American Educational Research Journal
Fall 1975, Vol. 12, No. 4, Pp. 435-447

Student Ratings of Instruction: Validation Research[1]

PETER W. FREY
Northwestern University

DALE W. LEONARD
Purdue University

WILLIAM W. BEATTY
North Dakota State University

Correlations between instructional ratings and exam performance at three universities indicated that three rating factors (student accomplishment, presentation clarity, and organization-planning) correlated highly with educational achievement. These rating factors were derived by a factor analysis of a 21-item questionnaire. Separate analyses of the relationship between instructional ratings and student characteristics indicated that the student's grade point average and math aptitude score (SAT) did not systematically vary with his ratings. However, more senior students (number of terms in college) rated instructors more favorably than their less experienced classmates.

[1] This research was supported by the Center for the Teaching Professions, Northwestern University, with funds from the Kellog Foundation and by Endeavor Information Systems, Inc.

With widespread adoption of teaching evaluation systems based on student ratings, many concerned educators have raised questions concerning the relationship between achievement of educational objectives and student ratings of instruction. Although there is quite a long history of research on student ratings of instruction (cf., Costin, Greenough, & Menges, 1971; McKeachie, 1969; and Menges, 1973, for reviews), the validity of student instructional ratings is still a subject which fosters lively debate. A widely publicized report by Rodin and Rodin (1972) provided data which purported to show that student ratings correlated negatively with an objective criterion of "good teaching." These results have been contested on methodological grounds (cf., Rodin, Frey, & Gessner, 1975). Subsequent replications of the basic design using regular faculty rather than teaching assistants and using more sophisticated rating forms have reported a positive relationship rather than a negative relationship (Frey, 1973; Gessner, 1973; Sullivan & Skanes, 1974).

The present research studies were conducted in order to examine the reliability of previous findings. In each study, an introductory course with a large enrollment was subdivided into multiple sections and each section was taught by a regular faculty member. Each of the sections had a common course syllabus, a common textbook, and a common final exam. Ratings of the course and instructor which were collected in class during the last two weeks of the term were compared to the average final exam performance in each section. The correlation between these two measures were examined to determine the validity of the student ratings.

METHOD

Instructional ratings were obtained from students enrolled in multi-section courses at Northwestern University, Purdue University, and North Dakota State University. At Northwestern, the course consisted of 16 sections of introductory calculus. The students were primarily freshmen. At Purdue, the course consisted of 10 sections of educational psychology. The students were sophmores, juniors, and seniors. At North Dakota State, the course consisted of 5 sections of introductory calculus and the students were primarily freshmen and sophomores. At all three universities, the students chose their sections in the normal fashion during course registration. Standard math ability scores (SAT) were available for the Northwestern students and were analyzed to determine if non-random section assignment produced ability differences among sections.

At all three schools, the instructor for each section had primary responsibility for the organization and conduct of classroom activities. All sections used a common course syllabus, common textbook, and a common final exam. The relative performance of each section (i.e., the class average) on the common final exam was used as the external criterion of student achievement. The operational definition of "good teaching," therefore, was the teacher's relative success at preparing his students for the common final exam.

Instructional ratings were obtained at each institution by distributing a questionnaire to the students in class during the last two weeks of the term. The Endeavor Instructional Rating Form[2] was used at all three institutions. This rating form has 21 items which are each marked on a seven-point scale. The response fields for 14 of the items are based on a frequency dimension (never, seldom, sometimes, often, always) while the response fields for the other 7 items are based on the degree of agreement (definitely no, no, yes, definitely yes). The form was designed to provide information about seven aspects of the teaching situation; three items elicit information about each aspect. These seven aspects are clarity of presentations, work load, personal attention, class discussion, organization-planning, grading, and student accomplishment.

