DOCUMENT RESUME

ED 290 395 HE 021 086

AUTHOR Cochran, Thomas R.; Gravely, Archer
TITLE Measurement Issues in Student Evaluations of Instruction and Their Impact on Decision Making.
PUB DATE 30 Oct 87
NOTE 19p.; Paper presented at the Annual Conference of the Southern Association for Institutional Research and the Society for College and University Planning (New Orleans, LA, October 28-30, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty; Decisionmaking; *Faculty Evaluation; Higher Education; Rating Scales; Semantic Differential; *Statistical Analysis; *Student Evaluation of Teacher Performance; Teacher Effectiveness
IDENTIFIERS *Mean (Statistics); *Median (Statistics); University of North Carolina Asheville

ABSTRACT

Methods were examined by which the results of student evaluations of instruction may be presented as one indicator of teaching effectiveness for faculty personnel committees in order to best answer the question of teaching competence. Specifically, the study examined two measures of central tendency, the median and the mean, to determine which method best addressed the question of competence versus incompetence, rather than the teaching style used. The data utilized were the student evaluations of instruction for the 1986-87 year at the University of North Carolina at Asheville. The evaluation data for each class were merged with the personnel system to capture faculty demographics. A total of 210 faculty were included. The results indicated that the median better addresses the questions of competence than does the mean, given the limitations of ordinal data combined with the highly negatively skewed distribution. Those faculty rated between 3.5 and 4.0 on a five-point scale often are subject to unwarranted scrutiny by personnel committees concerning their competence. If the median were used, those questions would not occur. Recommendations are made regarding (1) feedback procedures by department chairs and deans; (2) peer review systems; (3) collection of data on valued teaching activities; and (4) personnel procedures that weight teaching, research, and service. (KM)

MEASUREMENT ISSUES IN STUDENT EVALUATIONS OF INSTRUCTION AND THEIR IMPACT ON DECISION MAKING

Thomas R. Cochran and Archer Gravely

University of North Carolina at Asheville

Paper presented at the Annual Conference of the
Southern Association of Institutional Research
October 28-30, 1987
New Orleans, Louisiana

INTRODUCTION

Much has been written during the past fifteen years concerning the use and abuse of Student Evaluations of Instruction and the role they should play in decisions relative to faculty tenure and promotion (Scheck, 1978; Rodin, 1982; Cross, 1987; and Seldin, 1987 among others). Faculty governance groups have debated and re-written Student Evaluations of Instruction procedures as if it were an ongoing activity. While the jury is still out as to the validity of these instruments, their use in making personnel decisions of faculty occurs annually at most institutions. In fact, at a large number of institutions it is the only assessment device used to measure faculty teaching effectiveness. The purpose of this paper is to examine ways in which the use of Student Evaluations of Instruction can be improved to maximize their effectiveness as a decision making tool for assessing teaching performance. At teaching institutions such issues are at the very heart of the mission of the institution.

Why is the Student Evaluation of Instruction process so maligned? As a tool for decision-making, Student Evaluations of Instruction ratings are frequently used inappropriately. All colleges and universities have a strong need to rank faculty by different levels of teaching effectiveness and reward them

appropriately. The problem is that the student evaluation process is not capable of making distinctions among the competent. Student evaluations are only able to identify those faculty who are having problems in the classroom. This limitation is based on two factors. First, nearly all faculty are reasonably competent and the resulting scores are highly skewed to the positive. This finding is well documented in the evaluation literature and holds true for various types of evaluation instruments and scaling techniques. Second, the resulting data are ordinal rather than interval or ratio. Ordinal data convey a rank order relationship and the difference between a "5" and a "4" may not be the same as the difference between a "3" and a "4". As a result, the computation of a mean score is inappropriate.

These measurement distinctions are frequently ignored in social science research. When one is testing hypotheses using group measures, these measurement issues are not critical; but in the assessment of individual faculty for merit, tenure, and promotion decisions, the limitations of ordinal data must be respected. A common practice is for institutions to compute mean student evaluation of instruction scores to two decimal places. These data are then used by university administrators and faculty personnel committees to assess teaching

effectiveness. These data are frequently used to make fine distinctions among various levels of competence when in fact the data can only discriminate between the competent and incompetent.

