courses.

The first data set containing 12,983 cases included graduate and undergraduate courses.

Several steps too detailed to be discussed here were taken to create a clean, unambiguous data set. The final step in the procedure was to differentiate among cases according to the rank of the course instructor—professor, associate professor, assistant professor, instructor, and lecturer—resulting in a final sample of 5,077 cases. Each case represented a single course taught by an individual instructor in a given semester. It should be noted that this means that a given instructor may have been included in the data set multiple times, depending on the number of courses he or she taught.

Procedure

Data analysis was exploratory: although reports in the literature provided helpful context for the exploration, the approach was trial and error, with the goal of developing a set of procedures that could be used in a systematic, confirmatory study using data from a subsequent academic year. The general process was an iterative cycle of exploratory analysis, reflection, reconsideration of the literature, and discussion with colleagues.

As a result of this exploratory process, a model for analysis emerged, summarized by eight procedures.

1. Partition items from the CIS Basic Form¹ to fit the three use clusters reflected in research literature, so that data analyses could be performed.
2. Compute correlation statistics for the relationships between all sets of CIS Basic Form items and class grade point averages, the latter computed by dividing the sum of the grades received in the course by the number of students graded, assigning 4 for an *A*, 3 for a *B*, and so forth. Do not include in the computations course withdrawals and credit/no credit grades.

¹ The CIS (Course Instructor Survey) Basic Form is the evaluation instrument used by the largest number of faculty members at the institution studied. A copy of the Basic Form may be found in the appendix.

3. Compute effect size differences for all faculty ranks using the group mean and standard deviation for the *total sample* as the comparison base.
4. Compute effect size differences for the ranks of associate professor, assistant professor, instructor, and lecturer, using the mean and standard deviation of the *rank of full professor* as the comparison base.
5. Compute regression coefficients, with the CIS Basic Form *overall instructor* item as the predicted variable and all other items, except the overall course item, as predictor variables.
6. Compute regression coefficients, with the CIS Basic Form *overall course* item as the dependent variable and all other items, except the overall instructor item, as the independent variables.
7. Compute ANOVA, with the CIS Basic Form item “workload” as the dependent variable and all other Basic Form items as independent variables.

These procedures provided several views of the data, allowing judgment of the appropriate weight of evidence in support of—or not—research hypotheses. While use of these procedures could not establish a conclusive cause and effect relationship, the cumulative weight of the evidence they provided facilitated useful interpretation.

Results

The CIS Basic Form items were sorted on the basis of the two ratings purposes “improving teaching” and “administrative decision” and the six dimensions suggested by Cohen (1981). The third category on the CIS Basic Form encompasses the items for the self-reported “student achievement” data. The description of this reorganization shown in Table 2 categorizes the Basic Form items by purpose, dimension, and student achievement report. Unless otherwise

noted, the rating scale for each item was 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, and 5 = *Strongly Agree*. Except for the dimension “skill,” the CIS Basic Form items all seemed to have a clear association with one of the two purposes, the dimension cluster with it, and student achievement report. In addition, the “mean course grade,” computed directly from university grade records, was included as a measure of student achievement.

