A report is presented on an effort to use the statistical technique of factor analysis in evaluating the curriculum of a school of public health. Results indicate that the quality of teacher performance can be evaluated independently of the "popularity" of the teacher.

STUDENT EVALUATION OF A PUBLIC HEALTH CURRICULUM

Bernard Caffrey, Ph.D., and Glen E. Kost

MANY teachers routinely evaluate their courses to obtain feedback from their students. The value of such a practice has been demonstrated by Tuckman and Oliver,[1] who reported that student evaluations were related to improvement in teacher performance, while evaluation by superiors brought about no improvement. Warren[2] questioned the validity of such evaluations, calling them "a teacher popularity poll among the students." Others, however, feel that such a practice can serve a very useful purpose.[3,4] It is not unusual for medical students and students in schools of public health to end the academic year with the presentation of some citation for outstanding teachers and, in some cases, criticism of poor teachers and courses. Such evaluations are seldom objective and dispassionate, and may describe the "popularity poll" of which Warren wrote.

Isaacson, et al.,[5] discussed a scale used to evaluate teacher performance. Application of the statistical technique of factor analysis showed that the scale measured relatively independent dimensions of teacher performance. The fact that student evaluation of teacher performance must be considered from a multidimensional point of view might help to explain some of the controversy concerning such evaluations. If, for example, a teacher were rated on a number of items, a single summary score might actually be a composite of many facets of teacher performance. The ideal would be to identify these different facets, evaluate the teacher relative to them, and obtain scores on each of these facets or dimensions.

Caffrey[6] used the Isaacson scale to assess the possible sources of bias in evaluating teachers. He found that evaluation of teachers was free from bias due to personal characteristics of the teacher (friendliness, sarcasm, and the like) or the student (sex, grade in the class, grade point average). A factor analysis of the data revealed that independent dimensions were involved in the evaluation. The dimension termed "Teaching Ability" was the primary factor derived, followed by "Feedback to Students," "Overload of Students," and "Structure of the Class." There was very little correlation among these factors. Such a technique of measuring the different dimensions of teacher performance helps to increase the reliability and validity of such evaluations.

In the spring of 1969 the faculty curriculum committee of the Tulane University School of Public Health proposed that an evaluation of the courses be carried out. The president of the student body was appointed to direct the proj-

ect, which received the approval of the dean of the school. An initial evaluation attempt was made, but the rate of response from the student body did not reach the 50 per cent level. Although the results were summarized and presented to the curriculum committee, questions of objectivity and validity of the survey led to plans for a more structured and objective type of evaluation. Since this evaluation may serve as a model for similar evaluations in other schools, it will be discussed in detail here.

Method

Consideration of a number of scales led to the selection of items from the Isaacson, et al.,[5] scale, which was modified in accordance with the research of Caffrey.[6] The scale was to be brief (20 items), to elicit the maximum response rate from the students. Thus items related to only three factors were selected: *teaching ability* (12 items); *student-teacher interaction* or rapport (3 items); and *overload or structure* (3 items) from the Isaacson-Caffrey scale. Two items were added to assess the adequacy of library and laboratory facilities. The meaning of these factor names may be better understood by an inspection of Table 1, which lists the items that were selected to measure these dimensions.

The first item of the scale referred to a general evaluation of the all-around teaching ability of the teacher, and the last item (20) referred to the over-all value of the course. Responses to these two items were scaled on seven-point scales from "1=very poor," to "7=outstanding, excellent." Responses to the other 18 items were scaled on a five-point scale from "1=strongly disagree," to "5=strongly agree." The form concluded with an invitation to "'please feel free to make any other comments or suggestions" on the reverse side of the form.

The evaluation form was given to the students in two phases. The first phase, given three weeks before final tests were to begin, concerned the courses taught during the first semester; the second, given eight days later, concerned the second semester courses.

Table 1—Items related to four factors selected for assessment in a school of public health curriculum evaluation

Factor I. Teaching ability:
1. How would you rate the instructor in general (all around) teaching ability?
2. Class time was well spent.
3. The instructor was skillful in observing student performance.
5. He put the material across in an interesting way.
8. He tried to increase the interest of class members in his subject.
11. He made it clear how each topic fit into the course.
13. He explained why he did things.
15. He stimulated the intellectual curiosity of his students.
17. He changed his approach to meet new situations.
18. He was aware of it when students failed to follow him in class.
19. He explained clearly and his explanations were to the point.
20. How would you rate the over-all value of this course?

Factor II. Student-teacher interaction:
6. He listened attentively to what class members had to say.
10. The students argued with one another or with the instructor, not necessarily with hostility.
12. The students frequently volunteered their own opinions.

Factor III. Overload or structure:
4. The instructor assigned very difficult reading.
7. He followed an outline closely.
14. He planned the activities of each class period in detail.

Factor IV. Adequacy of facilities:
9. Library resources were adequate for students in this class.
16. Equipment and laboratory facilities in the class were adequate.

