Florida Journal of Educational Research
Fall 1985, Vol. 27,
No. 1, Pp. 19-26

The Agreement and Structure of Colleague Ratings

Maria M. Llabre
University of Miami

Harry W. Forgan
University of Miami

ABSTRACT. The interrater reliability and the factor structure of colleague ratings of university faculty were studied. Two approaches to rating colleagues were compared: a global rating and an analytic rating in four areas of responsibility. Reliability indices indicated that means based on four raters using the analytic method were reliable. Results of a factor analysis indicated the presence of two factors: research and publications and, to a lesser extent, teaching.

Judgements of faculty effectiveness are made annually by administrators in order to award merit pay, promotion, and tenure. These judgements are typically based upon evaluations of faculty work in four areas including teaching; advising; contributions to the department, school and university; and research and publication. Evaluations of teaching effectiveness are often obtained from students as well as colleagues, but evaluations of the other three areas are usually obtained only from colleagues. Although there is general agreement about the types of faculty responsibilities that are to be rated, there are widespread differences in how raters are chosen, the degree of their familiarity with the person being judged, the leniency of the evaluation, the raters' attitudes toward the evaluation process, and the methods used to obtain ratings (Centra, 1980). These factors can influence the reliability of the ratings

obtained. In addition, Fenker (1975) identified faculty opposition to the use of colleague ratings for determining merit pay, promotion, and tenure since there is competition among the faculty performing the ratings for the administrative awards.

Since the administrative decisions to be made about faculty are important, and colleague ratings are used often for these decisions, the accuracy of colleagues' ratings should be studied. When student and colleague ratings of teaching effectiveness have been compared, the two groups have been found to be in general agreement although colleagues' ratings were not as reliable (Blackburn & Clark, 1975; Centra, 1975). Doyle and Crichton (1978) reported good convergent validity and somewhat less adequate discriminant validity for colleague ratings of teaching. These same authors found colleague rankings to be better in convergent and discriminant validity than colleague ratings. Centra (1980) questioned the validity of colleague ratings for teaching performance; however, he believes they are useful for evaluating publications and research.

At the University of Miami in Coral Gables two different methods of colleague ratings have been used. The purpose of this study was to compare the reliability of the scores obtained using two different methods. In addition, the extent to which different dimensions of performance could be identified using one of the methods was examined.

Method

The subjects (N=46) were faculty members from four departments in the School of Education and Allied Professions at the University of Miami. The departments involved were: Educational Psychology; Elementary Education; Educational Leadership and Instruction; and Health, Physical Education, and Recreation.

Within each department, every faculty member rated every other member using both methods. The ratings were done in the Spring of two consecutive years. Method 1, used in 1981, consisted of a global 5-point scale ranging from "outstanding" to "unsatisfactory."

The rating on this scale was to represent the overall performance of the faculty member. Method 2, used in 1982, consisted of a 7-point analytic scale which included four areas of responsibility: (1) teaching. (2) advising, (3) contributions to the Department, School, and University, and (4) contributions to the profession. For each area the rater specified a weight, between one and ten, which indicated the extent to which that area should be counted toward the total evaluation score. Certain constraints limited the possibility of the weights so that teaching must get a weight between four and seven, while each of the other three areas must get a weight between one and four. The result of Method 2 is a scale with increased variability compared to the one obtained using Method 1.

The first question addressed was whether the increase in variance resulted in an increase in true variance, i.e., true differences among faculty, or an increase in error variance. Reliability estimates were obtained for the scores from Method 1 and Method 2 for each department separately using a generalizability approach (Cronbach et al., 1972). To conduct this analysis, the data were summarized in the form of a rater by ratee matrix. Since faculty did not rate themselves, the diagonal elements were blank and were replaced by the mean score for that ratee. The reliability estimates computed as true variance were the variability among ratees. Error variance was made up of both the systematic variance of raters as well as the variance of the interaction of raters by ratees. Estimates were obtained for the reliability of a single score and the average score. The latter score is the one that was actually used in making merit decisions; however, since the average score was based on different numbers of scores for each department, these reliabilities are not comparable across departments.

The second question addressed was whether colleague ratings under Method 2 reflect separate dimensions of a professor's role or a more general effect. A factor analysis was done using the combined information for all four departments. To carry out this analysis, the following was done. First the scores from six raters

were chosen at random for each person. Two separate scores, each the total of 3 raters, were obtained under each of the four areas for each person. These eight totals were the variables used in the factor analysis of 46 cases. A principal axis solution was obtained using communality estimates in the diagonal and rotated to varimax criterion. Four factors were specified for the rotation.