Table 2
Sorting of CIS Basic Form Items by Purpose and Dimension

Purpose	Dimension	Basic Form Item
Instructor Development: to improve general teaching abilities over time (Cohen, 1980)		<i>Skill: the overriding quality to which students respond when rating instructors</i>
		2: The instructor communicated information effectively
		6: This course will be (or has been) of value to me
		<i>Rapport: an instructor’s empathy, friendliness, approachability, and accessibility</i>
		3: The instructor showed interest in the progress of students
		<i>Structure: how well the instructor planned and organized the course</i>
		1: The course was well organized
		<i>Difficulty: the amount and difficulty of the work</i>
		9: The workload was ... 5=excessive, 4=high, 3=average, 2= light, 1=insufficient
		<i>Interaction: students are encouraged to share ideas and be involved in class</i>
	5: The instructor made me feel free to ask questions, disagree, and express my ideas	
	<i>Feedback: the instructor’s concern with the quality of students’ work</i>	
	4: The tests and assignments were usually returned promptly	
Administrative decisions involving promotion, tenure, and merit (McKeechie, 1979)		<i>Instructor</i>
		7: This instructor was ... 1=very unsatisfactory, 2=unsatisfactory, 3=satisfactory, 4=very good, 5=excellent
		<i>Course</i>
	8: This course was ... 1=very unsatisfactory, 2=unsatisfactory, 3=satisfactory, 4=very good, 5=excellent	
Student achievement of educational goals (McKeechie, 1979)		<i>Overall grade point average</i>
		10: My UT GPA is ... 1=<2.00, 2=2.00-2.49, 3=2.50-2.99, 4=3.00-3.49, 5=3.50-4.00
		<i>Probable course grade</i>
		11: My probable grade is ... 1=A, 2=B, 3=C, 4=D, 5=F
		<i>Course grade</i>
	Mean course grade data for all students in one class from university records	

Next, correlations among the items related to the “improving teaching,” “administrative decisions,” and “student achievement” clusters were computed. The large N associated with this analysis affected the capacity to interpret the correlation values computed. Zero-order

correlations are shown in Table 3.

Inspection of the correlations of the “improving teaching” and “administrative decision” items with the “student achievement” items prompts five observations.

- Correlations with students’ self-reported cumulative grade point averages are very low across all items.
- Concerning students’ predicted probable course grades and mean course grade point averages, correlations with the “improving teaching” and “administrative decision” items are low to moderate and tend to parallel each other, which would be expected in view of the correlation of 0.79 between the students’ probable course grades and mean course grades.
- For all items, correlations with students’ probable course grades are higher than those with mean course grade point averages.
- Correlations between rating for course workload and all other items, except students’ self-reported cumulative grade point average, are low and negative.
- Concerning correlations between “improving teaching” items and “administrative decision” items, moderate to high correlations were found with all items, except with course workload.

In summary, as outcome measures, the “student achievement” items had a weak relationship with items related to both of the other two clusters, and the “administrative decision” items had a strong relationship with the “improving teaching” items.

To determine ratings differences associated with instructor rank, effect sizes² were computed using the mean and standard deviation for all ranks as the comparison group and then

² Effect size in this context would be the difference between the rank mean and the group mean divided by the group standard deviation.

performing a second analysis using the mean and standard deviation for the professor rank as the comparison group (Tables 4 and 5).

As shown in Table 4, the effect size for the rank of professor was consistently below the group mean and the rank of assistant professor exhibited the strongest positive effect size. That trend holds in general for the "improving teaching" items, the "administrative decision" items, and the "student achievement" items. The lecturer rank effect sizes were closest to the group mean, followed by the associate professor effect sizes. The instructor effect sizes exhibited wide variation, with mostly negative effect sizes for the "improving teaching" and "administrative decision" items and positive effect sizes for the "student achievement" items.

Concerning comparison of the effect sizes of student ratings for faculty ranks compared to the mean for the rank of professor (Table 5), it is noteworthy that the rank of assistant professor exhibited the largest positive effect size on items related to all three clusters. However, the rank of instructor exhibited the largest effect size for mean course grade.

For the next analysis, multiple regression was used to identify the best predictors for the "administrative decision" cluster: overall instructor and overall course items. In both cases (Tables 6 and 7), the best predictors were items from the "improving teaching" cluster. As shown in Table 6, instructor skills items concerning "communication" and "interest" were the two best predictors of overall instructor rating, though it seems likely that the instructor skill item concerning "value" could be included as a third predictor. The best predictors of overall course rating were instructor skills items concerning "value" and "communication" (Table 7). It is noteworthy that the mean course grade point average was not a significant predictor, indicating that overall evaluations are not influenced by grades for the course as a whole.

