FACULTY EVALUATION: Reliability of Peer Assessments of Research, Teaching, and Service

Lawrence S. Root

In this paper, assessments of faculty performance for the determination of salary increases are analyzed to estimate interrater reliability. Using the independent ratings by six elected members of the faculty, correlations between the ratings are calculated and estimates of the reliability of the composite (group) ratings are generated. Average intercorrelations are found to range from 0.603 for teaching, to 0.850 for research. The average intercorrelation for the overall faculty ratings is 0.794. Using these correlations, the reliability of the six-person group (the composite reliability) is estimated to be over 0.900 for each of the three areas and 0.959 for the overall faculty rating. Furthermore, little correlation is found between the ratings of performance levels of individual faculty members in the three areas of research, teaching, and service. The high intercorrelations and, consequently, the high composite reliabilities suggest that a reduction in the number of raters would have relatively small effects on reliability. The findings are discussed in terms of their relationship to issues of validity as well as to other questions of faculty assessment.

Performance evaluation is one of the most difficult aspects of supervision. In part, this can be attributed to the personal difficulty of having to disappoint someone who had expected a better rating. More fundamentally, however, evaluations are problematic because of the conflicting objectives of the process. On the one hand, an evaluation is intended to indicate the extent to which work meets a particular standard or expectation for performance. On the other hand, evaluations are part of a larger personnel goal: stimulating improvement in that performance. Unfortunately, these two objectives may be in conflict, with the result that the "summative" aspects of evaluation undercut their "formative" or developmental objectives (French-Lazovik, 1981, p. 74). The difficulty of providing critical feed-

Lawrence S. Root, School of Social Work, The University of Michigan, Ann Arbor, MI 48109-1285.

Research in Higher Education Vol. 26, No. 1

back without engendering counterproductive behavior (e.g., alienation and discouragement) is evidenced in the plethora of "how-to" manuals destined for the bookshelves of personnel directors.

The same dilemmas exist in the university, but the situation is complicated by the nature of academic work. Evaluation of faculty performance raises additional difficulties. First, performance is often difficult to measure. The "products" of faculty work—an educational experience for students, development of knowledge, and service to the institution and community—are varied in form and function, exacerbating problems of equitable treatment of colleagues. For example, research is often discussed in terms of publications, but even that seemingly straightforward measure has various interpretations. Across disciplines, how do the one- or two-page articles reporting on experimental outcomes in the physical sciences compare with the lengthy interpretative essays more common in the humanities? How should chapters in a book be compared with refereed journal articles? And, for that matter, what *is* a "refereed journal?" (Miller and Serzan, 1984).

Beyond issues of definition, assessments of the *quality* of research—its originality and its contribution to the field—raise profound and contentious questions in the review process. Even the peer review process itself has come under question, with concerns raised about consistency and bias in such basic areas as selection of articles for publication in professional journals (e.g., Ceci and Peters, 1982).

The challenges of faculty evaluation have received considerable attention in the professional literature, although the lion's share has gone to the evaluation of teaching (e.g., see Centra, 1979; Doyle, 1983; Eble, 1972; Hildebrand, Wilson, and Dienst, 1979). This is true despite the generally held view that "publications are paramount" in the determination of salary and promotion. (See Katz, 1973, p. 471; Tuckman and Gapinski, 1977; Kasten, 1984; for some exceptions, see Johnson and Kasten, 1983; Miller, 1978.)

In this article, we address the issue of reliability in collegial assessment of faculty performance. To do this, we examine a system of peer review for salary determination developed and used in a graduate professional school of a large research university. The process is based upon independent assessments of faculty performance, followed by group discussion of the individual ratings. Our emphasis is on assessing the reliability of the independent judgements of colleagues in the areas of research, teaching, and service.[1] Although there may be some idiosyncracies associated with the setting, the findings are relevant to faculty evaluation policies and procedures in institutions of higher education more generally.

THE PROCESS OF FACULTY ASSESSMENT FOR ANNUAL SALARY INCREASES

In the early 1980s, under the leadership of a new dean, the school that is

our setting altered its governance structure by electing an Executive Committee, with three members representing the professorial ranks and three members elected at large. This Committee replaced the Faculty Council, which had been advisory to the dean. While the Faculty Council had traditionally been consulted in salary decisions, the decisions rested with the dean. Under the new Executive Committee, salary decisions moved more squarely under the purview of that elected group.

