RB-76-6

RESEARCH BULLETIN

STUDENT RATINGS OF INSTRUCTION AND THEIR RELATIONSHIP TO STUDENT LEARNING

John A. Centra

with the assistance of

Bert Rose
Memorial University
Newfoundland

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the authors. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
February 1976

Abstract

Student ratings of instruction were correlated with examination performance in 72 sections of seven courses. In two of the courses, students had been randomly assigned to sections. The pattern of correlations across the courses indicated that the global ratings of teacher effectiveness and of the value of the course to students were most highly related to mean exam performance (12 out of 24 product-moment and partial correlations were .58 or above). Ratings of course objectives and organization, and of the quality of lectures were also fairly well correlated with achievement. Ratings of other aspects of instruction, such as the teacher-student relationship or the difficulty/workload of the course, were not highly related to achievement scores.

Student Ratings of Instruction and Their Relationship to Student Learning

Studies of what student ratings of instruction really measure have frequently employed student achievement as a validity criterion. The hope is that ratings of teaching or course quality would be at least moderately related to how much students learn in a course.

In spite of a number of articles on the topic, there are still some unresolved issues, one of which involves what the relationships found may actually mean. Previous researchers have usually employed a predictive or criterion-related model to conceptualize the validity of student ratings (American Psychological Association, 1966). Most of these studies included a multisection course with a common final examination. For each section, mean student ratings were correlated with mean final exam performance at the end of the course. Students had generally selected their sections or teachers rather than being assigned to them on a random basis, requiring some kind of statistical adjustment to compensate for initial differences in student ability or achievement. For example, in an early study, Elliot (1950) adjusted for academic aptitude and found moderate correlations between the adjusted achievement scores and ratings of some of the areas of instruction. Cohen and Berger (1970); Morsh, Burgess, and Smith (1956); and McKeachie, Lin, and Mann (1971) also reported some moderate correlations between ratings and learning, but again students had not been assigned to teachers on a random basis.

A negative adjusted correlation was reported by Rodin and Rodin (1972) who also had not been able to assign students randomly to teachers. Their study, however, had other features that might account for the negative results as several critics have pointed out (Doyle, 1975; Frey, 1973; Gessner, 1973).

The predictive model provides useful evidence of the correlational relationships between ratings and learning. Correlational evidence alone, however, is subject to varying interpretations. Although different teaching techniques or course characteristics as rated by students may influence students to learn more, other explanations are possible. It could be argued, for example, that differences in student achievement are due not to teacher effects but to differences in students uncontrolled by whatever pretest had been used. Differences in final exam scores for students at a given pretest performance level may, as one possibility, be due to differences in student motivation rather than in teacher effectiveness. Highly motivated students might not only do better than predicted but may seek out teachers with good reputations and rate them higher regardless of teaching performance.

Another possibility is that a class with high achieving students might inspire a teacher to put more into the course, so that student achievement would influence teacher behavior rather than the reverse.

If one assumes that different teaching practices as rated by students should affect learning, then a research design in which students had been assigned randomly to each section of a course is highly advantageous.

But because of the practical difficulties in assigning college students to sections on a random basis, only one study has thus far been reported that employed this approach. Sullivan and Skanes (1974) correlated instructor ratings and final exam scores in 130 sections for 10 first-year courses in which students at Memorial University (Newfoundland) had been randomly assigned. The average correlation was .39 between an overall rating of the instructor and achievement scores. Eight of the 10 courses had correlations higher than .32. Correlations were highest for full-time faculty and lowest for inexperienced part-time instructors.

Another issue that warrants further study is how student ratings of different dimensions of instruction are related to student achievement. Most of the previously cited studies reported correlations between global ratings of teaching effectiveness and student achievement. For those that also included an analysis of specific instructional dimensions (e.g., course organization or instructor preparedness), the results were generally inconsistent (e.g., Sullivan and Skanes, 1974).

The purpose of the present study was to investigate further the relation between achievement and student ratings. The Student Instructional Report (SIR), which elicits general ratings of the teacher or the course as well as ratings of specific instructional practices, was used in the study (Centra, 1972). Included for analysis were courses in which students had been randomly assigned to sections, and courses in which prior achievement in the subject matter was statistically adjusted.