Table 3
Correlation Values for All Basic Form Items Plus Mean Course Grade Point Average for Fall 2003–Spring 2004

		Improving Instruction							Administrative Decision		Student Achievement		
		Organized	Communicated	Interest	Promptness	Expression	Value	Workload	Instructor	Course	My GPA	Prob Grade	Class GPA
Organized	Pearson Correlation	1	.836	.615	.613	.587	.736	-.082	.793	.774	.058	.217	.157
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	5075	5075	5075	4935	5075	5075	5075	5075	5075	4939	4933	5075
Communicated	Pearson Correlation	.836	1	.778	.539	.761	.855	-.127	.933	.877	.073	.357	.309
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5076	5076	5076	5075	4940	4934	5076
Interest	Pearson Correlation	.615	.778	1	.481	.841	.790	-.070	.846	.789	.106	.448	.433
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5076	5076	5076	5075	4940	4934	5076
Promptness	Pearson Correlation	.613	.539	.481	1	.457	.493	-.131	.543	.533	.072	.181	.137
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	4935	4935	4935	4935	4935	4935	4935	4935	4935	4934	4928	4935
Expression	Pearson Correlation	.587	.761	.841	.457	1	.735	-.167	.809	.745	.052	.401	.361
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5076	5076	5076	5075	4940	4934	5076
Value	Pearson Correlation	.736	.855	.790	.493	.735	1	-.051	.880	.928	.104	.398	.363
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5077	5077	5077	5076	4940	4935	5077
Workload	Pearson Correlation	-.082	-.127	-.070	-.131	-.167	-.051	1	-.107	-.130	.055	-.255	-.163
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5077	5077	5077	5076	4940	4935	5077
Instructor	Pearson Correlation	.793	.933	.846	.543	.809	.880	-.107	1	.920	.110	.396	.355
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000
	N	5075	5076	5076	4935	5076	5077	5077	5077	5076	4940	4935	5077
Course	Pearson Correlation	.774	.877	.789	.533	.745	.928	-.130	.920	1	.110	.444	.391
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	5075	5075	5075	4935	5075	5076	5076	5076	5076	4939	4934	5076
My GPA	Pearson Correlation	.058	.073	.106	.072	.052	.104	.055	.110	.110	1	.363	.327
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	N	4939	4940	4940	4934	4940	4940	4940	4940	4939	4940	4934	4940
Prob Grade	Pearson Correlation	.217	.357	.448	.181	.401	.398	-.255	.396	.444	.363	1	.788
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
	N	4933	4934	4934	4928	4934	4935	4935	4935	4934	4934	4935	4935
Class GPA	Pearson Correlation	.157	.309	.433	.137	.361	.363	-.163	.355	.391	.327	.788	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	
	N	5075	5076	5076	4935	5076	5077	5077	5077	5076	4940	4935	5077

Table 4
Effect Size for Basic Form Items by Faculty Rank Compared to Group Mean for All Ranks

Basic Form Items by Cluster	Professor	Associate	Assistant	Instructor	Lecturer	Mean Effect
Improving Instruction						
Organized	-0.12	-0.02	0.24	-0.24	0.07	-0.01
Communicated	-0.13	0.08	0.25	-0.22	0.05	0.01
Interest	-0.18	0.05	0.32	-0.11	0.08	0.03
Promptness	0.01	-0.06	-0.02	-0.02	0.03	-0.01
Expression	-0.18	0.1	0.36	-0.15	0.04	0.03
Value	-0.1	0.09	0.21	-0.28	0.02	-0.01
Workload	-0.02	-0.08	0.08	0.09	0.02	0.02
Mean Effect Size	-0.10	0.02	0.21	-0.13	0.04	0.01
Administrative Decision						
Instructor	-0.1	0.07	0.2	-0.2	0.03	0.00
Course	-0.07	0.12	0.17	-0.36	0.00	-0.03
Mean Effect Size	-0.09	0.10	0.19	-0.28	0.02	-0.01
Student Achievement						
My GPA	0.14	0.05	-0.18	0.08	-0.12	-0.01
Prob Grade	-0.1	0.1	0.19	0.17	-0.02	0.07
Class GPA	-0.13	0.08	0.24	0.26	-0.01	0.09
Mean Effect Size	-0.03	0.08	0.08	0.17	-0.05	0.05
Summary						
Mean Effect Size	-0.09	0.06	0.19	-0.11	0.02	0.01