Results

The reliability coefficients obtained for each of the departments under the two rating methods are included in Table 1. Coefficients are provided for one observation and for an average of two, four, six and N observations. Method 2 resulted in higher reliability coefficients than did Method 1 for the Elementary Education and Educational Leadership Departments. Method 1 resulted in better reliability for the Educational Psychology Department. Coefficients were similar using the two rating methods in the Health, Physical Education and Recreation Department, although reliabilities were slightly higher for Method 2.

Table 2 contains the factor loadings and the percentage of common variance accounted for by each factor for the four factor solution. It seems that raters did judge their colleagues on separate dimensions, however, the last two factors accounted for very little of the common variance.

Discussion

Of the two methods of colleague ratings that were compared in this study, the analytic approach, Method 2, resulted in high reliability estimates for most departments. However, the estimates did not reach an acceptable level when only one or two colleagues were considered in the rating regardless of the method used. In most departments, using four or more colleagues yielded reliable mean ratings. The pattern in the Educational Psychology Department, which was different from that in other departments, may be partly explained by the composition and

Table 1

Reliability Coefficients for Estimates of Mean Ratings

	Rating Method 1: Global 5-point Scale for Overall Performance				Rating Method 2 : 7-point Scale for Four Separate, Weighted Areas of Responsibility			
	Number of Observations							
Department	1	2*	4	N	1	2	4	N
Educational Psychology (n=15)	.300	.462	.632	.866	.128	.226	.369	.678
Elementary Education (n=10)	.162	.278	.435	.658	.554	.713	.833	.926
Educational Leadership (n=10)	.215	.354	.523	.751	.301	.463	.632	.812
Health Physical Education and Recreation (n=6)	.472	.641	.782	.843	.480	.649	.787	.847

*Scores are average of 2, 4, or N observations

Table 2

Factor Loadings for Four Factor Solution and the Percentage of Common Variance for Each Factor

Area of Responsibility	Randomly Selected Rater Groups	Factors 1	2	3	4
Teaching	T1*	.17	.81	.37	.17
	T2	.10	.78	.28	.32
Advising	A1	.10	.37	.83	.32
	A2	.12	.44	.70	.31
Contributions to the Department, School, and University	D1	.39	.25	.34	.72
	D2	.24	.37	.37	.76
Contributions to the Profession	P1	.88	.05	.04	.21
	P2	.94	.18	.14	.14
Percent of Common Variance for Each Factor		68.7	20.2	6.7	4.4

*T1 represents the scores obtained from three randomly selected raters, and T2 represents the scores obtained from the other three raters for each individual.

organization of that department. Unlike the other departments, it is made up of four separate areas, and the faculty offices are not in close proximity. A consequence of this arrangement is that colleagues are not always familiar with the work of others and they depend on the examination of supporting evidence for their ratings. The type and availability of evidence was not comparable for the faculty in this department. Clearly, when evidence was not available on which to base the ratings, the more impressionistic approach (Method 1) produces more reliable ratings.

The results of the factor analysis indicate that raters were judging their colleagues on separate dimensions. Of the four possible dimensions provided by Method 2, only two accounted for a large proportion of variance. The first factor, accounting for about two thirds of the common variance, is defined in terms of contributions to the profession (P1, P2). This finding supports Centra's (1980) position that colleague judgment maybe more influential in the areas of research and publications than in areas that directly involve students. The second factor, defined in terms of teaching, accounted for about 20 percent of the common variance. This finding suggests that colleagues do identify individual differences in teaching performance, independent of performance in research and publications. Once differences in these two areas are noted, other responsibilities of faculty members do not result in major dimensions of individual differences as perceived by colleagues.

References

Blackburn, R. T. & Clar, M. J. (1975). An Assessment of faculty performance: Some correlates between administrator, colleague, student and self ratings. Sociology of Education, 48, 242-256.

Centra, J. A. (1975). Colleagues as raters of classroom instruction, Journal of Higher Education, 46, 327-337.

Centra, J. A. (1980). Determining Faculty Effectiveness, Washington, DC: Jossey-Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Doyle, K. O. & Crichton, L. I. (1978). Student, peer, and self-evaluations of college instructors. Journal of Educational Psychology, 70, 815-826.

Fenker, R. M. (1975). The evaluation of university faculty and administrators: a case study. Journal of Higher Education, 46, 665-686.

AUTHORS

MARIA M. LLABRE, Associate Professor, Department of Educational and Psychological Studies, University of Miami, Coral Gables, Florida 33124

HARRY W. FORGAN, Professor, Department of Teaching and Learning, University of Miami, Coral Gables, Florida 33124