Criteria for assessment for salary review parallel those established for promotions. These criteria represent a broad view of the range of faculty activities which are to be included in an annual assessment. The guidelines for annual reviews are articulated in a 14-page memo which includes the criteria for each of the three areas as well as suggestions of the types of information which a faculty member might provide to enable the Executive Committee to assess his or her work.

Each faculty member prepares a report of activities during that year, and this report serves as the basis upon which the Executive Committee determines its ratings. These annual reports vary greatly in detail and length. After the initial year of this new evaluation system, a form was introduced to encourage more uniform reporting. Its use, however, is not mandatory. The availability of the form has increased consistency of reporting, but large differences remain in the way individual faculty members explicate and document their activities.

Faculty members append a variety of documents to support their narrative reports. For research, these generally include copies of materials published during the year, grant proposals prepared, and copies of works in progress. Faculty also include documentation of particularly meritorious reviews, such as awards for scholarship.

The documentation of teaching takes varied forms. The Executive Committee has access to statistics on workload (what courses were taught, size of the classes, number of advisees, participation on preliminary examination and dissertation committees). Most faculty include course outlines and syllabi in their annual reports as well as other teaching material which illustrates the pedagogical approach taken. Student evaluations of courses are also usually submitted as part of the record. Other aspects of teaching are more idiosyncratic and are documented in differing ways (e.g., collaboration in teaching, curricular development, tutorials offered).

The documentation of service is the least consistent of the three areas. Evidence in the form of the *product* of the service in the university (e.g., committee reports, proposals) and recognition from outside groups (e.g., awards for service) are usually included.

Once the annual reports are submitted to the dean, they are made available to the Executive Committee. In the year studied (assessments in the spring of 1985 for the 1984–85 academic year), each Committee member

independently rated each faculty member in the three areas. These individual ratings were then submitted to the Dean's Office prior to group discussion. The data analyzed in this paper are those ratings.

The Committee uses an eight-point scale for the assessment of faculty performance:

0	unsatisfactory
1	marginal
2	satisfactory
3	(intermediate rating)
4	substantial
5	(intermediate rating)
6	superior/outstanding
7	exceptional

This point system is directly related to salary increments. Teaching and research are of equal value while service counts as one-half of that value. To achieve this weighting, the ratings in teaching and research are multiplied by two. Thus there is a maximum of 14 points for research, 14 for teaching, and 7 for service. In the determination of salary increments, each point above 10 (the equivalent of a "satisfactory" rating in each of the areas) results in a fixed dollar increment (determined subsequently in light of budgetary availability). For example, a rating of four in research, three in teaching, and three in service would result in an aggregate rating of 17 (8, 6, and 3). If the salary increment for that year was $250 for each point over 10, the increase would be $1750 (that is, seven points above the 10-point cutoff; seven multiplied by $250).

The specifics of the committee process have evolved over time. In preparation for the 1984–85 evaluations, the Executive Committee implemented some basic training techniques in an effort to establish a common understanding of the ratings. The initial step involved reviewing and discussing the criteria. Some continuity came from earlier experience; four of the six members had been involved in the procedure the previous year. In addition, cases from that year were selected to illustrate the levels of performance which had previously been associated with high and low ratings. This exercise was intended to encourage greater continuity and consistency over time as well as to clarify the criteria employed.

