DOCUMENT RESUME

ED 254 554 TM 850 165

AUTHOR Dielman, T. E.; Horvatich, Paula K.
TITLE terrater Reliability and Internal Consistency of :dent and Staff Ratings of Medical Instruction.
PUB DATE Ma. 85
NOTE 12p.; Paper presented at the Annual Meetinç of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; Graduate Medical Education; Higher Education; *Interrater Reliability; *Medical School Faculty; Peer Evaluation; *Rating Scales; Student Evaluation of Teacher Performance; *Teacher Evaluation; *Test Reliability
IDENTIFIERS *University of Michigan Medical School

ABSTRACT

The purposes of this study were to establish the interrater reliability, dimensionality, and internal consistency of an instruction evaluation instrument used at The University of Michigan Medical School. Using the nine-item rating scale, 1,758 student ratings and 88 staff ratings were gathered on 61 faculty. Interrater agreement ranged from .28 to .53 among students, from .11 to .70 among staff, and from .28 to .70 between students and staff. Separate factor analyses of student and staff data showed all items except one exhibited high loadings on a single factor. The eight items forming this factor were summed with unit weighting to form a total "Teaching Quality" index score for each instructor. Internal consistency (Cronbach's alpha) coefficients for this index were .92 for student raters and .93 for staff raters. Although this evidence for reliability is compelling, the question of validity remains. Validation methods depend on how the instructional ratings are used. The instructor evaluation form is appended. (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED254554

Interrater Reliability and Internal Consistency of Student and Staff Ratings of Medical Instruction

T. E. Dielman, Ph.D.
The University of Michigan Medical School

Paula K. Horvatich, Ph.D.
The Medical College of Virginia

ABSTRACT

A nine-item rating scale was developed and employed to rate the teaching effectiveness of 61 medical school faculty. A total of 1,758 student ratings and 88 staff ratings were gathered. Interrater agreement on the nine items ranged from .28 to .53 among students, from .11 to .70 among staff, and from .28 to .70 between students and staff. Separate factor analyses of the student and staff data showed all items except one to exhibit high loadings on a single factor. The alpha coefficients for a "Teaching Quality" index were .92 and .93 for students and staff, respectively.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. E. Dielman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

TM 850 165

BEST COPY

Interrater Reliability and Internal Consistency of Student and Staff Ratings of Medical Instruction

T. E. Dielman, Ph.D.
The University of Michigan Medical School

Paula K. Horvatich, Ph.D.
The Medical College of Virginia

Objectives

The objectives of the study were to establish the interrater agreement and internal consistency of ratings of medical instruction. Three types of interrater agreement were of interest: (1) agreement among students, (2) agreement among staff raters, and (3) agreement between student and staff raters. Internal consistency (Cronbach's alpha) coefficients were calculated separately for student and staff raters.

Perspective

Student ratings of instruction have been employed to assess strengths and weaknesses of courses, the teaching effectiveness of individual instructors, and for faculty development purposes for many years. Many questionnaires and rating scales and systems have been developed to elicit student opinion. However, as pointed out by Irby, et al. (1976) most student rating instruments have been designed for single-instructor courses. In medical schools, many basic science and preclinical sequences are taught via the lecture method by multiple instructors, each representing an area of specific content expertise. Morris (1976) suggested that items for student rating instruments need to be designed to fit the particular instructional format (e.g., lecture format). Therefore, Irby, et al. (1977) and Morris (1976) recommended that student rating questionnaires focus on the following aspects, among others, of lecture effectiveness: (1) organization and clarity of presentation, (2) appropriate use of media, (3) ability to stimulate interest, (4) appropriate level of difficulty, (5) paced to facilitate student note-taking, (6) clarity of

BEST COPY

objectives, (7) relevance to educational goals, and (8) relationship of examinations to the lecture.

Using a 13-item student evaluation instrument that included the instructional aspects suggested above, Irby, et al. (1977) explored students' ability to judge teaching effectiveness in multi-instructor courses typical of medical education. Their study demonstrated that students did discriminate among instructors and that these discriminations were stable over time, indicating that medical students' ratings of instructor effectiveness in multi-instructor courses provide reliable feedback. Reporting on the number of ways to establish the reliability of student evaluation instruments, Doyle (1975) reviewed several studies that included measures of internal consistency. Of these, one study used Cronbach's alpha and obtained internal consistency coefficients that ranged from .80 to .89. The same study also examined the internal consistency of colleague ratings which ranged from .65 to .86. Doyle (1975) recommended that student ratings be supplemented by other sources of information such as colleague ratings and noted the deficiency of the literature in reporting the reliability of these additional sources. According to Costin, et al. (1971) another way to validate student ratings is to compare them with ratings made by colleagues. Although few such studies have been conducted, positive but low correlations between colleagues' and students' ratings have been demonstrated. Both Doyle (1975) and Costin, et al. (1971) concluded that peer ratings were less reliable than students' ratings and that correlations between colleague and student ratings are low because colleagues do not experience the same amount of exposure to an individual instructor's teaching as do students. Moreover, getting colleagues to spend the amount of time it would take to observe an instructor's course makes peer rating studies impractical.

