DOCUMENT RESUME

ED 093 938 TM 003 756

AUTHOR Shapiro, Steven L.; Stein, Barry A.
TITLE Student Evaluations Related to Frequency of Testing.
PUB DATE [72]
NOTE 17p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; College Students; College Teachers; *Teacher Evaluation; *Testing
IDENTIFIERS *Frequency of Testing

ABSTRACT

An investigation was designed to determine the effects of frequency of testing on student performance. As a result of this experiment a relationship between student evaluations of teachers was noted. The results indicated increases in teacher ratings as test frequency increased. It is assumed that the findings were produced by the independent variable, frequency of testing, not by differences in subject matter, teachers, or methodology. (BB)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 093938

STUDENT EVALUATIONS RELATED TO FREQUENCY OF TESTING

By Steven L. Shapiro
Barry A. Stein

An interesting finding concerning student evaluations of teachers was noted recently as a result of an experiment designed to determine the effects of differences of frequency of testing on the performance of students enrolled under two different admissions policies in an urban community college. The two policies are open admissions, which admits all high school graduates regardless of average, and selective admissions, which in N.Y.C. prior to Fall 1970, required most high school graduates to have a 75 per cent average or higher to gain admission to a community college.

DESIGN

The experimental sample was selected from those students registering for Business Organization and Management for the Fall 1972 semester at

---

Dr. Steven L. Shapiro and Dr. Barry A. Stein are assistant professors in the department of business, Queensborough Community College, Bayside, N.Y. 11364. The authors thank Professor Sheldon Somerstein, chairman of the business department for his outstanding cooperation throughout the experiment.

TM 003 750

Queensborough Community College.[1] The sample was distributed into twelve classes, divided into three treatment groups - four receiving 10 tests, four 5 tests and four 3 tests during the term. Room and time assignments were made at random. Each group was given the same 150 multiple-choice questions during the semester to measure learning. Each class took the same 100 item multiple-choice final examination.

There were four instructors teaching the twelve classes in the investigation; each teaching a 10 test, 5 test and 3 test class (see Table 1). The instructors met weekly with the experiment leader to discuss the topics to be covered and methodology to be used.

Although all three treatment groups were composed of students from two different high school academic levels (below 75% and 75% or above), the groups were proven to be comparable by a two way analysis of variance on the variables high school average and reading and English expression scores.

---

[1]Queensborough Community College is a branch of the City University of New York.

RESULTS AND CONCLUSIONS

Through analysis of the data, it was noted that the students taking more than three test (experimental groups) had significantly higher (.01) final examination scores and final course grades than those being given three tests during the semester (control group). To be precise, open admissions freshmen did best when tested ten times during the term while regular freshmen achieved most when tested five times throughout the semester (see Table 2).

It was further observed that the students in the experimental groups rated their instructors higher in all categories of the student evaluation form used throughout the college. Although the nature of the evaluation instrument precludes identification of individual students, the overall findings found in Table 3 indicate surface validity and again raise a question that has been pondered for many years: What really is the relationship between students' achievement and teacher ratings?

In this experiment, the objective test results which show increased learning by students taking five or ten tests as opposed to three, agree with

the subjective student evaluations of teachers. Although differences in subject matter, teachers and methodology affected the evaluations, all three groups were affected due to selective manipulation in setting up the investigation. It is assumed that the findings were produced by the independent variable, frequency of testing, and not by differences in subject matter, teachers or methodology.

The findings of this study support D.N. Elliot who in a study of a large introductory chemistry course, concluded that ". . . there is probably, in general, a positive relationship between the ratings given an instructor by his students and their achievement. . ."[2] They also are compatable with H.H. Remmers who, in essentially the same experimental design, concluded that ". . . there is warrant for ascribing validity to student ratings . . . as measured by what students actually learn of the content of the course."[3]

Some investigators have found a negative correlation between the

---

[2]D.N. Elliot, Purdue University Student Higher Education, 70, 5 (1950).

