SANGAMON STATE UNIVERSITY

Fall Faculty Evaluation

Technical Paper No. 2

"Student Evaluation of Faculty Teaching Performance: Analysis of Four Years' Data"

by Professor Jerry A. Colliver
and Robert M. Wesley

Sponsored by the Division of Academic Affairs
Office of the Vice President

Sangamon State University

Springfield, Illinois 62708 MEMORANDUM

TO: The University Community and Its Observers

FROM: John H. Keiser, Vice President for Academic Affairs

DATE: February 11, 1976

SUBJECT: STUDENT EVALUATION OF FACULTY TEACHING PERFORMANCE

Several years ago the Division of Academic Affairs published a paper entitled "A Report on Student Evaluation of Faculty Teaching Performance at Sangamon State University" written by Professor Jerry A. Colliver.

The attached report by Professor Colliver and Robert M. Wesley is a follow-up on the same subject entitled "Student Evaluation of Faculty Teaching Performance: Analysis of Four Years' Data." While the sponsorship does not necessarily imply total agreement with the details of the paper, it does indicate an endorsement of its quality and usefulness. Certainly, the subject is one of interest to all elements of the university community.

JHK:sjk
enc.

STUDENT EVALUATION OF FACULTY TEACHING PERFORMANCE: ANALYSIS OF FOUR YEARS' DATA

Jerry A. Colliver and Robert M. Wesley

January 20, 1976

The authors want to acknowledge the invaluable assistance and advice of Mr. William Gorrell, the Director of Institutional Research. Mr. Gorrell gave generously of his time and resources to the completion of the project. The authors also want to thank Mr. Terry Powell, a member of the staff of the Office of Institutional Research, for his technical assistance.

STUDENT EVALUATION OF FACULTY TEACHING PERFORMANCE: ANALYSIS OF FOUR YEARS' DATA

Jerry A. Colliver and Robert M. Wesley

January 20, 1976

INTRODUCTION

Student evaluation of faculty teaching performance has provided an important source of evidence in making faculty personnel decisions at Sangamon State University. Students have been asked to rate the competency and the teaching ability of their instructors; specifically, near the end of each term, students have rated faculty using the following two items:

(1) Do you think this teacher is competent in the content or matter offered in this course?

exceptionally competent		satisfactory		incompetent
5	4	3	2	1

(2) Overall, do you consider this person a good teacher?

excellent		good		poor
5	4	3	2	1

The ratings on these two items have provided a major source of input in making decisions concerning salary, promotion, tenure, and retention over a four year period from the 1971-1972 academic year to the 1974-1975 academic year.

The history and rationale underlying the development of the two item evaluation process are presented in Technical Paper No. 1 (Colliver, 1972). Also reported there are the results of an extensive analysis of the data obtained during the 1971-1972 academic year using the two item evaluation process. In

general, the analysis of that data was concerned with (1) the reliability of the two item evaluation process, (2) the relationship between evaluations obtained using the two item procedure and evaluations obtained using other evaluation procedures, and (3) the effects on the evaluations of certain extraneous variables such as class size, structure, faculty characteristics, etc.

The two items were administered for the three subsequent academic years, 1972-73 through 1974-75, and a record of the resultant data has been maintained. The research reported in this paper was motivated by a desire to look at the stability (reliability) of the evaluations obtained from the two item evaluation process over a four year period. In addition, the present paper reports on the amount of agreement between the ratings given on the teaching item and the ratings on the competency item. Finally, the effect of rewording the teaching item on the ratings is also reported.