Note. Significance levels were not computed for the mean differences in the table.

Table 5
Effect Size for CIS Basic Form Items and Mean Class GPA by Faculty Rank Compared to Full Professor for Fall 2003 and Spring 2004

Basic Form Items by Cluster	Associate Professor	Assistant Professor	Instructor	Lecturer	Mean Effect Size
Improving Instruction					
Organized	0.10	0.24*	0.12	0.19*	0.16
Communicated	0.20*	0.23*	0.09	0.17*	0.17
Interest	0.22*	0.30*	0.06	0.25*	0.21
Promptness	-0.07	-0.02	-0.03	0.02	-0.03
Expression	0.26*	0.34*	0.03	0.20*	0.21
Value	0.19*	0.20*	0.18	0.12*	0.17
Workload	-0.06	0.09	0.12	0.05	0.06
Mean Effect Size	0.12	0.20	0.08	0.14	0.14
Administrative Decision					
Overall Instructor	0.12	0.19*	0.10	0.12*	0.13
Overall Course	0.19	0.17*	0.28*	0.07	0.18
Mean Effect Size	0.06	0.18	0.19	0.10	0.16
Student Achievement					
Overall GPA	-0.09	-0.17*	-0.06	-0.26*	-0.15
Probable Course Grade	0.19*	0.18*	0.26*	0.07*	0.18
Mean Course Grade ¹	0.21*	0.24*	0.38*	0.12*	0.24
Mean Effect Size	0.10	0.08	0.19	-0.02	0.09
Summary					
Mean Effect Size All	0.12	0.17	0.13	0.09	0.13

Note. The full professor rank was the control group.

¹ Probable course grade: $r = 0.79$.

* Statistically significant different control group mean ($p = .05$).

Table 6
Regression Coefficients for Predicting Overall Instructor Rating

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	-.826	.042		-19.635	.000
	Communicated	.511	.011	.499	46.382	.000
	Interest	.227	.011	.190	21.353	.000
	Value	.204	.010	.178	20.614	.000
	Expression	.126	.010	.098	12.117	.000
	Organized	.072	.010	.061	7.439	.000
	My GPA	.043	.007	.026	5.863	.000
	Promptness	.012	.006	.011	2.005	.045
	Workload	-.002	.006	-.001	-.303	.762
	Prob Grade	-.012	.012	-.007	-1.063	.288
	Class GPA	.005	.008	.005	.680	.496

Table 7
Regression Coefficients for Predicting Overall Course Rating

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	-.806	.045		-17.810	.000
	Value	.674	.011	.605	63.210	.000
	Communicated	.177	.012	.178	14.969	.000
	Prob Grade	.093	.012	.059	7.512	.000
	Organized	.130	.010	.114	12.515	.000
	Workload	-.057	.006	-.047	-9.549	.000
	Interest	.044	.011	.038	3.827	.000
	Promptness	.027	.006	.024	4.164	.000
	Class GPA	.026	.009	.023	3.028	.002
	Expression	.020	.011	.016	1.818	.069
	My GPA	-.008	.008	-.005	-1.014	.310

For the final analysis using ANOVA, the construct “workload” was recoded using a three-level categorical scale by collapsing “Excessive” and “Heavy” into one level, placing “Average” as a second level, and collapsing “Light” and “Insufficient” as a third level of a categorical variable for groups, with all other CIS Basic Form items and the course grade point average being dependent variables. The ANOVA results are shown in Tables 8-10 and 12-19.