Method

As with the Sullivan and Skanes (1974) study, courses at Memorial University in Newfoundland were included in the present study. The data were collected during the Winter 1975 term at the university. First year students are required to take several introductory courses, each of which is taught as a two semester sequence. Teachers in seven of these course areas agreed to participate in this study. In two of these courses, Chemistry 1000 and Biology 1010, students were enrolled in the first part of the sequence and had been assigned to teachers on a random basis. In the remaining five courses, students were in the second part of the sequence. During the first term they too had been assigned to sections on a random basis but many had changed teachers for the second term in order to accommodate their class schedules or because they preferred another teacher.

The number of sections varied between 7 and 22 for the seven courses. The total was 72 sections taught by 44 teachers. None of the teachers was a graduate teaching assistant, and almost all were experienced teachers.

A final examination in each course was constructed by an examination committee which included Junior Division faculty (those teaching the introductory courses) and Senior Division faculty members, who taught the later courses. The procedure is as follows:

> Early in each semester, the examination committee makes up and circulates a model of the final examination, but the final examination itself is not circulated and its specific

> content is not known to any except the committee members until the examination is written. The committee also sets out guidelines to be used on marking each answer. All answers are board-marked with a small group of faculty members marking one question on all papers. The marks on the final examination are then tabulated simply by adding the raw scores on each question (Sullivan & Skanes, p. 585).

The final examination score, therefore, should be comparable from one section of a course to another.

Item reliabilities and a factor analysis of the Student Instructional Report are reported elsewhere (Centra, 1973a; Linn, Centra, & Tucker, 1975). Nine items or groups of items based in part on the factor structure were included in the analysis. An important item was the overall rating of the teacher's effectiveness: "Compared to other instructors you have had (secondary school and college), how effective has the instructor been in this course?" (Responses on a 5-point scale varied from one of the most effective to one of the least effective.) Other global items included a rating of the "overall value of the course" to the student, and a rating of the "general quality of the lectures." Five items were included in a rating of the teacher-student relationship (e.g., the instructor's availability for consultation with students); four were included in a rating of course objectives and organization (e.g., agreement between announced objectives and what was taught); three items were included in the dimension termed course difficulty and workload, and two items were included in a rating of both reading assignments and examinations. Finally, a single item which had not correlated highly with other items on the form assessed student effort (I have been putting a good deal of effort into this course).

The mean ratings on SIR items or on groups of items for class sections were correlated with mean student examination scores. For the five courses in the second term of the sequence, first term final examination scores were partialed out of both the second term examination scores and the ratings by students, resulting in partial correlation coefficients between ratings and achievement with previous achievement held constant. Global ratings and examinations scores were also plotted for the one course with the largest number of sections and teachers.

Results

The correlations between student ratings of the nine SIR variables and final examination scores are presented in Table 1. Listed first are correlations for the two courses in which students were assigned randomly, Chemistry 1000 and Biology 1010. Also given are product-moment correlations and partial correlations for Psychology 1001, Biology 1011, Mathematics 1011, Physics 1051, and Chemistry 1001. While the correlations were sometimes substantial, they were not always statistically significant,

----

Insert Table 1 about here

----

due possibly to the small number of sections for several of the courses. For the two courses in which students had been randomly assigned to sections, the highest correlations with achievement were for the ratings of the value of the course to students, followed by the ratings of teacher effectiveness and ratings of lecture quality. Course difficulty and workload correlated lowest with final examination scores.

For the five courses, there are some large differences between the product-moment and partial correlations, but because of the small number of class sections for each course it is difficult to know what this means. The partial correlations, in particular, will be unstable because of the sample size and differences in variances. Only for Psychology 1001, which included the largest number of class sections (22) were the product-moment and the partial correlations very similar. Ratings and test scores in psychology were significantly correlated for six of the nine SIR variables. Ratings of course difficulty and workload, student effort, and reading assignments had the three lowest correlations.