RATINGS OF FACULTY BY A SIX-PERSON EXECUTIVE COMMITTEE

In terms of the outcome of this process, the overall ratings of faculty

TABLE 1. Means and Standard Deviations of Ratings

Raters	Research Mn	Research SD	Teaching Mn	Teaching SD	Service[a] Mn	Service[a] SD	Total Mn	Total SD
A	7.3	3.22	7.8	1.91	2.7	1.53	17.6	4.28
	(0.441)[b]		(0.245)		(0.567)		(0.243)	
B	7.3	2.29	8.0	2.31	3.8	1.40	19.2	4.73
	(0.314)		(0.289)		(0.368)		(0.246)	
C	7.9	3.59	8.5	2.00	2.8	1.52	19.0	4.77
	(0.454)		(0.235)		(0.543)		(0.251)	
D	7.5	3.80	8.8	1.92	3.5	1.31	19.5	5.49
	(0.507)		(0.218)		(0.374)		(0.282)	
E	7.6	2.98	8.2	2.10	4.4	1.34	19.8	4.53
	(0.392)		(0.256)		(0.305)		(0.229)	
F	7.7	2.60	8.4	1.63	3.1	0.91	19.0	3.42
	(0.338)		(0.194)		(0.294)		(0.180)	
Avg.[c]	7.6	0.21	8.3	0.33	3.4	0.59	19.0	0.69
	(0.028)		(0.040)		(0.174)		(0.036)	

[a]The ratings for research and teaching are doubled in the rating process to correspond with their weighting in the determination of salary increments. The mean ratings for service are *not* multiplied by a factor of two.
[b]Coefficients of variation (standard deviation divided by mean) are displayed in parentheses.
[c]These standard deviations represent the variance among the mean ratings of the six raters.

members ranged from 12 to 26, with an average of about 19. If we assume the pay increment for each point above 10 to be $250 (which was not the actual dollar equivalent in that year), the lowest merit increase in salary would have been $500, $4,000 the highest, with a median of $2250 (mean of $2240).

Summary measures of the ratings of Executive Committee are displayed in Table 1. The average ratings given by members in each of the three areas and for the total of the three show a high level of consistency across raters. The averages for research range from a low of 7.3 to a high of 7.9. The overall average for research for the six raters is 7.6, reflecting a basic rating of 3.8, just below "substantial," before being multiplied by 2 to yield the aggregate weighted measure.

The ratings for teaching are somewhat higher, averaging 8.3 (just over 4 on the 8-point scale). It should be noted that the standard deviations of the ratings by the individual raters tend to be smaller for teaching than for research, indicating that five of the six raters (all but rater B) found less variation in the teaching than in research. Comparing the coefficients of variation, which control for differences in the respective means, the variance for rater B is also lower for teaching.

TABLE 2. Correlations between Ratings in Research, Teaching, and Service

Ratings	Correlation Coefficient (Pearson's *r*)	Rank-Order Corr. Coeff. (Kendall's Tau)
Research × Teaching	0.185	0.130
Research × Service	0.197	0.112
Teaching × Service	0.277	0.205

The average ratings of service by all but rater E are lower than for the other two areas. For direct comparability with the ratings in teaching and research, service ratings must be multiplied by two.

The total ratings (sum of the ratings in each of the areas) reflect the mean overall ratings of each faculty member by the individual raters. Comparing the coefficients of variation, we see that for four of the six raters there is less variation in the total score than in any of its three constituent parts. For the other two raters, only the variation in their ratings of teaching is lower than that of their total ratings.

Previous studies suggest that there is little or no correlation between the performance levels of faculty members in the three areas of research, teaching, and service (Centra, 1979, p. 34; Machalak and Friedrich 1981, pp. 594–5). In order to explore this, an analysis was undertaken to examine the relationship between assessments of performance in the three areas of research, teaching, and service. Correlations between the areas were calculated, treating the ratings both as interval and ordinal measures. These findings are displayed in Table 2.

The correlations suggest very weak positive associations between the ratings of individual faculty members in the three areas. The slightly larger coefficients obtained when the data are treated as interval measures (rather than ordinal) suggest that there may be somewhat more association of the ratings for those at the extremes of the ratings. Examining the five cases with the highest and the five with the lowest research ratings, we have the results shown in Table 3. The higher research ratings find some reflection in teaching and service, but the differences are small. While there are certainly examples of individuals who excel in all areas, systematic patterns of relationship between research, teaching, and service are not evidenced in the data emerging from these assessments.