A key question in the Student Evaluation of Instruction debate is that of competence. Rodin (1982) offered the concept of the Journeyman Principle as one way of focusing the issue. She noted that individuals interested in becoming practitioners of a skilled trade or profession (in this case college teaching) must undergo rigorous training. Those who complete the training are known as journeymen and are assumed competent to perform satisfactorily the range of tasks ordinarily required. They are not required to prove competence, since journeyman status itself attests to such competence. It is the judgement of incompetence that is made on the basis of special evidence. In relation to the present paper, the Journeyman Principle would suggest a radically different approach to the interpretation and use of student evaluation scores. The Journeyman Principle provides a good conceptual perspective for dealing with the limitation of student ratings. Backed by the Journeyman Principle, student evaluations can help to determine competence versus incompetence but not stylistic differences among the competent. In other words, not how does one teach, but rather the effectiveness of their teaching. In support of this viewpoint McBean and Lennox

(1982) suggest that the greatest value of student evaluations, from an administrative viewpoint, is the "flagging" of courses and/or professors that are in difficulty. However, Scheck (1982) takes a more extremist position when he asserts that it is immoral to use student evaluations for the administrative purposes of tenure and promotion of faculty. Yet, the fact remains that two-thirds of the four-year liberal arts colleges use student ratings of instruction in the evaluation of faculty (Cross, 1987).

One additional issue should be noted in evaluating a faculty member's teaching as part of the personnel process. Several articles have appeared recently (Seldin, 1987; Weimer, 1987) which stress the importance of multiple measures of teaching effectiveness. There seems to be a clear consensus that student evaluations of instruction used as the sole measure to teaching effectiveness are very inadequate. The need for additional methods of evaluating instruction seems to be the only sensible long range solution for properly assessing teaching. Since the student evaluation of instruction should remain as one of multiple data points in overall faculty evaluation, student evaluations of instruction must be used in an appropriate manner.

The present study examines methods by which the results of Student Evaluations of Instruction may be presented as one

indicator of teaching effectiveness for faculty personnel committees in order to best answer the question of teaching competence. This paper puts aside the question of using teaching evaluations for the purpose of improving instruction and concentrates on determining the appropriate role of student evaluations in personnel matters.

Specifically, this study examined two measures of central tendency, the median and the mean, to determine which method best addressed the question of competence versus incompetence. Typically, a Likert type scale is used for the student to rate the course and instructor. By using the median instead of the mean as the measure of central tendency it is the contention of this paper that personnel committees will be encouraged to judge whether or not the teaching is competent versus incompetent and not the teaching style used.

METHOD

The present study was based on the student evaluations of instruction completed during the 1986-87 year. The University of North Carolina at Asheville has a standardized evaluation form and procedures for every department. Completed evaluation forms are scored by the Office of Institutional Research with the