It was expected that “workload” would have nonlinear characteristics, and indications of

nonlinearity can be seen in the mean ratings for each workload category for each ANOVA. The differences in the means associated with the workload categories are small, but the pattern is observable.

Path analysis of variables

Path analyses using the method developed by Baron and Kenny (1986) were conducted for the propositions that (1) course value predicts overall course rating mediated by overall instructor rating and (2) instructor communication predicts overall instructor rating mediated by overall course rating.

As the results of the first path analysis, shown in Table 8, indicate, course value (X) was a statistically significant predictor for overall course rating (Y), as was overall instructor rating (M), the latter mediating the relationship between course value and overall course rating. These findings were confirmed by two other analyses: the bootstrapping method proposed by Preacher and Hayes (2004) and the Sobel test (Sobel 1982).

Table 8.
Direct and Total Effects of Course Value (X) as a Predictor for Overall Course Rating (Y) Mediated by Overall Instructor Rating (M).

Coefficient	s.e.		t	Sig (two)
b(YX)	.8742	.0067	130.2183	.0000
b(MX)	.9532	.0052	184.2437	.0000
b(YM.X)	.7630	.0147	51.7728	.0000
b(YX.M)	.1469	.0151	9.7535	.0000

Table 9 shows the results of the second path analysis, which was used to assess the possibility that the effects of instructor communication on overall instructor rating were mediated by the overall course rating. Again, the results of all three tests were consistent in indicating that overall course rating mediates the relationship between instructor communication and overall instructor rating.

Table 9.

Direct and Total Effects of Communication (X) as a Predictor for Overall Instructor Rating (Y) Mediated by Overall Course Rating (M).

Coefficient	s.e.		t	Sig (two)
b(YX)	.9532	.0052	184.2437	.0000
b(MX)	.8742	.0067	130.2183	.0000
b(YM.X)	.4531	.0088	51.7728	.0000
b(YX.M)	.5570	.0087	63.8742	.0000

According to Baron and Kenny (1986), the mediation effects observed in both cases are only partial, because there is significant correlation between the independent and dependent variables after controlling for the effects of the mediator variables. In the case of overall course rating, the mediator variable—overall instructor—had a greater effect upon overall course rating than did the independent variable, course value (Table 8). On the other hand, the predictor variable—communication—had a greater effect upon the overall instructor rating than did the mediator variable, overall course (Table 9).

The relationship between workload and overall GPA

The descriptive statistics for student-reported overall GPA, probable course grade, and course grade point average are shown in Tables 10, 11, and 12, respectively. While the means varied for the student-reported overall GPA (Table 10), the confidence intervals overlap, making the differences less significant. From inspection of the confidence intervals for probable course grade (Table 11) and course grade point (Table 12), it can be seen that courses in which students rated workload insufficient had the highest mean grade point averages. Conversely, courses in which students rated the workload excessive had the lowest mean grade point averages. It can also be seen in Tables 14-21 that, in courses with the highest mean grade point averages and a workload rated insufficient, the remaining Basic Form items tended to elicit more positive

ratings than they elicited in courses in the two other workload categories.

In order to determine how students rated the course workload across instructor ranks, a cross tabulation with percentages was prepared (Table 13). Within workload categories, lecturers received the highest percentage ratings for all categories, and professors received the second highest percentages in the same categories. Within each instructor rank, the majority workload rating was average. Except for lecturers, the trend in the workload rating by rank was average to excessive.

Table 10
Mean of Student-Reported Overall GPA by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	151	3.200	.3699	.0301	3.141	3.259	2.0	4.0
Average ²	3374	3.146	.3572	.0062	3.134	3.158	1.4	4.0
Excessive ³	1415	3.195	.3743	.0100	3.175	3.214	1.8	4.0
Total	4940	3.161	.3633	.0052	3.151	3.171	1.4	4.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data. $F < .001$.