When we look at the individual raters, we find few clear patterns (see Table 1). Rater A consistently provided the lowest ratings, but the differences between these averages are not large, and all of the raters have compa-

TABLE 3. Research Ratings

Average For	Five Highest	Five Lowest
Research	11.4	3.3
Teaching	8.1	7.2
Service	3.9	3.2

rable levels of variation in their ratings of individual faculty members, suggesting that there was a common sense of the range of performance. Ranking the raters in terms of their average ratings for the three areas, we have the results shown in Table 4. The differences between the overall level of assessments, are very small, and the comparison of average ratings and the associated measures of variation are more striking in their consistency than in the differences observed.

ASSESSING INTERRATER RELIABILITY

In order to assess the extent to which the raters agree in their evaluation of individual faculty members, correlation coefficients were calculated for each pair of raters in the scores assigned in the three areas and for the total ratings. These correlation coefficients are displayed in Table 5. Cases for which there were missing data were excluded from this analysis. In most instances, data were missing because of some singularity in the situation of that faculty member (e.g., special administrative responsibilities which influenced teaching workload; interdisciplinary appointments).

The paired correlation coefficients suggest high levels of interrater reliability, particularly in assessments of research, for which they range from a low of 0.740 (between raters B and A) and a high of 0.915 (raters E and C).[2] The paired correlations are somewhat lower for teaching and service, al-

TABLE 4. Average Ratings

	Research	Teaching	Service
Highest	C	D	E
	F	C	B
	E	F	D
	D	E	F
	B	B	C
Lowest	A	A	A

TABLE 5. Correlation Matrices for Ratings of the Six Raters (Executive Committee)

A. Research

(n = 29 cases with no missing values)

Raters					
B	0.740				
C	0.844	0.844			
D	0.823	0.842	0.846		
E	0.882	0.871	0.915	0.887	
F	0.846	0.857	0.886	0.823	0.848
Raters	A	B	C	D	E

B. Teaching

(n = 25 cases with no missing values)

Raters					
B	0.605				
C	0.459	0.631			
D	0.538	0.716	0.517		
E	0.653	0.586	0.516	0.798	
F	0.503	0.619	0.545	0.746	0.614
Raters	A	B	C	D	E

C. Service

(n = 27 cases with no missing values)

Raters					
B	0.224				
C	0.704	0.451			
D	0.624	0.611	0.764		
E	0.691	0.339	0.701	0.664	
F	0.775	0.331	0.774	0.741	0.709
Raters	A	B	C	D	E

D. Total Rating

(n = 25 cases with no missing values)

Raters					
B	0.700				
C	0.734	0.840			
D	0.753	0.898	0.838		
E	0.830	0.804	0.818	0.837	
F	0.693	0.801	0.780	0.850	0.732
Raters	A	B	C	D	E

TABLE 6. Mean Correlation Coefficients Between Raters

Raters	Research	Teaching	Service	Total
A	0.827	0.552	0.604	0.742
B	0.831	0.631	0.391	0.809
C	0.867	0.534	0.679	0.802
D	0.844	0.663	0.681	0.835
E	0.881	0.633	0.621	0.804
F	0.852	0.605	0.666	0.771
Average Intercorrelation	0.850	0.603	0.607	0.794
Composite Reliability[a]				
6 raters:	0.971	0.901	0.901	0.959
3 raters:	0.945	0.820	0.820	0.920

[a]The reliability of n raters is calculated using the generalized Spearman-Brown formula (e.g., see Guilford, 1954, p. 354).

though still quite high in comparison with expectations for interrater reliability (Cohen and McKeachie, 1980). The correlations for the total scores again fall between the others. The service ratings of rater B present the only apparent anomaly in the paired correlations, with B showing generally lower correlation coefficients (except when paired with rater D).

In order to see more clearly the differences in the judgments of the individual raters, mean correlation coefficients were calculated (see Table 6). For the ratings in the area of research, all of these means are above the 0.800 level. For teaching, the mean coefficients range from 0.534 to 0.663. With the exception of rater B, all of the coefficients for service are over 0.600. Considering all of the paired correlation coefficients, the average intercorrelations for research, teaching, and service were 0.852, 0.603, and 0.607, respectively.

The reliability of the composite ratings by the six committee members was calculated using the generalized Spearman-Brown prophecy formula (Guilford, 1954, p. 354). This statistic can be understood as the correlation of present composite ratings with those expected from another group of raters selected from the same universe. The square of this statistic suggests the proportion of the true variance which is explained by the composite ratings.