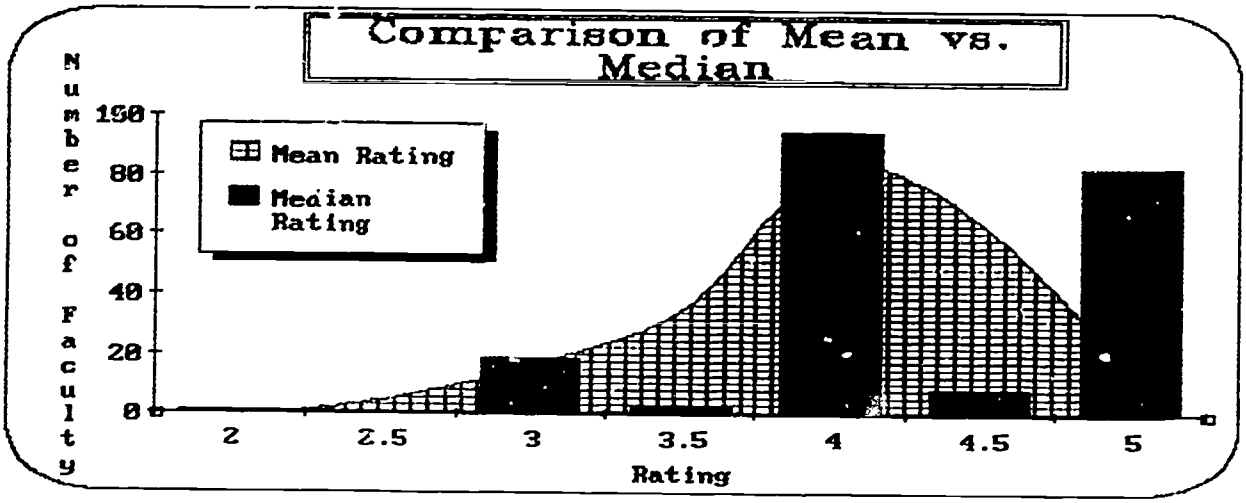
aid of an optical scanner. The evaluation data for each class were merged with the Personnel system to capture faculty demographic type information. The unit of analysis in the present study was the instructor rather than a particular course. Instructors with more than one class evaluation had the multiple class measures collapsed into a single set of instructor mean/medians for each item. A total of 210 faculty were included in the population. The analysis focused on comparing the mean versus the median distribution of evaluation scores. The results to follow were based on a five-point Semantic Differential scale designed to measure overall instructor effectiveness.

RESULTS

The primary comparison examined in the present paper was the use of the mean versus the median as a measure of central tendency in analyzing Student Evaluations of Instruction. Figure 1 below presents the comparison of the Mean versus Median rating for all the faculty during the 1986-87 Academic year. In the in the case of the median, the midpoint faculty score was 4 while for the mean, the average score was 4.19. The distribution of the mean versus median scores points out a positive difference in

the median rating for most faculty as compared to the mean rating of faculty.

Figure 1 -- A comparison of the Median versus the Mean in Overall Faculty Ratings on the Student Evaluation of Instruction.



Tables 1 and 2 examine these differences in greater detail. From Table 1 one can note that less than 11% of the faculty received overall median scores below 4. Most of the faculty were

Table 1 - Distribution of Overall Instructor Median Scores on the Student Evaluations of Instruction.

RANGE	FREQUENCY	CUM. FREQ.	PERCENT	CUM. %
2.00 -2.49	1	1	.48%	.48%
2.50 -2.99	0	0	0%	48%
3.00 -3.49	19	20	9.05%	9.52%
3.50 -3.99	3	23	1.43%	10.95%
4.00 -4.49	95	118	45.24%	56.19%
4.50 -4.99	9	127	4.29%	60.48%
5.00	83	210	39.52%	100.00%

rated at 4 (45%) or 5 (40%). In contrast, Table 2 presents the distribution of mean scores for faculty in which over 28% of the faculty received scores below 4.00. Nearly 40% of the faculty were rated between 4.00 and 4.49, while 32% of the faculty were rated between 4.50 and 5.00. Again the most noteworthy