¹ Insufficient = 1.00–2.49

² Average = 2.5–3.49

³ Excessive = 3.5–5.00

Table 11
Mean of Student-Reported Probable Course Grade by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	151	3.771	.2297	.0187	3.734	3.808	3.0	4.0
Average ²	3371	3.402	.3369	.0058	3.391	3.414	2.3	4.0
Excessive ³	1413	3.285	.3881	.0103	3.265	3.305	1.7	4.0
Total	4935	3.380	.3603	.0051	3.370	3.390	1.7	4.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data. $F < .001$.

¹ Insufficient = 1.00–2.49

² Average = 2.5–3.49

³ Excessive = 3.5–5.00

Table 12

Mean of Course Grades by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	157	3.681	.3421	.0273	3.6272	3.735	2.0	4.0
Average ²	3480	3.212	.4760	.0081	3.1957	3.227	1.7	4.0
Excessive ³	1440	3.133	.5247	.0138	3.1057	3.1600	.8	4.0
Total	5077	3.204	.4954	.0070	3.1901	3.2174	.8	4.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data. $F < .001$.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 13
Student Rating of Course Workload by Faculty Rank, Fall 2003 and Spring 2004

		Instructor Rank					Total	
			Professor	Associate Prof	Assistant Prof	Instructor	Lecturer	
Workload Group	Insufficient	Count	25	25	18	12	77	157
		% within Workload Group	15.9%	15.9%	11.5%	7.6%	49.0%	100.0%
		% within Instructor Rank	1.7%	3.2%	2.3%	6.4%	4.2%	3.1%
		% of Total	.5%	.5%	.4%	.2%	1.5%	3.1%
	Average	Count	1079	560	538	118	1185	3480
		% within Workload Group	31.0%	16.1%	15.5%	3.4%	34.1%	100.0%
		% within Instructor Rank	72.6%	71.9%	69.2%	63.1%	64.2%	68.5%
		% of Total	21.3%	11.0%	10.6%	2.3%	23.3%	68.5%
	Excessive	Count	383	194	222	57	584	1440
		% within Workload Group	26.6%	13.5%	15.4%	4.0%	40.6%	100.0%
		% within Instructor Rank	25.8%	24.9%	28.5%	30.5%	31.6%	28.4%
		% of Total	7.5%	3.8%	4.4%	1.1%	11.5%	28.4%
Total		Count	1487	779	778	187	1846	5077
		% within Workload Group	29.3%	15.3%	15.3%	3.7%	36.4%	100.0%
		% within Instructor Rank	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	29.3%	15.3%	15.3%	3.7%	36.4%	100.0%

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Counts are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 14
Average Rating of Course Organization by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	155	4.282	.4958	.0398	4.203	4.361	2.3	5.0
Average ²	3480	4.244	.4703	.0080	4.228	4.259	1.5	5.0
Excessive ³	1440	4.171	.5711	.0150	4.141	4.200	1.0	5.0
Total	5075	4.224	.5028	.0071	4.210	4.238	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 15
Average Rating of Instructor Communication by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	156	4.392	.4625	.0370	4.319	4.465	2.5	5.0
Average ²	3480	4.196	.5512	.0093	4.178	4.214	1.4	5.0
Excessive ³	1440	4.085	.6329	.0167	4.052	4.117	1.0	5.0
Total	5076	4.171	.5766	.0081	4.155	4.186	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 16
Average Rating of Instructor Interest by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	156	4.468	.3860	.0309	4.407	4.529	2.8	5.0
Average ²	3480	4.280	.4782	.0081	4.265	4.296	1.9	5.0
Excessive ³	1440	4.237	.5320	.0140	4.209	4.264	1.0	5.0
Total	5076	4.274	.4931	.0069	4.260	4.287	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 17