[3]H.H. Remmers, F.D. Martin, D.N. Elliot, Purdue University Student Higher Education, 66, 17 (1949).

amount learned from an instructor and the students' evluation of his teaching performance. Rodin and Rodin in a study of 293 students in an undergraduate calculus course, concluded that ". . . the instructors with the three lowest subjective scores received the three highest objective scores while the instructor with the highest subject rating was lowest on the objective measure."[4]

R.H. Knapp found evidence that student evaluations, to a large extent, tend to reflect the personal and social qualities of an instructor, "who he is" rather than "what he does."[5] The results of this investigation indicate that "what he does" and not "who he is" determines to some degree the results of the student evaluation. Testing frequency seems to have been measured rather than individual teaching abilities. Each of the four participating instructors received their best evaluations as a result of increasing exam frequency. Collectively, they did not receive their highest rating in any of the ten categories from the 3 test group. The results indicate that students

---

[4] M. Rodin and B. Rodin, Science 177, 4055 (1972).

[5] R.H. Knapp, The American College, N. Sanford, Ed. (Wiley, New York, 1962), pp. 290-311.

taking 10 tests rate the instructors highest in categories 3, 4, 6, 7, 8, 9, and 10. Students in the 5 test group evaluate the instructors best in categories 1, 2 and 5. While it must be noted that the teachers themselves may have placed more emphasis on the experimental classes due to the nature of the study, these findings do show a pattern which indicates the importance of course modification in affecting student evaluations of teachers.

VARIATIONS IN RATINGS

The student evaluation of faculty is used today for purposes of re-hiring, promotions and tenure. It can be an important determinant in the relative success or ultimate failure of a teacher's career. In examining the individual categories more closely, it appears that students are measuring their image of achievement rather than teacher performance. By scanning column A we see that although all four teachers were required to cover the same topics during the term and all students received the same 150 multiple-choice items, the differences in teacher ratings are quite evident. Overall mean final examination scores differ significantly by approximately three points between the 10

test and 3 test group (see Table 2) while student evaluations differ by as much as 36.9 per cent (category 7) between the same two groups.

Interestingly, the greatest variation occurs in the category: Evaluates students' work. Students in the 10 test group who were constantly evaluated, indicated this on the rating form. While it was true, the students took the same number of test items during the semester as those evaluated less frequently and the mean total items correct during the term was 100.89 for the 10 test group and 101.4 for the 3 test group.

The category: How would you describe instructor to others?, shows the second greatest variation of ratings (35.9 per cent). This category is a particularly important one. It shows that when the four teachers in the experiment gave three tests during the semester, 24 per cent of the students rated them excellent. When five tests were used, 53 per cent responded excellent and when 10 tests were given, 61 per cent described the instructors as excellent. Accepting possible variations in teacher motivation toward the various groups, the percentages still strongly favor the instructors when they used increased

test frequency.

Although the category: Rate your own performance; shows a relatively small variation of ratings (10.2 per cent), it is important to recognize that students have a better self-concept when undergoing higher frequency testing. It is even more evident when columns A and B are combined and the variation increases to 25.7 per cent.

The student evaluations of instructors were conducted prior to the final examination. At that time, students were not aware of final exam grades or final course grades. The only evaluations of students were in the form of exam grades. The mean total items correct from these exams for the three groups had a variation of .509 (the difference in mean total items correct between the 10 test group and 3 test group) which is remarkably small when considering the number of students involved and the number of test items administered. The 258 students responding (81.6% of the 316 finishing the semester in the twelve classes) should have rated the instructors practically the same in each frequency group since achievement had been virtually the same up to that point.

Since final examination mean scores show significant differences between the groups, higher ratings as exam frequency increases indicate a relationship between student learning (achievement) and teacher ratings. This positive relationship, however, goes further than merely stating a possible correlation. The real question becomes: Are teacher ratings subject to actual student achievement which may be created by one or more course variables?