Students were advised to allow some time between their evaluation of individual courses, so that the evaluations would be relatively independent of each other. Students were requested to refrain from signing the forms, and they were advised that none of the teachers in question would see their actual forms but would receive the results of their evaluation. To encourage maximum participation the students were assured that they would receive the results of the evaluation, and that the results would also be made available to the administration and the teachers themselves.

When the forms were given to the students for the second phase, they were accompanied by a cover page with the statement: "This questionnaire is a reliable measure of student opinion regarding the curriculum." This was to be answered on the five-point scale (1= strongly disagree to 5=strongly agree) used on the evaluation form. Since the students had already completed the forms for the first semester, this item was designed to assess their general attitudes toward the form.

The data were analyzed using the IBM 7044 of the Tulane University computing center. To facilitate rapid processing, the information on the forms was immediately transferred to data-processing cards. Each course and teacher was given a code number. The course and teacher to whom these code numbers referred were known only to the class president. The data were analyzed "blindly," i.e., with no knowledge of which course or which teacher was being evaluated. Since some courses had multiple teachers, and some teachers had multiple courses, this permitted an evaluation of the curriculum by courses and by teachers. In cases where a single course had more than one teacher, the student was requested to attempt to identify *one* teacher with the course, and to evaluate the course relative to that teacher.

All of the data were first entered into a principal components factor analysis followed by a varimax rotation. This statistical technique uses the correlations among items as a basis for defining the dimensions (factors) which can be used to explain the responses to all of the items. If there were no correlations among item responses there would be as many factors as there are items. In that case, however, the factor analysis would be useless and most of the variance among the items would be attributable to "error" variance. The value of factor analysis is to reduce a complex array of correlations to its simplest dimensions, while at the same time accounting for a relatively high proportion of the variance among the items. In this way, the researcher can evaluate the information statistically to determine whether or not he is mixing "apples and lemons" in obtaining summary scores for a group of items. The computer program used for the factor analysis[7] produces factor scores for individuals. These factor scores are in standard score format with a mean of 0.00 and a standard deviation of 1.00. The use of standard scores permitted an instant statistical evaluation of the relative position of an individual course on the factor. Thus, if a course received a mean rating of +0.90 on Factor I, this is instantly seen as 0.9 standard deviations above the mean. Reference to the tables for the normal curve found in any standard statistics book shows that the mean rating for that course was above 81.6 per cent of the other courses.

The data cards were grouped according to courses, and means, standard deviations, and standard errors of the mean calculated for the 20 items and the two factor scores. The means and standard deviations were also calculated for all 973 forms. These served as the "population" values if a teacher wished to compare responses to items for a particular course with those for all other courses.

In order to evaluate the individual

items of the rating form, frequency counts were also obtained for all the items. This was done for all the forms taken as a unit, and for each of the individual courses. In this way the frequency of a given response for a particular item could be seen relative to the "population" responses for that item. Although the standard deviations for individual items reflect the degree of variability in the responses, a consideration of the actual frequency counts has some value for descriptive purposes.

Each item of the rating form was evaluated in three ways: (1) by its contribution to the factor scores; (2) by its mean and standard deviation; and (3) by a frequency count of responses to that item. In this way the reliability of the form could be estimated by judging the reliability of individual items. In order to assess the validity of the factor scores, two members of the student committee read the comments made by the students concerning 10 core courses. These comments were evaluated as negative or positive, and a frequency count obtained for them. Chi-square tables were constructed to evaluate these comments relative to the obtained factor scores.

Results

Of the 90 full-time students registered for the first semester, 81 (90%) returned their completed forms. Of the 95 registered for the second semester, 78 (82.1%) returned them. The overall response rate (159/185) was 85.9 per cent. On the basis of the high-response rate, little bias due to characteristics of nonrespondents should be found in the evaluation. A total of 973 forms was returned.

In the second phase of the evaluation, the students were asked to rate the form itself. In reply to the statement: "This questionnaire is a reliable measure of student opinion regarding the curriculum," the students answered as shown in Table 2. The fact that 61/78 (78.2%) returned the cover sheet helps support the conclusions to be drawn from Table 1. Only 3 per cent (2) of the students felt that the scale was not reliable (the statement should have said "valid," as that was the sense intended). Although 44.3 per cent replied that they were "not sure," their failure to disagree indicates a readiness to let the final results speak for them.

The varimax rotation of a principal components factor analysis for the 973 forms revealed that the 20 items could be accounted for by two main factors and a third (weak) factor. The correlations (loadings) of the 20 items with the two principal factors are presented in Table 3. Factor I can best be described as an *ability and proficiency* factor, with high correlations among items that were chosen to measure that dimension. Factor II had high correlations with the three items chosen to measure the dimension *student-teacher interaction* or rapport. The correlation between the two factor scores computed for the 973 completed forms was 0.12, indicating that the two factors were relatively independent. Thus the two factors measure essentially distinct dimensions of teacher performance.