RESULTS

The responses to all the rating forms at Northwestern (n = 421) and at Purdue (n=218) were factor analyzed to determine if a common factor structure would hold for responses from two different schools representing two different disciplines (i.e. calculus and educational psychology). The rating data from North Dakota were not factor analyzed since the sample size (n=139) was too small to obtain a reliable inter-item correlation matrix. The results of these factor analysis procedures are presented in Table 1 and Table 2. The factor loadings clearly indicate the presence of seven factors in both instances. The two factor matrices are also highly similar suggesting that the obtained structure is reliable.

Validity of the Ratings

The correlation between each section's final exam performance and the instructor's ratings can be determined in several different ways. The results from the Purdue and North Dakota State studies are presented in terms of the correlation between the

[2] Sample copies of this form are available on request from the first author.

TABLE 1

Factor Analysis of Endeavor Instructional Rating Form[a]

	Factor Matrix						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
I. Clarity of Presentations							
7. presentations clarified material	.72	−.04	.24	.22	.25	.10	.34
11. presented clearly and summarized	.74	−.01	.30	.09	.36	.14	.27
16. good use of examples	.78	.00	.19	.21	.19	.23	.22
II. Work Load							
1. had to work hard	−.03	.91	−.03	−.13	.02	−.05	−.08
10. required a lot of time	.00	.95	−.01	−.03	.07	−.06	−.04
20. heavy work load	−.01	.92	−.06	−.04	.09	−.08	−.04
III. Personal Attention							
3. listened and willing to help	.27	−.14	.69	.36	.18	.12	.16
9. able to get personal help	.19	.02	.72	.29	.14	.05	.24
21. concerned about student difficulties	.32	−.03	.66	.35	.16	.20	.13
IV. Class Discussion							
6. discussion was welcome	.09	−.11	.26	.81	.06	.14	.13
13. encouraged to participate	.10	−.05	.13	.88	.07	.11	.12
18. encouraged to express ideas	.22	−.06	.27	.78	.05	.11	.17
V. Organization-Planning							
2. planned in advance	.12	.14	.25	−.09	.81	.08	.14
14. detailed course schedule	.22	.07	−.04	.22	.76	.15	.20
17. activities were orderly scheduled	.22	.02	.17	.10	.80	.19	.09
VI. Grading							
5. fair and impartial grading	.03	−.18	.42	.08	.14	.64	.25
12. grading reflected performance	.17	−.05	.03	.15	.14	.88	.17
15. grading indicated accomplishments	.16	−.06	.06	.14	.17	.89	.14
VII. Student Accomplishment							
5. understand advanced material	.10	−.18	.14	.16	.21	.19	.79
8. ability to analyze issues	.28	.02	.14	.21	.08	.18	.78
19. increased knowledge and competence	.28	−.05	.19	.09	.17	.16	.79

[a] Principal factor solution with varimax rotation. The data represent sixteen calculus sections at Northwestern University (n = 421).

observed final exam average for each instructor's students and his instructional ratings. Math SAT scores were available for almost all of the Northwestern students so an adjusted (linear regression) final exam score for each section was used for the correlational analysis. This regression procedure provides an improved measure of how much the students learned in the course.

The Purdue results exclude one instructor who withheld his

TABLE 2

Factor Analysis of Endeavor Instructional Rating Form[a]