The reliability of the six raters in this study is presented in Table 6, and the expected reliability for three raters is also included. For each of the three areas and for the total rating, the composite reliabilities of the six raters are over 0.900. A reduction in the number of raters has little effect on the composite reliability of rating in research or on the total rating because of the very high average intercorrelations. The impact is larger for teaching and

service, but the reliability of the composite ratings of three raters is still relatively high.

DISCUSSION

There are several implications of this analysis for quantitative aspects of faculty evaluation. Beyond this, there are implications for the more qualitative questions of evaluation—the nature of the assessment and rewards within academe.

The most striking finding is the high degree of agreement observed between the raters. When taken together, the composite ratings of the six raters represent a very reliable indication of faculty performance. When we look at the three rating areas, we find the strongest agreement in the area of research, but even the lower correlation coefficients seen in the assessments of teaching and service combine to provide reliable composite ratings.

One practical implication of the high interrater reliability is the possibility of a reduction in the number of raters with only a modest loss of composite reliability. Given the extensive time necessary for performing these assessments, reducing the number of raters represents a significant time savings. The figures in Table 6 suggest that such a change in procedures would retain a high level of composite reliability while halving the number of evaluations to be performed.

The question of "how much reliability is enough" remains a matter of judgment. The marginal increase in reliability gained with the addition of raters falls off rapidly, depending upon the level of interrater reliability. Figure 1 shows the increase gained in reliability with additional raters when the intercorrelations between raters are 0.20, 0.40, 0.60, and 0.80. With high intercorrelation, the curve is almost horizontal, indicating small increases in the reliability of the composite ratings. At lower levels of intercorrelation, the curves don't begin to level off (the marginal additions to composite reliability are very small) until more raters are involved. Three raters are often considered a lower limit for effective decision making in an evaluative exercise (French-Lazovik, 1981, p. 83). Beyond that, the decision about how many should be involved is a function of the interrater reliability and the acceptable level of error.

Two cautionary notes should be made in this discussion of high interrater reliability. First, the results reported here represent a single year of evaluations. Data from prior deliberations are not available, and it may be that the results are attributable in part to the particular individuals involved or other aspects which may be peculiar to the assessment situation. The findings call for follow-up study to assess their robustness.

A second area of caution in the interpretation of the findings involves the

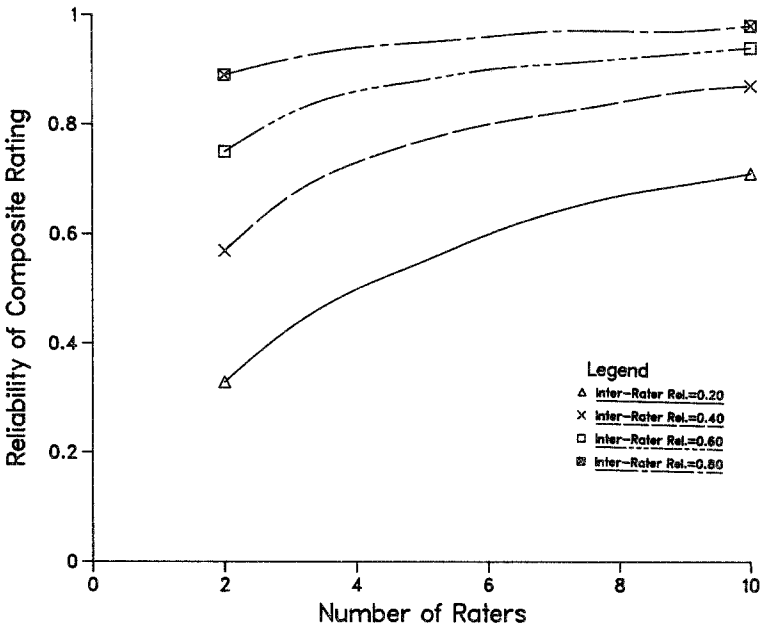

FIG. 1. Reliability of composite ratings with different average intercorrelations.

tendency to interpret reliability measures as a measure of validity. The high levels of reliability observed in the case studied in this paper, while a necessary condition for a good measure, do not indicate that the assessments themselves are valid. In fact, high interrater reliability in judgments of a complex phenomenon may signal shortcomings in the operationalization of that measure, a failure to capture the reality of the phenomenon being studied.