"THE GREATER LEARNING IMAGE"

In this study, the conclusions of Elliot and Remmers are substantiated while those of Rodin and Rodin and Knapp are not. There does seem to be a relationship between achievement and ratings but certainly not a simple one. The ratings seem to have been made, to a great extent, according to the students perception of learning throughout the semester. This "greater learning image" may be the result of several factors. Perhaps increased test frequency as opposed to three major examinations reduced test anxiety and made no one test critical. Another reason may have been the personal contact between teacher and student which developed as a result of constant item discussion

and grade distribution throughout the term. A third possiblity could have been that as a routine of testing was established, students liked having fewer topics on each exam, and found studying to be easier. All or any of these factors may have had a much greater effect on the student evalautions than the actual increased achievement which is evidenced by the final examination mean scores.

SUMMARY

The variable testing frequency seems to have a great influence on teacher evaluations. Results show tremendous increases in ratings as test frequency increases. "Who the teacher is" as opposed to "what he does" seems to be unimportant. What is important is how student achievement and evaluation of faculty are affected by frequency of testing. There is a significant relationship between test frequency and student achievement on the final examination. From this standpoint, the appraisal instrument used to evaluate the teachers is valid to some extent. There appears to be a striking relationship between the "greater learning image" created in this investigation

by increased test frequency and student evaluation of instructors.

With the student evaluation becoming a more and more important part of the success or failure of the college teacher, there is no doubt that a great many questions concerning these evaluations are still only vaguely answered. The researcher must no longer be concerned with only teacher effectiveness but also concentrate on the ingredients that contribute to the overall effectiveness of instruction.

The instructor, on the other hand, anxious for a high rating in category 9 as well as all the others, must begin to seek out various means of stimulating the ratings. In the sense that this may lead to experimentation and educational advances, this is fine. If, however, this pursuit leads to the use of teaching gimmicks, designed only to improve image and not instruction, the teacher evaluation idea fails. Those using the evaluations must read them carefully and always remember that while numbers don't lie, they do sometimes exaggerate!

Table 1

NUMBER OF CLASSES AND TYPES OF TESTS GIVEN BY EACH TEACHER

	10 Tests	5 Tests	3 Tests	Total Number of Business Org. & Mgmt. Classes Taught
Teacher A	1	1	1	3
Teacher B	1	1	1	3
Teacher C	1	1	1	3
Teacher D	1	1	1	3
Total classes	4	4	4	12

TABLE 2

FINAL EXAMINATION AND FINAL COURSE GRADE RESULTS
COMPARED TO TOTAL ITEMS CORRECT

High School Average	Frequency of Exam	Number of Observations	Final Examination		Final Course Grade		Mean Total Items Correct During Semester
			Mean	S.D.	Mean	S.D.	
Under 75%	10 Tests	57	72.895	10.304	2.309	0.726	97.5813
Under 75%	5 Tests	59	69.678	10.505	1.988	0.829	93.9045
Under 75%	3 Tests	49	67.939	10.792	1.691	0.842	94.5815
75% or above	10 Tests	51	75.549	11.481	2.655	0.967	104.5797
75% or above	5 Tests	44	78.318	8.515	2.799	0.699	111.4153
75% or above	3 Tests	56	74.318	10.697	2.534	0.891	107.3570
Over & under 75% combined	10 Tests	108	74.1483	. .	2.4724	. .	100.8861
Over & under 75% combined	5 Tests	103	73.3591	. .	2.3345	. .	101.3848
Over & under 75% combined	3 Tests	105	71.3140	. .	2.1406	. .	101.3951

TABLE 3

COMPARISON OF RESULTS OF STUDENT EVALUATION OF INSTRUCTORS

Number of Responses		Percentage of Response	
3 Test Group	89	3 Test Group	84.76%
5 Test Group	85	5 Test Group	82.76%
10 Test Group	84	10 Test Group	77.78%