Although conclusions for Table 1 must be drawn cautiously because of the small number of classes for most courses, the pattern of correlations indicates that the examinations scores were significantly related to several of the SIR variables. Ratings of overall teaching effectiveness and the value of the course to students, in spite of consisting of only a single item each (and hence a less reliable measure), were both fairly well correlated with achievement: 12 out of the 24 product-moment and partial correlations were .58 or above. Ratings of course objectives and organization, and of the quality of lectures, were also fairly well correlated with achievement: 14 out of the 24 correlations were .47 or above. Ratings of the teacher-student relationship, of the course examinations, and of student effort were not strongly correlated with achievement: the median correlation was about .30. The weakest or most inconsistent correlations with achievement were for ratings of reading assignments and for course difficulty and workload.

A scatterplot of mean student achievement and ratings of overall teacher effectiveness for the 22 psychology classes is presented in Figure 1 in order to study ratings of different sections for the same teacher. The 22 classes were taught by 9 teachers, with each teaching at least two sections. Of particular interest are the class sections that fall some distance from the regression line. Teachers A, B, and C each have one section off the regression line and at least one other section very near the line. For both teachers A and C, the final examination score for one of their sections is lower than would be predicted by ratings; for teacher B, the final examination score for one section is higher than predicted by ratings. Teacher X, however, is especially noteworthy. All three of the classes for teacher X had lower achievement scores than their ratings would lead one to expect. In other words, relative to other teachers, X's ratings would not be an accurate reflection of how much students had learned. The ratings would,

----

Insert Figure 1 about here

----

nevertheless, by a very good estimation of student learning for the other eight teachers. The average rating for each teacher--that is, the average of all sections taught by a teacher--correlated very highly with their examination scores: the rank correlation for all nine teachers was .67 (the rank correlation for the eight teachers was .82).

Although not presented, a scatterplot of achievement and ratings of the value of the course to students in the psychology course was similar to Figure 1. Teacher X's three sections received lower achievement scores than expected from student ratings of course value.

Why teacher X was rated relatively high cannot be determined from the information available. Perhaps he or she chose to emphasize material other than that covered by the examination; or perhaps the teacher could inveigle students into thinking they had learned, as Naftulin, Ware and Donnelly (1973) illustrated in one study.

Discussion

Because of the relatively small number of sections for several of the courses, the patterns of correlations across courses rather than ratings for any one course should be emphasized. This pattern indicates that the student ratings most highly related to achievement test scores were the global ratings of teachers, of lectures, and of courses, as well as the ratings of course objectives and organization. Included among the latter group of items were student ratings of the extent to which the instructor had accomplished course objectives. In a sense, then, this dimension probably reflected student estimates of what they had learned in the course (relative to the objectives), as well as the extent to which the course had been organized (e.g., the instructor was well-prepared for class). Ratings of reading assignments and of the level of course difficulty and workload tended to have the lowest correlations with achievement. Apparently students' views of how arduous a course was, or how excellent the text or reading assignments were had little to do with how much they learned. Of course, the text and readings were probably standard from section to section and thus might not be expected to relate to differences in learning between sections. Correlations between achievement and the other SIR variables--teacher-student relationship, course examinations, and student effort--fluctuated from course to course but were generally very modest (the median correlation was about .30). These results are somewhat similar to those reported recently by Frey, Leonard, and Beatty (1975), although their instrument did not include global ratings. Correlations for the global ratings of teacher effectiveness were slightly higher than those reported by Sullivan and Skanes (1974) at the same university.

These patterns of correlations across the different SIR rating areas were somewhat similar for most of the courses, regardless of whether or not students had been assigned to sections randomly. In particular, the correlations of the global ratings with achievement were generally higher than those of the more specific instructional practices for both the random and nonramdom groups. An issue raised earlier in this paper was that the nonrandom assignment of students to sections (the predictive model) left open the possibility of alternative explanations of the results. That, of course, is still a possibility; other factors may well have influenced the correlations. These other factors, however, were less likely to influence correlations for the random group.

Global student ratings of teacher effectiveness or course value may be more valid estimates of student academic achievement because they are not tied to a specific instructional style. Some instructional practices likely work well for some but not all teachers, as the results also suggest. Not all teachers, for example, have to develop close student relationships to facilitate learning in their course. For some it is part of their teaching style, and it may well contribute to their effectiveness as measured by ratings or student achievement. Other teachers may use other practices that account for their effectiveness.