RELIABILITY

Review of Reliability as Assessed for 1971-72 Data

The reliability of the two item evaluation process was first assessed using the data obtained for the 1971-72 academic year and is reported in Technical Paper No. 1. As reported there, reliability coefficients were computed which indicated (1) the amount of agreement among the mean evaluations for different classes and (2) the amount of agreement between the mean evaluations for the winter and spring quarters. First, a single mean evaluation was available for each class which was the mean of the ratings given by all students in that class on the competency and the teaching items combined. Since most faculty members taught two classes each quarter, one winter quarter class was arbitrarily chosen for each faculty member and called class 1; the remaining class was called class 2. Similarly, one spring quarter class was arbitrarily chosen and called class 3 while the remaining class was called class 4. The correlations (reliability coefficients) were computed between all possible pairings of the four class means. The correlations and the number of pairs of classes each correlation was based upon are presented below where the subscripts of r indicate the two classes that were correlated:

$r_{1,2} = .55$, $n = 55$;

$r_{1,3} = .59$, $n = 75$;

$r_{1,4} = .65$, $n = 49$;

$r_{2,3} = .36$, $n = 55$;

$r_{2,4} = .53$, $n = 42$;

$r_{3,4} = .49$, $n = 49$.

All of the correlations were significant at the .01 level. The reliability of the quarter means was determined by correlating the winter quarter means and the spring quarter means. The correlation between the quarter means was found to be $r = .62$ ($n = 75$; $p < .01$).

The reliabilities of the class means were generally above .50; the average reliability was computed using the Fisher's z transformation method for averaging weighted correlations and was found to be $r_{ave} = .60$. The reliability of the quarter mean was .62. It was felt that the reliability of the quarter mean was generally higher than the reliabilities of the class means because quarter means were based on more information. Since personnel decisions were based to a considerable extent upon faculty members' grand means which were means based on the data for all classes in both quarters for each faculty member, the reliability of the grand mean was expected to be even greater than that of class or quarter means because the grand means were based on even more information than either of these means. Consequently, the Spearman Brown formula for determining the reliability of a test of doubled length was applied to the reliability of the quarter mean ($r = .62$) to project the reliability of the grand mean. The projected reliability of a grand mean based upon two quarters of evaluation data was found to be $r = .77$. Finally, since the grand means would be based upon evaluations from all courses in three quarters rather than two if this evaluation procedure were to be used in subsequent years, the reliability of a grand mean based upon data from three quarters was projected using the Spearman Brown formula and found to be $r = .83$.

Reliability of Grand Means across 1971-72 and 1972-73 Academic Years

The two item evaluation process has been used for the three academic years subsequent to the 1971-72 academic year. Consequently, at the outset of this research, it was thought that it would be possible to directly assess the reliability of the two item evaluation process over a four year period. This would have provided an opportunity to directly assess the reliability of the

grand mean by looking at the amount of agreement among the grand means for the four academic years and, although of lesser importance, it would have also provided a check on the accuracy of the projected reliabilities obtained from the use of the Spearman Brown technique. However, complete data are available for only two academic years, 1971-72 and 1972-73. Evaluation was mandatory in 1971-72 and 1972-73 and data are available from virtually all classes at the University for those two academic years. Unfortunately, the evaluation data for the two other academic years, 1973-74 and 1974-75, was not obtained systematically and, as a result, is incomplete and probably biased making it difficult, if not impossible, to meaningfully interpret the results of the reliability analysis of this data. The reasons for the incomplete data for these two academic years and the ambiguous results of the reliability analysis of this questionable data are presented in the following section. Consequently, the primary evidence bearing on the reliability of the grand means between years is based on the data obtained for the 1971-72 and the 1972-73 academic years.

The reliability of the grand mean was assessed between years two different ways. (1) First, the reliability was assessed using the grand means of the 75 faculty members for whom data were available for <u>both</u> the 1971-72 and the 1972-73 academic years. The reliability was determined by computing the correlation between the grand means for the two academic years. The reliability coefficient was found to be $r = .61$ ($n = 75$; $p < .01$). (2) In addition, the reliability of the between year grand means was determined for the sample of 75 faculty for whom the reliability of the quarter means for the 1971-72 academic year was computed as reported in Technical Paper No. 1 and reviewed in the preceding section. Of these 75 faculty members, data was not available for 6 of these faculty for the

1972-73 academic year. Consequently, the reliability of the grand means was also determined by computing the correlation between the 1971-72 and the 1972-73 grand means for the remaining sample of n = 69 faculty members. This reliability coefficient was found to be r = .63 (n = 69; p < .01).