Average Rating of Promptness of Returning Assignments by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	149	4.317	.5084	.0417	4.235	4.400	1.9	5.0
Average ²	3375	4.276	.4856	.0084	4.259	4.292	1.2	5.0
Excessive ³	1411	4.164	.5750	.0153	4.134	4.194	1.3	5.0
Total	4935	4.245	.5160	.0073	4.231	4.259	1.2	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 18

Average Rating of Student Ability to Express by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	156	4.533	.3391	.0272	4.480	4.587	3.3	5.0
Average ²	3480	4.397	.4268	.0072	4.383	4.412	2.3	5.0
Excessive ³	1440	4.266	.5222	.0138	4.239	4.293	1.0	5.0
Total	5076	4.364	.4584	.0064	4.352	4.377	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 19

Average Rating of Course Value to Student by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	157	4.392	.4658	.0372	4.319	4.466	2.8	5.0
Average ²	3480	4.245	.4930	.0084	4.229	4.261	1.7	5.0
Excessive ³	1440	4.214	.5623	.0148	4.185	4.244	1.0	5.0
Total	5077	4.241	.5136	.0072	4.227	4.255	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 20
Average Overall Instructor Rating by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	157	4.371	.5079	.0405	4.291	4.451	2.6	5.0
Average ²	3480	4.167	.5631	.0095	4.149	4.186	1.4	5.0
Excessive ³	1440	4.075	.6469	.0170	4.041	4.108	1.0	5.0
Total	5077	4.147	.5893	.0083	4.131	4.164	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Table 21
Average Rating of Overall Course by Course Workload, Fall 2003 and Spring 2004

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Insufficient ¹	156	4.185	.5530	.0443	4.098	4.273	2.5	5.0
Average ²	3480	3.942	.5458	.0093	3.924	3.960	1.7	5.0
Excessive ³	1440	3.838	.6290	.0166	3.806	3.871	1.0	5.0
Total	5076	3.920	.5746	.0081	3.904	3.936	1.0	5.0

Note. The rating scale for the course workload item was 1 = insufficient to 3 = average to 5 = excessive. Means are for course level data.

¹ Insufficient = 1.00 – 2.49

² Average = 2.5 – 3.49

³ Excessive = 3.5 – 5.00

Discussion

The focus of this exploratory study was the CIS Basic Form used at a major Research I university in the southwest, an instrument for instructor and course evaluation very similar to those used at a majority of U.S. universities and colleges. Examination of the structure of the Basic Form, of item data from it, and of associated data concerning student course achievement has produced reassuring findings and has been the catalyst for an analytical model.

The structure of the Basic Form and its selection of items seem to fit well within the typical purposes and dimensions of student ratings of instruction as described in the literature. The six dimensions within the purpose of “improving teaching” are each addressed by at least one item. Items that address the “administrative decision” purpose are commonly used by students in making decisions regarding course selection. These findings suggest that the Basic Form is well designed for its purpose: to provide reliable information for faculty, administrators, and students.

The design of the Basic Form is further supported by the results of the analysis of rating data collected from its items. The fundamental relationships of the items between and within clusters, as expressed by the correlation coefficients, tend to follow patterns reported in the literature. So while the relationship between course grades and the ratings for items in the “improving instructor” and “administrative decision” clusters seems lower than those in some reports in the literature, it is consistent with the general pattern of weak relationships. Even though, therefore, it would seem that teaching effectiveness should result in student achievement, course grades continue to be a relatively ineffective predictor of effectiveness, perhaps because instructors assign grades in many ways, using many variables. Course grade may just be too variable a measure to be useful.

Furthermore, the conundrum is that a strong, positive correlation between students' ratings and course outcomes is widely taken to be evidence that student ratings have been manipulated by grading leniency—instead of elicited by good teaching and high student achievement—placing the validity of all student ratings in jeopardy. Yet, taken as a whole, the results of the effect size calculations and the “workload” ANOVA results do not support a grading leniency hypothesis.