	Factor Matrix						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
I. Clarity of Presentations							
7. presentations clarified material	.70	−.06	.31	.30	.27	.14	.29
11. presented clearly and summarized	.72	−.02	.28	.21	.35	.16	.25
16. good use of examples	.77	.07	.17	.26	.24	.24	.19
II. Work Load							
1. had to work hard	−.02	.90	−.01	.05	.06	−.09	.05
10. required a lot of time	.04	.94	−.01	.06	.05	−.05	−.04
20. heavy work load	−.01	.94	.02	.03	.00	−.09	−.04
III. Personal Attention							
3. listened and willing to help	.20	−.02	.80	.22	.24	.18	.25
9. able to get personal help	.19	.01	.74	.34	.17	.27	.14
21. concerned about student difficulties	.37	.04	.65	.28	.15	.27	.17
IV. Class Discussion							
6. discussion was welcome	.22	.01	.25	.84	.13	.10	.09
13. encouraged to participate	.13	.10	.12	.87	.10	.18	.18
18. encouraged to express ideas	.23	.07	.31	.74	.25	.12	.17
V. Organization-Planning							
2. planned in advance	.31	.08	.19	.09	.72	.13	.27
14. detailed course schedule	.10	.07	.12	.19	.81	.26	.13
17. activities were orderly scheduled	.36	.05	.21	.19	.76	.15	.20
VI. Grading							
5. fair and impartial grading	.12	−.17	.30	.15	.20	.72	.16
12. grading reflected performance	.20	−.07	.16	.14	.17	.80	.30
15. grading indicated accomplishments	.15	−.11	.16	.14	.18	.82	.30
VII. Student Accomplishment							
4. understand advanced material	.09	−.09	.20	.10	.22	.28	.76
8. ability to analyze issues	.24	−.01	.17	.17	.19	.19	.80
19. increased knowledge and competence	.23	.06	.11	.19	.14	.26	.81

[a] Principal factor solution with varimax rotation. The data represent ten educational psychology sections at Purdue University (n = 218).

final exam scores; the correlational analysis is therefore based on the nine sections taught by the other six teachers. The data from North Dakota State involve three instructors teaching five sections of calculus. The Northwestern study initially involved 16 sections of introductory calculus. Four of these sections were eliminated from our analysis for several different reasons. The instructor for one section died during the week before the final exam. A second instructor insisted on returning the midterm

exam to his students while they were responding to our questionnaire. Two other sections employed special teaching procedures; one was taught as a very small section in a dormitory rather than in a regular classroom and the other emphasized computer applications rather than the regular syllabus material. The results for the remaining 12 sections (representing 11 instructors) were included in our analysis.

The results of this correlational analysis are summarized in Table 3.[3] Three factors showed a fairly strong correlation with final exam performance: student accomplishment ($r=.59$), presentation clarity ($r=.58$), and organization-planning ($r=.51$). The students' perception of the work load did not correlate well with final exam performance ($r=-.28$).

Instructional Ratings and Student Characteristics

At Purdue, the students also provided information on their grade point average (GPA) and the number of semesters they had been in college (seniority). At Northwestern, information was available, as was noted previously, on each student's math SAT score. A correlation coefficient was calculated within each section between each of these measures for each student and the ratings made by the student. A summary of the mean correlations across sections for each measure is presented in Table 4.[4] These data indicate that the ratings did not systematically covary with grade point average. The students with higher math SAT scores showed a tendency toward low ratings on the work load factor and high ratings on the student accomplishment factor. Student seniority (number of semesters in school) showed a strong positive relationship to all of the instructional rating factors except for work load. The experienced students were clearly more lenient in their ratings than their younger classmates.

DISCUSSION

The present results demonstrate that the instructional rating questionnaire employed in our studies has a replicable factor structure and that three of these factors (student accomplishment, presentation clarity, and organization-planning) correlate positively with an external criterion of "good teaching." Overall, the correlations presented in this report were weaker than those reported earlier in a study (Frey, 1973) employing a similar research design. The major difference between our present

[3] A detailed section-by-section table of the final exam scores and ratings on each factor for each study is available from the authors on request.

[4] A table of the correlations for each section are available from the authors on request.

TABLE 3

Correlations between Each of the Seven Instructional Rating Factors and Final Exam Performance

Institution	*Number of Sections*	*Presentation Clarity*	*Work Load*	*Personal Attention*	*Class Discussion*	*Organization-Planning*	*Grading*	*Student Accomplishment*
Purdue	9	.81	.12	.42	.40	.85	.28	.53
North Dakota State	5	.74	−.40	.67	.64	.70	.24	.64
Northwestern	12	.18	−.57	.06	.03	−.02	.38	.61
Mean		.58	−.28	.38	.36	.51	.30	.59

TABLE 4

Mean Within-Section Correlations between Instructional Ratings and Three Student Characteristics

	Presentation Clarity	*Work Load*	*Personal Attention*	*Class Discussion*	*Organization-Planning*	*Grading*	*Student Accomplishment*
Math SAT Score	.00	−.32	.00	.03	.01	.12	.27
Grade Point Average	−.05	−.01	.00	.09	−.13	.07	.00
Seniority	.61	−.05	.56	.39	.48	.50	.54

studies and the prior one was the time and manner of data collection. In the present studies, the data were collected in class prior to the final exam while in the prior study the data were collected by a mail survey two months after the final exam.