Perhaps an explanation is in order of why the sample sizes in the two preceding reliability analyses are not the same (n = 75 and n = 69, respectively). In the latter analysis, 75 faculty members had data available for both quarters of the 1971-72 academic year but only 69 of those faculty also had data available for the 1972-73 academic year. On the other hand, in the former analysis, it was reported that 75 faculty had data available for both the 1971-72 and the 1972-73 academic years. However, there was no stipulation that these faculty had to have data for both quarters of 1971-72. Consequently, there were some faculty who had data for one quarter of 1971-72 but not both quarters although they did have data for both academic years. This resulted in the differences in sample sizes for these two analyses.

It should be noted that the between year reliability of the grand mean did not even approach the magnitude of the reliabilities as projected by the Spearman Brown formula. Using the Spearman Brown formula, the reliability of a grand mean based on data for two quarters was projected to be .77 and the reliability of a grand mean based on data for three quarters was projected to be .83. It was anticipated that the reliability of the grand mean computed directly by correlating the 1971-72 grand means with the 1972-73 grand means would fall somewhere between these projected reliabilities since the 1971-72 grand means were based on two quarters' data while the 1972-73 grand means were based on three quarters' data. This did <u>not</u> occur. In fact, the between year grand mean reliabilities computed on the two samples are r = .63 and r = .61 and are identical with the quarter mean reliability of r = .62 although it was

predicted that the between year reliability of the grand mean would represent some increase over the between quarter reliability because the grand means were based on more data than the quarter means. It is suggested that the additional stability contributed by more data for the grand means was offset by stability in the quarter means due to course sequences taught by a given faculty member and possibly due to a tendency of students within a given year to continue course work with a given faculty member. In test theory terms, it could be said that the within year factors contributed as much to the true score variance of the quarter means as the additional data for the entire year, acting analogous to a lengthened test, contributed to the true score variance of the grand means. At any rate, it is interesting and important to note that the grand mean reliability across years was not accurately predicted by the Spearman Brown projections; in fact, there was no indication of increased reliability with additional data. In conclusion, the reliabilities of the quarter means and the grand means consistently indicate that there is between 60 and 65% true score variance in the faculty evaluation summaries.

Reliability of Grand Means over Four Year Period

The initial plan for this research was to determine the reliability of the grand means obtained with the two item evaluation process across the four academic years from 1971-72 through 1974-75. Unfortunately, as indicated above, complete data was not available for the 1973-74 and 1974-75 academic years. Due to a severe snow storm during the last week of the semester in December of 1973, classes were dismissed and many classes were not evaluated. In addition, University policy regarding evaluation was changed that year making it possible for faculty to use alternative evaluation procedures. The snow storm seemed to set a precedent of not using the two item evaluation procedure which was legitimized by University policy. As a result, for example, the number of faculty

using the two item evaluation form dropped sharply from 208 in 1973-74 to 124 in 1974-75. This decline occurred in spite of the fact that the size of the faculty increased slightly over those years. The problem is compounded by the fact that evaluations are available for only some of the classes of those faculty who did use the two item evaluation process. Thus, faculty who used the two item evaluation process were self selected and within this self selected sample, faculty selected the classes they would evaluate with the two items. These two sources of bias make the data difficult if not impossible to meaningfully interpret.