Ratings of teaching provide a good example. Interrater reliability in classroom observation by peers is typically much lower than the levels reported here, low enough to suggest extreme caution in their use in summative evaluations (Centra, 1979, p. 75). The ratings of teaching in the case described herein were based primarily on the description by the faculty member, supported by course materials, measures of teaching workload, and student evaluations. This evidence may exclude some of the more elusive dimensions which make the rating of teaching problematic.

Other questions of validity may also enter into the interpretation of these results. For example, the low correlations observed among research, teaching, and service suggest that there is little or no relationship between performance in these areas. An alternative explanation could arise from the assessment process itself. It could be argued that there is a tendency for

raters to impose an implicit compensatory adjustment in their ratings so that those who receive low ratings in one area may be treated more generously in other areas. Conversely, a high score in research may predispose the rater to be more critical of performance in teaching and service to avoid providing some faculty with very high salary increases.

It is difficult to estimate the extent to which such compensatory adjustments may influence the overall ratings. The coefficients of variation for the individual raters (in Table 1) provide some evidence. This statistic indicates that there is less variation in the overall ("total") assessments of faculty members than in the assessments for the three areas. One interpretation of this is that the actual "overall performance," taking into account research, teaching, and service, is actually more even. Differing areas of strength tend to balance themselves out. On the other hand, the lower observed variation in overall assessments may be interpreted as evidence for the existence of compensatory adjustments.

Beyond the specifics of this evaluation setting, there are broader questions which arise concerning the use of merit pay in universities. It has been argued that merit pay may be ineffective and, indeed, counterproductive for faculty (McKeachie, 1979). On the other hand, an across-the-board salary program can introduce its own set of organizational and personnel problems (Keaveny and Allen, 1983).

If we accept the basic premise that monetary rewards have a place within the university, implementation of any reward system may lead to unwanted second-order effects, particularly in the area of teaching. The increased use of student evaluations, for example, has been the subject of ongoing debate concerning their utility (e.g., Cohen, 1983; Dowell and Neal, 1982; Feldman, 1977; Kulik and McKeachie, 1975; Marsh, 1984). Aside from whether or not student evaluations are an accurate reflection of the quality of teaching, the fact of their use may have negative effects on faculty. For example, grade inflation may be fueled by the popular belief that low grades are associated with more negative evaluations by students. Such a view finds some documentary support (Powell, 1977). In a more general sense, it is reported that the use of student evaluations lowers faculty morale and faith in the university administration and influences instructors to lower their expectations for student performance (Ryan, Anderson, and Birchler 1980).

These unintended effects reinforce the need to examine carefully merit pay systems in the university. For such a system to be successful the participants must believe that increased efforts on their part will result in real returns (Greene and Wallace, 1984). For the faculty member who has not been "productive" in the past with regard to publication, this premise may not hold. In such cases, alternative reward structures which recognize the diverse nature of the faculty contributions may be appropriate (McKeachie 1979).

A successful merit pay system also depends on the belief in its fairness and the fairness of its application. Traditionally, salary and promotion decisions

have been made "in an intuitive manner with seldom any clear understanding of the weights they are attaching to various criteria" (Katz, 1973, p. 476). More recently there has been a greater emphasis on developing more explicit standards and procedures. This change has been stimulated in part by the expectations of procedural fairness in judicial arenas (Lee, 1985). One basis for defining "justice" is procedural: justice is that which emerges from a process which is fair (e.g., Rawls, 1971). In the context of the university, creating open and reliable procedures for salary determination is one step in creating a fair system. Ensuring that the criteria used are adequate reflections of performance in the areas of research, teaching, and service is a second critical component in the development of an effective and equitable compensation program.

Acknowledgments. I am grateful to James A. Kulik, Research Scientist in the University of Michigan's Center for Research on Learning and Teaching, for his excellent advice on the methodological issues associated with interrater reliability, and to William Birdsall and Tony Tripodi for their thoughtful critiques of the manuscript.