The results presented in this report combined with the prior study (Frey, 1973) represent five multi-section classes, three institutions, and two disciplines. The average correlations across these five samples provide a fair indication of the external validity of each of the seven rating factors. Given the reliable pattern of our results, it seems appropriate to conclude that at least three of the rating factors provide a valid index of which teachers are most successful at preparing students for their final exam.

Our analyses of the relationship between instructional ratings and student characteristics suggested that neither grade point average nor aptitude scores vary systematically with the ratings. In contrast, student seniority does seem to affect the ratings. With the exception of the work load factor, there is a reliable trend for the more senior students to give higher ratings to their instructors. This effect is consistent with the previous weak tendency reported by Rosenshine, Cohen, and Forst (1973) for more experienced students to give higher ratings. This apparent increase in "leniency" as the student gains more experience in school could be attributed to several factors. One hypothesis is that students lower their expectations as they encounter bad teachers during their first few years in college. This notion could be formalized in terms of adaptation level theory by postulating that prior experiences with bad teaching result in a readjustment of the adaptation level toward a lower value. A second plausible hypothesis is that the juniors and seniors in the course represent a different population than the freshman and sophomores. The upper classmen may be taking the course as an elective while the freshman and sophomores are taking it as part of their major. If this were the case, the more lenient ratings by the upperclassmen could be attributed to less ego-involvement. Alternatively, the upperclassmen may be better prepared for the academic demands of the course and rate it more highly because they do, in fact, get more out of the course than the underclassmen. A fourth possible explanation could relate to a mortality factor such that the positive attitude of the upperclassmen accounts for their continued enrollment at Purdue as well as their higher teacher ratings. Students with less positive attitudes would be less likely to become juniors or seniors.

The clear relationship between instructional ratings and seniority suggests that it is reasonable to develop separate norms for freshman and upperclass courses. The present data suggest that the ratings for an upperclass course should generally be higher than those for a freshman course. The absence of any

systematic relationship between instructional ratings and GPA suggests that the common procedure of providing separate rating summaries for "A" students and "C" students is probably not necessary.

REFERENCES

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, *41*, 511–535.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. *Science*, 1973, *182*, 83–85.

Gessner, P. K. Evaluation of instruction. *Science*, 1973, *180*, 566–569.

McKeachie, W. J. Student ratings of faculty. *American Association of University Professors Bulletin*, 1969, *55*, 439–444.

Menges, R. J. The new reporters: Students rate instruction. In C. R. Pace (Ed.), *New directions in higher education: Evaluating learning and teaching*. San Francisco: Jossey-Bass, 1973.

Rodin, M. & Rodin, B. Student evaluations of teachers. *Science*, 1972, *177*, 1164–1166.

Rodin, M., Frey, P. W., & Gessner, P. K. Student evaluation. *Science*, 1975, *187*, 555–559.

Rosenshine, B., Cohen, A., & Furst, N. Correlates of student preference ratings. *Journal of College Student Personnel*, 1973, *14*, 269–272.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, 1974, *66*, 584–590.

Some Remarks on "Student Ratings: Validation"

CRAIG S. SCOTT

Oregon State System of Higher Education

The number of studies dealing specifically with the validity of student ratings of instructor performance has increased dramatically since the middle 1960's. The increase is due in part to the growing realization that students are in the most advantageous position to observe actual in-class activities, and that their collective input should, therefore, be used as a legitimate source of information when comprehensively assessing the teaching function. Calls for increased use of student input, for a variety of