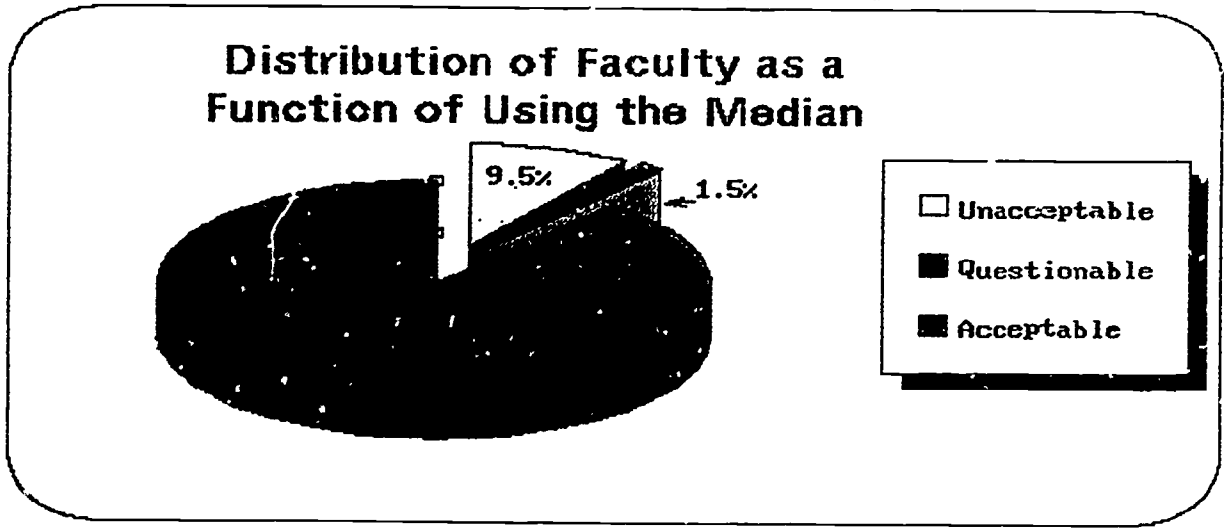
Table 2 - Distribution of Overall Instructor Mean Scores on the Student Evaluations of Instruction.

RANGE	FREQUENCY	CUM. FREQ.	PERCENT	CUM. %
2.00 -2.49	1	1	.48%	.48%
2.50 -2.99	5	6	2.38%	2.86%
3.00 -3.49	17	23	8.10%	10.95%
3.50 -3.99	36	59	17.14%	28.10%
4.00 -4.49	83	142	39.52%	67.62%
4.50 -4.99	63	205	30.00%	97 62%
5.00	5	210	2.38%	100.00%

difference was the percent of faculty rated below 4 on each of the two measures.

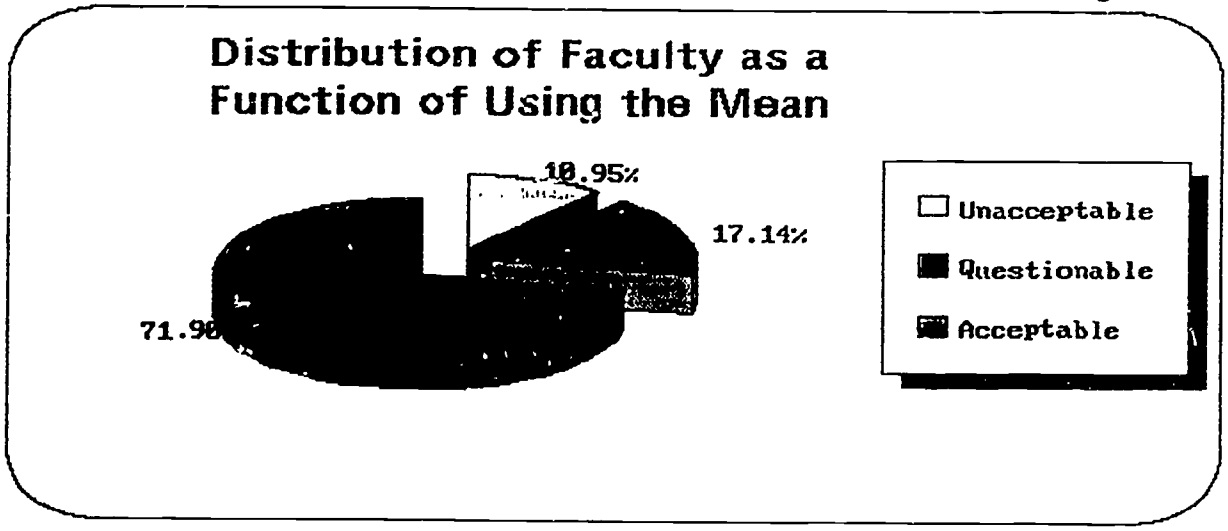
The significance of that difference can best be shown in Figures 2 and 3 below. Typically, universities will identify any one who below the average or midpoint of a distribution as being less than satisfactory. Anyone markedly below the average or