Category	A	B	C	D	E
1. Communicates subject to students	Very clearly	Fairly clearly	Somewhat confusing	Very confusing	Dn/da[a]
3 Test Group	36/40.5%[b]	44/49.4%	7/7.9%	2/2.2%	0/0%
5 Test Group	64/75.2%	19/22.4%	2/2.4%	0/0%	0/0%
10 Test Group	59/70.2%	25/29.8%	0/0%	0/0%	0/0%
2. Demonstrates interest and enthusiasm.	Great extent	Somewhat	Rarely	Not at all	Dn/da
3 Test Group	37/41.5%	38/42.7%	11/12.4%	3/3.4%	0/0%
5 Test Group	66/75.2%	16/18.8%	2/2.4%	0/0%	0/0%
10 Test Group	63/75%	21/25%	0/0%	0/0%	0/0%
3. Knows subject.	Very well.	Fairly well	Poorly	Not at all	Dn/da
3 Test Group	62/69.7%	23/25.8%	2/2.25%	2/2.25%	0/0%
5 Test Group	65/76.4%	19/22.4%	0/0%	0/0%	1/1.2%
10 Test Group	71/84.5%	13/15.5%	0/0%	0/0%	0/0%
4. Uses class time productively.	Always	Frequently	Sometimes	Rarely	Dn/da
3 Test Group	40/44.9%	30/33.8%	14/15.7%	3/3.4%	2/2.2%
5 Test Group	54/63.5%	22/25.9%	3/3.5%	0/0%	6/7.1%
10 Test Group	56/66.67%	20/23.8%	2/2.4%	2/2.4%	4/4.8%
5. Helps students understand assignments and materials.	Always	Frequently	Sometimes	Rarely	Dn/da
3 Test Group	35/39.3%	27/30.3%	15/16.9%	12/13.5%	0/0%
5 Test Group	55/64.7%	22/25.9%	5/5.9%	1/1.2%	2/2.3%
10 Test Group	51/60.7%	25/29.8%	7/8.3%	1/1.2%	0/0%

TABLE 3 - Continued

COMPARISON OF RESULTS OF STUDENT EVALUATION OF INSTRUCTORS

CATEGORY	A	B	C	D	E
6. Encourages questions.	Always	Sometimes	Rarely	Discourages participation	Dn/da
3 Test Group	42/47.2%	32/36%	10/11.2%	4/4.5%	1/1.1%
5 Test Group	50/58.8%	29/34.2%	3/3.5%	0/0%	3/3.5%
10 Test Group	54/64.3%	27/32.1%	2/2.4%	0/0%	1/1.2%
7. Evaluates students' work.	Always	Sometimes	Rarely	Never	Dn/da
3 Test Group	35/39.3%	30/33.7%	12/13.5%	5/5.6%	7/7.9%
5 Test Group	53/62.4%	23/27.1%	3/3.5%	1/1.1%	5/5.9%
10 Test Group	64/76.2%	13/15.4%	3/3.6%	0/0%	4/4.8%
8. Encourages class participation.	Actively	Sometimes	Rarely	Discourages participation	Dn/da
3 Test Group	46/51.7%	30/33.7%	11/12.4%	2/2.2%	0/0%
5 Test Group	63/74.1%	16/18.8%	3/3.5%	1/1.2%	2/2.4%
10 Test Group	63/75%	20/23.8%	1/1.2%	0/0%	0/0%
9. How would you describe instructor to, others?	Excellent	Good	Average	Poor	Dn/da
3 Test Group	22/24.8%	39/43.8%	19/21.3%	9/10.1%	0/0%
5 Test Group	45/52.9%	34/40%	5/5.9%	1/1.2%	0/0%
10 Test Group	51/60.7%	26/31%	6/7.1%	1/1.2%	0/0%
10. Rate your own performance.	Excellent	Good	Average	Poor	Dn/da
3 Test Group	8/8.9%	37/41.6%	41/46.1%	3/3.4%	0/0%
5 Test Group	16/18.8%	45/52.9%	23/27.1%	0/0%	1/1.2%
10 Test Group	16/19.1%	48/57.1%	19/22.6%	1/1.2%	0/0%

[a] Doesn't know or doesn't apply.

[b] Number of responses/Percentage of response