Studies of this nature have not been reported in medical education. However, the multi-instructor courses in medical schools, in which an individual instructor may actually

BEST COPY

lecture just once, or at most several times, provide a better opportunity to compare colleague and student ratings of instruction than is practical in single-instructor university courses. The purposes of this study were to establish the interrater reliability, dimensionality, and internal consistency of an instrument used to evaluate instruction at The University of Michigan Medical School (UMMS). Three sets of interrater reliabilities were considered: (1) agreement among students, (2) agreement among professional staff, and (3) agreement between students and professional staff. Internal consistency coefficients (Cronbach alphas) were calculated separately for student and staff ratings.

Method

A total of 61 faculty members who lectured at least four hours in two large, introductory courses at UMMS were selected as ratees. The raters were a randomly selected sample of first and second year medical students attending UMMS during three consecutive terms, as well as four professional staff members from The Office of Educational Resources and Research at UMMS. During the first term, the random sample of students consisted of 63 first-year and 49 second-year students. During the second term, there were 61 first-year and 49 second-year students. During the third term, 28 first-year and 43 second-year students participated in the evaluations. The number of first-year students during the last term of the study decreased because one of the first-year courses was discontinued. Two professional staff raters attended each lecture session which was to be rated by professional staff. Therefore, a total of 88 ratings were made by the professional staff on 44 lecture occasions. A total of 1,758 ratings of 61 faculty were made by students. There were 34 lecture occasions rated by both students and professional staff. Throughout the study, the students rated each instructor on the nine-item rating form immediately after his/her last lecture. The professional staff observed the instructors' performance on seven of the nine items. The rating form is shown in Figure 1. Two items required judgments about examinations and content overlap which could not be ascertained from the lectures observed by professional staff.

BEST COPY

Results

The interrater reliabilities for each of the items were computed by calculating rho coefficients among all student raters and among all staff raters. The interrater reliabilities among students, among staff raters, and betwen student and staff raters are shown in Table I. The range of interrater reliabilities for the students was from .28 to .53 with a median coefficient of .37. All of the students' interrater reliabilities were significant beyond the .001 level, and all of them except one exceeded .30. The interrater reliabilities among staff members ranged from .11 to .70 with a median coefficient of .50. Five of the seven staff agreement coefficients were significant at least at the .01 level. Pearson product-moment correlation coefficients were computed for the average ratings of students and staff to determine reliabilities across rater types. The agreement between staff and students reached at least the .01 level of significance in six of the seven instances. The agreement coefficients between students and staff ranged from .28 to .70 with a median coefficient of .60.

The factor structure of the rating scale was examined separately for the student and staff ratings. All but one of the nine items were highly intercorrelated. The matrices of intercorrelations are shown in Table 2. The correlations above the diagnoal in Table 2 are the item intercorrelations based on staff ratings, and item intercorrelations based on student ratings are below the diagonal. In each case the factor analytic procedure was the iterative principal axis method with iteration ceasing when the communality estimates converged in the third decimal place. In both analyses all of the items except one loaded highly on the first general factor. The factor loadings based on student ratings and staff ratings are shown in Table 3. The items which formed the first factor were summed with unit weighting to form a total "Teaching Quality" index score for each instructor. As shown in the last two rows of Table 3, the Cronbach alpha coefficients for the index score were .92 and .93, respectively, for student and staff ratings when item

BEST COPY

three was excluded from the index. When item three was included in the index the Cronbach alpha coefficients were .89 and .90, respectively, for student and staff ratings.

Discussion

The results of the study indicate that at least five of the nine rating items exhibited acceptable levels of interrater reliability. Staff ratings on two items demonstrated low levels of interrater reliability. With the exception of one item, the ratings rendered by both students and staff could be combined into a unidimensional "Teaching Quality" index which exhibited a high degree of internal consistency. The items employed in the scale are sufficiently general that they are generalizable to most, if not all, lecture format courses. Although the evidence of reliability presented here is considered compelling, the question of validity remains. As usual, the question, "Validity for what purpose?" must be asked. If the purpose of rating instructional quality is to provide constructive feedback to instructors with the goal of achieving behavioral change, then a longitudinal study of instructional behavior, with intervening feedback for some instructors and not for others, is indicated. If a summative evaluation is desired, for academic promotion or other purposes, then the agreement between students and staff provides some evidence for consensual validity as well as interrater reliability. In such instances, however, it would be desirable to employ as many different sources of performance data as possible in addition to ratings by students and colleagues. Such methods might include membership on teaching committees at local, regional and national levels, student achievement data, alumni review, and/or other objective data concerning teaching activities.