Figure 2- Distribution of Faculty as a Function of Using the Median as the measure of typicality.



midpoint is typically judged to be deficient and the performance considered unacceptable. Figures 2 and 3 attempt to graphically

Figure 3- Distribution of Faculty as a Function of Using the Mean

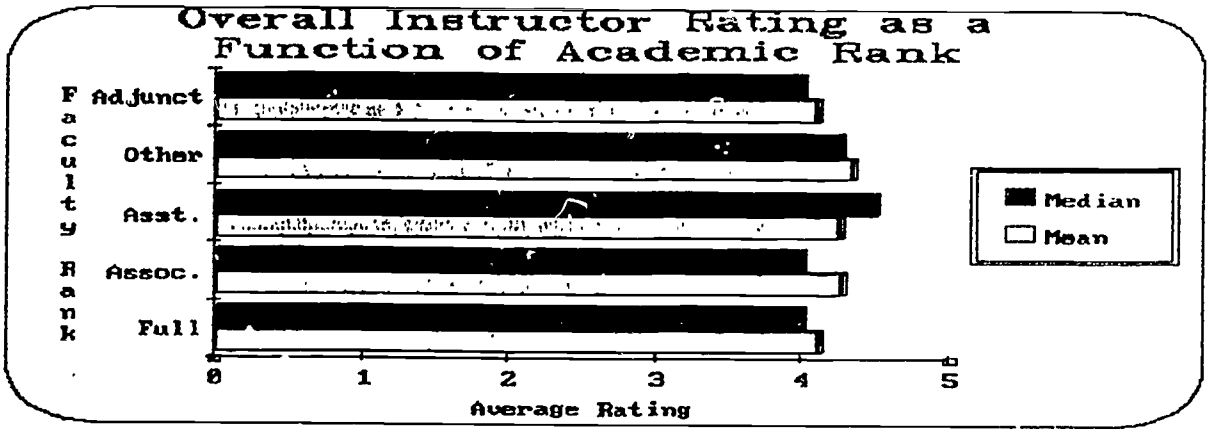


present that situation. Anyone with an overall score of 4 or

above was considered acceptable. Faculty rated between 3.50 and 4.00 were considered "questionable", while any faculty member with a score below 3.50 was judged to be deficient or "unacceptable". Examination of Figures 2 and 3 reveals that nearly the same percent of faculty were classified as unacceptable (9.52% for the Median vs 10.95% for the Mean). Most noteworthy, however, is the difference in the percent of faculty considered to be "questionable" when the mean and median are compared. Only 1% of the faculty are in question -- due primarily to the nature of the median as a measure of central tendency; while over 17% of the faculty are considered questionable when the mean is the measure of central tendency.

One additional area was examined relative to Student Evaluation of Instruction and the use of the mean versus the median -- that being tenure status and academic rank. Figure 4

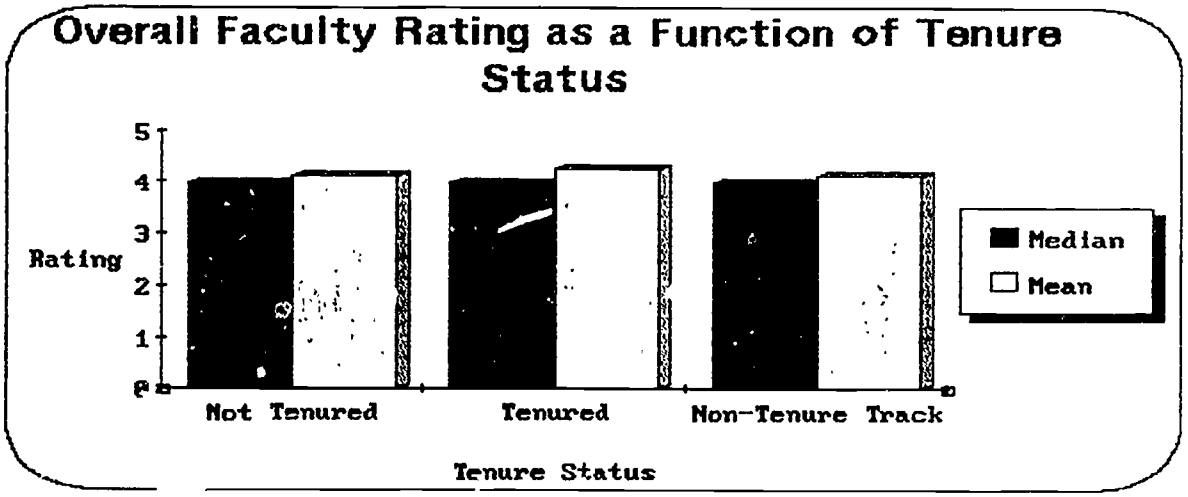
Figure 4- Overall Rating as a Function of Academic Rank