In spite of this bias and the problem of meaningfully interpreting the reliabilities obtained from this data, the results of the analysis of the data for the four years are presented here since it was felt that they should be reported and should become a part of the record of the evaluation process. Since grand means were not available for all faculty across all four years, the reliability of the grand mean was assessed three different ways. (1) Initially, all possible pairings of the grand means for the four academic years were formed for each faculty member. For example, for a given faculty member, the 1971-72 grand mean was paired with the 1972-73 grand mean, the 1971-72 grand mean was paired with the 1973-74 grand mean, etc. Of course, if data were not available for a given year, the pairings involving that grand mean would not be possible. The reliability coefficients were computed by correlating the pairs of grand means and are reported in Table 1. The number in parentheses to the right of each correlation is the number of pairs of grand means upon which the correlation was based. (2) In addition, the reliability of the grand mean was assessed by looking at the data available over the four year period for the 75 faculty for whom the reliability of the quarter mean was assessed on the 1971-72 data. A grand mean was computed for each year that data was available for each of these faculty members and the correlations between the grand means for pairs of years were computed as described above. The reliability coefficients are reported in Table 2 and the number of cases each

	1971-72	1972-73	1973-74	1974-75
1971-72				
1972-73	.61 (75)			
1973-74	.51 (62)	.45 (130)		
1974-75	.44 (38)	.42 (76)	.46 (101)	

Table 1

	1971-72	1972-73	1973-74	1974-75
1971-72				
1972-73	.63 (69)			
1973-74	.53 (57)	.47 (57)		
1974-75	.56 (34)	.54 (34)	.35 (34)	

Table 2

	1971-72	1972-73	1973-74	1974-75
1971-72				
1972-73	.63 (34)			
1973-74	.56 (34)	.59 (34)		
1974-75	.56 (34)	.54 (34)	.35 (34)	

Table 3

correlation was based upon is in parentheses to the right of the correlation. (3) Finally, data was available across all four academic years for only 34 faculty. Grand means were computed for all four years for each of these 34 faculty. The correlations for all possible pairs of these grand means are presented in Table 3. Each of these reliability coefficients are based upon 34 cases.

It should be emphasized that the reliability coefficents in tables 1, 2, and 3 are based upon biased data and that the bias most certainly operates so as to lower the reliabilities. In spite of this, it should be noted that the correlations in all three tables are significant at the .05 level.

RELATIONSHIP BETWEEN COMPETENCY RATINGS AND TEACHING RATINGS

All of the means - class means, quarter means, and grand means - considered in this paper and in Technical Paper No. 1 were obtained by pooling the ratings on the competency item and the teaching item. It is of interest to determine if there is a relationship between the ratings faculty received on these two items. This was accomplished by correlating the means of the competency ratings and the means of the teaching ratings.[1]

Specifically, a competency mean was obtained for each faculty member by computing the mean of the ratings given on the competency item by all students in all of the classes for the faculty member in the 1971-72 academic year. Similarly, a teaching mean was obtained for each faculty member. A competency mean and a teaching mean were also obtained for each faculty member for the 1972-73 academic year. Due to the biases in the 1973-74 and 1974-75 data discussed above, correlations were computed only for the 1971-72 and 1972-73 data.

The correlation between the competency and the teaching means for the 1971-72 academic year was found to be $r = .80$ ($n = 112$; $p < .01$) and the correlation for 1972-73 was $r = .79$ ($n = 179$; $p < .01$). Notice that the correlations for the two academic years are virtually identical.

The correlations might have been even higher except that there appeared to be a "ceiling effect" on the competency item ratings. The values in the table below support the notion of a ceiling effect in that it can be seen

---

[1] The nature of the relationship between the ratings on the competency and the teaching items should have been studied at the time the research was conducted for Technical Paper No. 1. Due to a number of practical problems (such as time pressure to complete the personnel decisions, lack of familiarity with the University's new computer facility, etc.), it was not possible to perform this analysis at that time.

that the University means on the competency item for both the 1971-72 and the 1972-73 academic years are quite high and that the variances and standard deviations are small. The restriction of variability would result in a diminished correlation of the competency item with any other variable (McNemar, 1969). Also consistent with the notion of a ceiling effect is the fact that a common objection to the competency item made by both students and faculty was that students are not qualified to judge the competency of a faculty member in "the content or matter offered in the course." Students were especially critical of this item and many indicated that they rated all faculty "5" on this item because they felt uncomfortable making a judgment outside what they felt to be their legitimate domain of evaluation.