NOTES

1. In this article, the term "research" is used to include the variety of activities included in "knowledge development," a term which may connote a broader range of pursuits than the term used at the university at which this study was undertaken.
2. The data in this article represent the universe of ratings, rather than a sample from that universe, therefore the concept of "statistical significance" does not formally apply. The familiarity of such measures, however, makes it useful as a general indication of the strength of a relationship. In this case, if a sample had been involved, coefficients greater than about 0.500 would be considered statistically significant at the $p < .01$ level.

REFERENCES

Ceci, S. J., and Peters, D. P. (1982). Peer review: A study of reliability. *Change* 14(6): 44–48.

Centra, J. A. (1979). *Determining Faculty Effectiveness.* San Francisco: Jossey-Bass.

Cohen, P. A. (1983). Comment on 'A selective review of the validity of student ratings of teaching.' *Journal of Higher Education* 54: 448–458.

Cohen, P. A., and McKeachie, W. J. (1980). The role of colleagues in the evaluation of college teaching. *Improving College and University Teaching* 28: 147–154.

Dowell, D. A., and Neal, J. A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education* 53: 51–62.

Doyle, K. O., Jr. (1983). *Evaluating Teaching.* Lexington, MA: Lexington Books.

Eble, K. E. (1972). *The Recognition and Evaluation of Teaching.* Washington, D.C.: American Association of University Professors.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: a review and analysis. *Research in Higher Education* 6: 223–274.

French-Lazovik, G. (1981). Peer review: documentary evidence in the evaluation of teaching. In J. Millman (Ed.), *Handbook of Teacher Evaluation,* pp. 73–89. Beverly Hills, CA: Sage.

Greene, R. J., and Wallace, M. J. (1984). Is there/can there be merit in merit programs? *ACA 1984 Conference Proceedings,* pp. 12–19. Scottsdale, AR: American Compensation Association.

Guilford, J. P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

Hildebrand, M., Wilson, R. C., and Dienst, E. R. (1971). *Evaluating University Teaching.* Berkeley, CA: University of California Center for Research and Development in Higher Education.

Johnson, M., and Kasten, K. (1983). Meritorious work and faculty rewards: An empirical test of the relationship. *Research in Higher Education* 19: 49–71.

Kasten, K. (1984). Tenure and merit pay as rewards for research, teaching, and service at a research university. *Journal of Higher Education* 55: 500–514.

Katz, D. A. (1973). Faculty salaries, promotions, and productivity at a large university. *American Economic Review* 63: 469–477.

Keaveny, T. J., and Allen, R. E. (1983). The implications of an across-the-board salary increase. *Research in Higher Education* 19: 11–24.

Kulik, J. A., and McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of Research in Higher Education,* Vol. 3, pp. 210–240. Itasca, IL: Peacock.

Lee, B. A. (1985). Federal court involvement in academic personnel decisions. *Journal of Higher Education* 56: 38–54.

Machalak, S. J., Jr., and Friedrich, R. J. (1981). Research productivity and teaching effectiveness at a small liberal arts college. *Journal of Higher Education* 52: 578–597.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76: 707–754.

McKeachie, W. J. (1979). Financial incentives are ineffective for faculty. In Lewis, D. R. and Becker, W. E. (Eds.), *Academic Reward in Higher Education,* pp. 3–20. Cambridge, MA: Ballinger.

Miller, A. C., and Serzan, S. L. (1984). Criteria for identifying a refereed journal. *Journal of Higher Education* 55: 673–699.

Miller, D. A. (1978). Criteria for appointment, promotion, and retention of faculty in graduate social work programs. *Journal of Education for Social Work* 14(2): 74–81.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. *Research in Higher Education* 7: 193–205.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Ryan, J. J., Anderson, J. A., and Birchler, A. B. (1980). Student evaluations: The faculty responds. *Research in Higher Education* 12: 312–333.

Tuckman, H. P., Gapinski, J. H., and Hagemann, R. P. (1977). Faculty skills and salary structure in academe: A market perspective. *American Economic Review* 67: 692–702.

Received November 14, 1986