presents the overall instructor rating on the mean and median as

a function of a[illegible]c rank. Differences among ranks were negligible on both the mean and the median. Figure 5 presents the overall faculty rating as a function of tenure status.

Figure 5- Overall Faculty Rating as a Function of Tenure Status



Again, the differences reported for both the mean and the median were negligible.

DISCUSSION

The purpose of this paper was to link Rodin's (1982) Journeyman Principle with the limitation of Student Evaluation scores. The mean/median measurement issue was addressed by comparing the distribution about each measure of central tendency. The results indicate that the median better addresses the question of competence than does the mean given the limitations of ordinal data combined with the highly negatively skewed distribution. All too often, those faculty who are rated between 3.5 and 4.0 on a five point scale are subject to unwarranted scrutiny by personnel committees concerning their competence. If the median were used those questions would not occur. Only those faculty below 3.5 would be considered as unacceptable -- a range that is nearly identical for either the mean or median. The question of competence versus incompetence would be addressed clearly. The debate over making fine distinctions among the competent would have to be resolved with additional data.

These findings support a position taken by DeCanio (1986) in which he states, "...the mis-specification associated with use of the standard linear model when the dependent variable is limited and qualitative can lead to inefficient estimates (and) predictions outside the range of possible responses." The use of

the median does seem to accurately identify those faculty who are not performing up to standards as suggested by McBean and Lennox (1982) while avoiding the artificial precision of mean ratings of faculty with ratings slightly below average. However, the Student Evaluations of Instruction alone do not determine teaching effectiveness. What then can be done to assess teaching effectiveness in order to provide fairer, more equitable personnel decisions about faculty teaching performance on college and university campuses?

First, it continues to be very important to provide the consumer (in this case students) with the opportunity to offer feedback about their experiences in the classroom. In today's budget conscious, accountability-oriented environment, universities cannot afford to ignore the views and attitudes of the people they serve. The key to this feedback is to use it primarily as a means of providing feedback to the instructor and to encourage students to comment and make their views known.

RECOMMENDATION: Department chairs and college deans should be encouraged to establish procedures by which Student Evaluations of Instruction are used as a source of feedback about a faculty member's performance in the classroom and as an indicator of competence.

As has been suggested by Weimer (1987) and Seldin (1987) student evaluations should be only one of several sources of information about teaching effectiveness. Cahn (1987) strongly

supports the idea of peer evaluation programs as a much more effective method of evaluating faculty styles of competence.

RECOMMENDATION: EACH ACADEMIC DEPARTMENT SHOULD ESTABLISH A SYSTEM OF PEER REVIEW WHICH WOULD INCLUDE CLASSROOM VISITATION, CRITIQUE OF WRITTEN MATERIAL USED IN THE CLASS, AND DEPARTMENTAL SEMINARS DESIGNED TO SHARE INFORMATION ABOUT TEACHING TECHNIQUES.