Student ratings are increasingly being used to make tenure and promotion decisions about faculty members (Bejar, 1975). If one assumes that ratings should bear at least a moderate relationship with student learning before they are used in this way, then the global ratings are more defensible than the ratings of specific practices. The latter could be more useful in helping instructors improve what they do by implying specific changes (Centra, 1973b), but the results of this study indicate that they may not always reflect student academic achievement. Even the global ratings may not adequately reflect the relative level of student achievement for all teachers, as the scatterplot for the psychology course in this study illustrated. In that instance the student achievement scores for one of the nine teachers were lower than would be predicted by the ratings. It is possible that for another sample, one or more teachers might be "underrated" rather than "overrated." Although global ratings and achievement were, in general, highly correlated for most of the courses in this study, the exceptions would appear to underscore the need to supplement the ratings with additional criteria of teaching performance, such as colleague input or available evidence of student learning.

References

American Psychological Association. Standards for educational and psychological tests and manuals. Washington, D. C.: American Psychological Association, 1966.

Bejar, I. I. A survey of selected administrative practices supporting student evaluation of instruction programs. Research in Higher Education, 1975, 3, 77-86.

Centra, J. A. The student instructional report: Its development and uses. SIR Report #1. Princeton, N. J.: Educational Testing Service, 1972.

Centra, J. A. Student instructional report number 3: Item reliabilities, the factor structure, comparison with alumni ratings. Princeton, N. J.: Educational Testing Service, 1973(a).

Centra, J. A. The effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973(b), 65, 395-401.

Cohen, S. A., & Berger, W. G. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. In Proceedings of the 178th Annual convention of the American Psychological Association, 1970, 5, 605-606.

Doyle, K. O. Student evaluation of instruction. Lexington, Mass.: D. C. Heath & Co., 1975.

Elliot, D. H. Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education, 1950, 70, 5-61.

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

Frey, P. W., Leonard, D. W., & Beatty, W. W. Student ratings of instruction: Validation research. American Educational Research Journal, 1975, 12, 435-447.

Gessner, P. K. Evaluation of instruction. Science, 1973, 180, 566-570.

Linn, R. L., Centra, J. A., & Tucker, L. R. Between, within, and total group factor analyses of student ratings of instruction. Multivariate Behavioral Research, 1975, 10, 277-288.

McKeachie, W. J., Lin, Y. G., & Mann, W. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal, 1971, 8, 435-444.

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. Journal of Educational Psychology, 1956, 47, 79-88.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. The Doctor Fox lecture: A paradigm of educational seduction. Journal of Medical Education 1973, 48, 630-635.

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

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.

Table 1

Correlations Between Mean Student Instructional Ratings and Mean Final Examination Scores,

for 72 Sections and 7 Courses

(decimals omitted)

	Correlations with End of Term Examination Scores											
	Students assigned randomly		Students not assigned randomly					Partial Correlations [a]				
Rating Areas	Chem 1000 N=7	Biol 1010 N=7	Psych 1001 N=22	Biol 1011 N=13	Math 1011 N=8	Physics 1051 N=7	Chem 1001 N=8	Psych 1001 N=22	Biol 1011 N=13	Math 1011 N=8	Physics 1051 N=7	Chem 1001 N=8
Overall teaching effectiveness	60	61	63**	53*	-19	69*	30	64**	23	87**	58	41
Value of course to student	73*	92**	41*	63**	-18	61	23	47*	26	96**	-32	48
Teacher-students relationship [b]	58	30	49**	39	-24	17	06	43*	-08	87**	09	18
Course objectives and organization [b]	53	45	49**	51*	-23	64	27	49**	18	81**	62	31
Reading assignments [b]	35	55	35	61*	04	-59	04	24	34	17	23	-06
Course difficulty and workload [b]	-15	11	-50	-13	-48	17	30	-46	-30	60	37	-42
Examinations [b]	31	81*	45*	56*	-55	-10	02	46*	46	19	-63	21
Lectures	76*	47	47*	55*	-16	11	53	46*	18	90**	-04	43
Student effort	79*	09	17	00	-05	-10	48	30	-27	76*	07	51

[a] Examination scores from the first term partialled out of both ratings and second term examination scores.

[b] Based on two or more items.

*.05 level, one-tailed test.

**.01 level, one-tailed test.