Rather, the most revealing finding is the difference in effect sizes for student ratings between the ranks of professor and assistant professor. The professor, widely held to be the gold standard of teaching effectiveness, had the lowest student rating effect sizes across all items, and the assistant professor consistently had the most positive effect sizes of all instructor ranks. The implication to be drawn from these data is that assistant professors are doing a better job of teaching than are faculty members of any other rank.

Although student achievement is assumed to be reflected in positive student ratings, student achievement is not used in administrative decisions. Instead, student ratings on “administrative decision” items are used in making decisions concerning promotion, tenure, post-tenure review, merit, and awards. For that reason, it seemed prudent to investigate the predictability of “administrative decision” items by the “improving teaching” items and “student achievement” items. It turns out that two or three items from the “improving teaching” cluster were the best predictors of ratings of overall instructor and overall course, and that course grades were the least influential predictors, providing a strong indication that improvement in instructor teaching abilities has more affect on summative ratings than the assignment of course grades.

The relationship of ratings of course workload to student achievement is important. The expected pattern would be that high-ability students would rate a low course workload as

insufficient, low ability students would rate a high course workload as excessive, and students whose abilities most closely matched to the course workload would rate it average. To the degree that student ratings of course workload followed that pattern, allayed would be the notion that grading leniency was a factor in student ratings. Granting that differences evident in the results are very small, the pattern from the ANOVA is consistent with the expected results. Students with the highest course grade point average rated the course workload as insufficient, and students with the lowest course grade point average rated the course workload as excessive. The greatest number of students rated the course workload as average. These results are consistent with the finding of a slightly negative correlation between course workload and the other Basic Form items. It was a little surprising that students with the highest mean course grades responded to the remaining Basic Form items more positively. Nevertheless, the implication is that students were making an honest evaluation of the instructional process.

The consistency in the results from the separate analyses, with no serious anomalies, provides coherence to the results, suggesting that the course-instructor survey program is well designed and is providing valid results unaffected by biasing factors. The indication is strong that improvement in instructor teaching abilities—specifically, in communication—has more affect on the summative ratings than does the assignment of course grades. In addition, the results strongly suggest that assistant professors as a group are doing the best job of teaching in the classroom.

Conclusion

This study was conceived as exploratory, to develop a model for analyzing the results of the course-instructor survey program at the university. Beyond that, the purpose was to use its analytic techniques to find out how the results related to the purposes of the survey, to course

grades, and to grading leniency bias. We believe that these purposes have been accomplished, and a model has been established to serve as a foundation for a confirmatory study to replicate the results.

We believe that the results of this study support five conclusions. First, the CIS Basic Form is well designed to accomplish the three purposes of students' ratings of instruction, with items that are sufficiently representative of those purposes. Second, student achievement data, such as probable course grade and mean course grade point average, are weakly related to ratings of teaching skills and to summative instructor and course ratings. Third, teaching skills—particularly communication skills—are strongly and positively related to summative instructor and course ratings and are the most influential factors in determining what those ratings are. Fourth, assistant professors as a group are demonstrating the best teaching skills of all instructor ranks. Fifth, the preponderance of evidence, including course workload evidence, supports the conclusion that students are providing ratings of instruction that are unbiased by grading leniency.

In sum, as an evaluation of teaching, student ratings of instruction are accurate and trustworthy.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, *51*, 1173-1182.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, *51* (3), 281-309.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? Journal of Educational Psychology, *92* (1), 202-228.

McKeachie, W. J. (1979, October). Student ratings of faculty: A reprise. Academe, 384-397.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, Instruments, and Computers, *36* (4), 717-731.

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhardt (Ed.), Sociological methodology 1982 (pp. 290-312). San Francisco: Jossey-Bass.