A third activity which would aid in focussing attention on the importance of the role of teaching at a college or university would be to follow a procedure first suggested by Scheck (1978). The idea of "counts" to describe scholarly activities, i.e. number of presentations, books, publications, etc., is commonplace in higher education. While the debate wages on as to the quality of the scholarship, counts serve a vital role in assessing the scholarly accomplishments of a the faculty member. Scheck's idea was to apply the same system to teaching. Instead of relying solely on Student Evaluations of Instruction, he proposed that counts be made on such activities as; the number of course preparations, new course innovations, courses taught outside the discipline, levels of courses taught, etc. The "counting" system has a way of establishing the priorities of an institution by the types of activities which are counted.

RECOMMENDATION: DEPARTMENT CHAIRS AND ACADEMIC DEANS SHOULD ROUTINELY COLLECT DATA ON THOSE TEACHING ACTIVITIES VALUED BY THE INSTITUTION. THESE COUNTS SHOULD BE AS CONSISTENT AS POSSIBLE ACROSS DISCIPLINES.

Finally, the institution must reaffirm the relative importance of teaching in the university. The "Holy Trinity" of academic life: **Teaching, Research, and Service**; should be weighed by decision-makers in a manner that is consistent with the mission of the university. In today's computer age such weightings could even be quantitatively applied to data about a faculty member in order to provide a profile of the individual in the context of the particular institution. For example, an undergraduate liberal arts school should weight teaching more heavily than research or service. At some schools undergraduate research is stressed, thus research may be weighted higher than service but less than teaching. At major research universities the weights may be on research more than teaching. Other schools may have a particular mission which would suggest that service be given greater weight.

RECOMMENDATION: ACADEMIC LEADERS SHOULD CLEARLY ESTABLISH THE RELATIVE IMPORTANCE OF TEACHING, RESEARCH, AND SERVICE, AND STRIVE TO CREATE PERSONNEL PROCEDURES WHICH WOULD BE ABLE TO APPLY THESE WEIGHTS TO INDIVIDUAL FACULTY.

In conclusion, Student Evaluations of Instruction are a way of life on most college campuses. This paper has attempted to offer an approach to the use of these evaluations which would minimize the distortion and misrepresentation of a faculty

member's performance when personnel decisions are involved. The median as the preferred representation of a faculty member's "typical" performance is both statistically more correct and less subject to false conclusions. The median provides a natural link to Rodin's (1982) Journeyman Principle where the issue focuses on competence versus incompetence. Finally, the method by which faculty instruction is assessed in relation to personnel matters calls for a variety of data points and an institutionally determined value placed on teaching at that institution. Clearly, the use of only the Student Evaluation of Instruction as the index of teaching effectiveness by personnel committees should be discontinued. A multiple set of data points on teaching should be used in a manner consistent with the institution's mission to insure that faculty are being evaluated in a manner which is best for the faculty and school.

REFERENCES

Cahn, Steven M. Opinion: Faculty Members should be Evaluated by their Peers, not by their Students. The Chronicle of Higher Education. October 14, 1987, Vol. 34(7), B2-B3.

Cross, Patricia K. "Teaching for Learning". AAHE Bulletin, April 1987, vol. 39(1), 3-7.

DeCanio, Stephen J. Student Evaluations of Teaching-- A Multinominal Logit Approach. Research In Economic Education. Summer 1986, 165-76.

McBean, Edward A. and Lennox, William C. "Issues of Teaching Effectiveness as Observed Via Course Critiques". Higher Education, November 1982, 645-55.

Rodin, Miriam J. "By a Faculty Member's Yardstick, Student Evaluations Don't Measure Up". Teaching Political Science, Summer 1982, 174-76.

Scheck, Dennis C. "The Use and Abuse of Student Evaluations of Teaching Effectiveness in Higher Education". College Student Journal: Monograph, Fall 1978, vol. 12(3 part 2), 1-13.

Seldin, Peter. "Evaluating Teaching Performance: Answers to Common Questions". AAHE Bulletin, September 1987, vol.40(1), 9-12.

Weimer, Maryellen G. "Translating Evaluation Results into Teaching Improvements". AAHE Bulletin, April 1987, vol. 39(1), 8-11.