BEST COPY

REFERENCES

Costin, F., Greenough, W., and Menges, R. Student Ratings of College Teaching: Reliability, Validity, and Usefulness. Review of Educational Research, 1971, 41(5), 511-535.

Doyle, K. Student Evaluation of Instruction. Lexington, Massachusetts: Lexington Books, 1975.

Irby, D., DeMers, J., Scher, M., and Matthews, D. A Model for the Improvement of Medical Faculty Lecturing. Journal of Medical Education, May 1976, 51, 403-409.

Irby, D., Shannon, N., Scher, M., Peckham P., Ko, G., and Davis, E. The Use of Student Ratings in Multi-Instructor Courses. Journal of Medical Education, August 1977, 52, 668-673.

Morris, V. A Positive Approach to the Utilization of Student Feedback in Medical Education. Journal of Medical Education, July 1976, 51, 541-545.

BEST COPY

Figure 1
INSTRUCTOR EVALUATION FORM

1. How would you rate the organization of the presentations made by this instructor?

7	6	5	4	3	2	1
Very Well Organized			Somewhat Organized			Confusing

2. To what extent did the material presented by this instructor overlap with material presented by other instructors in this course?

7	6	5	4	3	2	1
Too Much			Just Right			Too Little

3. At what rate did the instructor present his/her material?

7	6	5	4	3	2	1
Much Too Fast			Proper Rate			Much Too Slow

4. Given the content he/she was teaching, how interesting did the instructor make the material?

7	6	5	4	3	2	1
Exceedingly Interesting			Interesting			Boring

5. How effectively did the instructor use the teaching techniques which he/she selected?

7	6	5	4	3	2	1
Very Effectively			Somewhat Effectively			Not At All Effectively

6. How clear was your understanding of his/her expectations regarding what you were to learn?

7	6	5	4	3	2	1
Exceedingly Clear			Clear			Not At All Clear

7. How would you rate your understanding of the content resulting from his/her teaching?

7	6	5	4	3	2	1
Complete Understanding			Adequate Understanding			Total Confusion

8. What was the relationship between the content taught by this lecturer and the examination questions asked?

7	6	5	4	3	2	1
Nearly Perfect Relationship			Fair Relationship			No Relationship At All

9. What overall rating would you give this instructor compared to other medical school instructors you have had?

7	6	5	4	3	2	1
Outstanding			Average			Poor

BEST COPY

TABLE I

Interrater Reliabilities
(Rho coefficients)

Item		Among Students (N=1,758)	Among Staff (N=88)	Between Staff and Students (N=34)
1.	Organization of presentations	.37***	.50***	.70***
2.	Overlap with other courses	.53***	N/A	N/A
3.	Presentation rate	.28***	.11	.50**
4.	Interest of presentation	.42***	.61***	.67***
5.	Effectiveness of teaching techniques	.37***	.58***	.60***
6.	Students' understanding of expectations	.31***	.11	.28
7.	Students' understanding of content	.30***	.31**	.54***
8.	Relationship between content and exam questions	.31***	N/A	N/A
9.	Overall rating compared to other instructors	.47***	.70***	.69***

**p <.01
***p <.001

N=total number of observations

BEST COPY

TABLE 2

Missing Data Correlations among Items;
Staff Rating Coefficients above Diagonal,
Student Rating Coefficients below Diagonal

Item No.	1	3	4	5	6	7	9
1	---	.04	.57	.65	.71	.71	.73
3	-.06	---	-.01	-.41	.14	.00	-.15
4	.57	.03	---	.76	.80	.57	.84
5	.66	-.06	.74	---	.61	.53	.86
6	.55	-.13	.50	.56	---	.84	.68
7	.62	-.17	.63	.66	.68	---	.70
9	.72	-.04	.81	.80	.61	.73	---

BEST COPY

TABLE 3

Results of Principal Axis Factor Analysis and Cronbach Alphas Based on Items Rated by Both Students and Staff

	Students		Staff	
Item No.	$\underline{V}$ fp	$\underline{h}^2$	$\underline{V}$ fp	$\underline{h}^2$
1	.66	.64	.49	.71
3	.11	.21	.00	.00
4	.70	.70	.41	.71
5	.80	.76	.82	.73
6	.69	.57	.30	.73
7	.76	.71	.86	.70
9	.90	.85	.99	.87
Eigenvalue	4.44		4.46	
s^2 accounted for	63%		64%	
alpha (including item 3)	.89		.90	
alpha (excluding item 3)	.92		.93	

BEST COPY