1971-72	Mean	Variance	Standard Deviation
Competency	4.63	.08	.28
Teaching	4.40	.14	.37
1972-73			
Competency	4.51	.09	.30
Teaching	4.09	.19	.44

EFFECT OF REWORDING TEACHING ITEM ON TEACHING RATINGS

In the fall of 1972 some members of the University Evaluation Committee suggested changes in the wording of the teaching item. On the one hand, it was argued that since Sangamon State is primarily a teaching institution, most of the faculty were attracted to Sangamon State because of their interest in and concern for good teaching which suggests that in general they are probably somewhat superior teachers. As a result, if students compared Sangamon State faculty with faculty they had known at other institutions, it is reasonable to assume that the Sangamon State faculty will generally receive high ratings resulting in little differentiation among the faculty. In order to avoid this, it was suggested that the teaching item be reworded so that faculty would be rated only relative to other faculty at Sangamon State. In addition, several faculty on the evaluation committee objected to the wording of the teaching item which asked students to rate faculty along a 5-point scale from excellent (5) through good (3) to poor (1) because they felt this required students to make a value judgment. These faculty suggested that students be asked only to rate faculty relative to other faculty rather than to judge faculty as excellent, good, or poor.

In light of these two suggestions, the teaching item was reworded to read as follows:

What is your overall evaluation of this individual as an instructor (at SSU)?

among the best	better than average	average	poor	among the worst
5	4	3	2	1

For the sake of comparison, the original item is again presented here:

Overall, do you consider this person a good teacher?

excellent		good		poor
5	4	3	2	1

Both of these items appeared on the evaluation form administered during the 1972-73 academic year and students in all classes were asked to rate the faculty on both of these items.

Thus, it was possible to compute two teaching means for each faculty member using the 1972-73 data: one was the mean of the ratings given by all students in all the classes of a given faculty member on the original teaching item and the other was the mean of all the ratings received by a faculty member on the reworded teaching item. The effect of rewording the teaching item was assessed by looking at the agreement between the ratings given on the two items. This was determined by correlating the means on the original item and the means on the reworded item. The correlation was found to be $r = .97$ ($n = 179$; $p < .01$). A correlation of this magnitude indicates that the two teaching items are measuring exactly the same thing.

In addition, the means and standard deviations of the original and the reworded teaching items for all 179 faculty combined were found to be:

	original item	reworded item
mean:	4.18	4.09
standard deviation:	.42	.44

As can be seen the means and the standard deviations of the two items were virtually identical. Thus, in light of the correlation between the original and the reworded items and a comparison of the means and standard deviations of the items, it seems reasonable to conclude that the rewording had absolutely no effect on students' ratings of the faculty.

SUMMARY

There was moderate agreement within and between years on the mean ratings on the competency and teaching items combined. It had been previously reported in Technical Paper No. 1 that the reliability of a quarter mean was r = .62. In the present paper it is reported that the reliability of a grand mean (mean of an entire year's data) was virtually identical to that of a quarter mean (r = .61 and r = .63). The projected reliabilities of a grand mean predicted from the reliability of a quarter mean using the Spearman Brown formula were not supported by the data.

In addition, there was considerable agreement among mean ratings received by each faculty member on the several items asking about faculty performance. There was a correlation of r = .80 between the mean ratings on the competency and the teaching items. Also, the agreement between the mean ratings on the original teaching item and a reworded version of that item was nearly perfect; r = .97.

REFERENCES

Colliver, Jerry A. A report on student evaluation of faculty teaching performance at Sangamon State University. Technical Paper No. 1, Sangamon State University, 1972.

McNemar, Quinn Psychological statistics. (4th ed.) New York: John Wiley & Sons, Inc., 1969.