The expected third factor (*overload or*

Table 2—Frequency and percentage of responses to the statement: "This questionnaire is a reliable measure of student opinion regarding the curriculum"

Response	Frequency	Percentage
1. Strongly disagree	1	1.6
2. Disagree	1	1.6
3. Don't know	27	44.3
4. Agree	19	31.1
5. Strongly agree	13	21.3
Total	61	99.9

Table 3—Factor correlations, means, and communalities (h^2) of items used to evaluate courses and teachers

Item:	Factor I	Factor II	h^2	Mean
1. General teaching ability	0.73*	0.38	0.67	4.54
2. Class time well spent	0.74*	0.34	0.66	3.63
3. Observed student performance	0.66*	0.52*	0.71	3.32
4. Assigned difficult reading	...	...	0.03	2.36
5. Put matter across well	0.70*	0.43*	0.67	3.37
6. Listened attentively	0.46*	0.63*	0.62	3.76
7. Followed an outline closely	0.71*	...	0.54	3.56
8. Increased class interest	0.65*	0.48*	0.65	3.69
9. Library resources adequate	0.32	...	0.12	3.53
10. Students argued	...	0.45*	0.20	3.13
11. How topics fit in	0.81*	...	0.68	3.39
12. Students gave opinions	...	0.66*	0.47	3.72
13. Explained why he did things	0.69*	0.41*	0.64	3.50
14. Planned class in detail	0.78*	...	0.63	3.62
15. Stimulated curiosity	0.70*	0.50*	0.73	3.43
16. Laboratory facilities adequate	0.36	...	0.14	3.29
17. Changed his approach	0.61*	0.56*	0.68	3.29
18. Aware if class not following	0.63*	0.53*	0.68	3.25
19. Explained to the point	0.78*	...	0.68	3.52
20. Over-all value of the course	0.79*	0.39	0.78	4.50
% of variance accounted for	48.9	6.0	54.9	

* Items which define a particular factor. Loadings <0.30 not included.

structure) was not found. Item 4, related to this factor, did not correlate with any of the factors. The other two items, 7 and 14, were related to Factor I. It is interesting to note that these two items alone correlated negatively with Factor II. If a class was too structured, there was less chance for *student-teacher interaction.* The expected fourth factor was related to items 9 and 16, and described a dimension that encompassed *adequacy of facilities.* Although these two items were related to the next factor in order, it accounted for so little of the total variance that it was not analyzed. Perhaps these two items were not relevant for most of the classes, and thus there was no consistent pattern of replies to them.

The column of Table 3 titled h^2 gives the communality for each item. This figure indicates the degree to which an item belongs with the other items. The means for the individual items are also shown in Table 3. Not many of the students felt that the teachers *assigned very difficult reading* (item 4). Except for item 4, all of the means were above 3.13. (Note that items 1 and 20 were scaled on a seven-point scale.) The mean of the means (not including items 1, 4, and 20) was 3.47, showing a tendency for the students to affirm the items. This indicates a generally positive evaluation of the curriculum, since the items were scaled so that high scores meant a positive response.

The means, standard deviations, and

standard errors of the mean were calculated for all of the questionnaire items and factor scores (in standard score form) for each course listed on the curriculum for the two semesters. Frequency counts of the responses to each item were also computed for each course. These were to be mailed to each individual teacher before the end of the academic year. There were 84 courses which were evaluated by more than one student. These data were given to the curriculum committee and made available to all of the faculty members. A written explanation of how to interpret the data was also provided. Since evaluation of *courses* was the main object of the program, these statistics were computed only for courses, and not for the individual teachers. Since most courses were taught by a single teacher, he was able to evaluate his performance in the courses that he taught.

After the data were obtained, it appeared desirable to assess the validity of the procedure. The first question to be answered was: "What was the relation between the size of the class and the evaluations?" On an *a priori* basis the validity of the evaluation would be supported if the scores on Factor I (*ability and proficiency*) were not correlated significantly with class size. On the other hand, it could readily be hypothesized that class size would be negatively correlated with scores on Factor II (*student-teacher interaction*). Calculation of the product moment correlation between the mean factor scores and the number of students in the class showed that both of these hypotheses were supported. The correlation between Factor I scores and class size was -0.18 (not significant when $N=84$); that between Factor II scores and class size was -0.40 (p 0.001). Although the correlation between Factor II and class size was significant, it was not so high as to indicate that class size alone accounted for these ratings.

In another attempt to assess the validity of the factor scores, the written comments made by the students for ten of the core courses were content-analyzed. These comments were read by two officers of the student body and evaluated as being positive or negative (critical). These evaluations were made without any knowledge of factor scores obtained by the individual classes. The ten courses were ranked by their scores on Factors I and II. Those that were 0.4 standard deviations below the mean were called "low"; those that were 0.4 standard deviations above the mean were called "high"; those within these ranges were called "average." The expectation was that there would be fewer critical comments made about courses that received high scores on Factor I, but that the scores on Factor II would not be related to the frequency of critical comments.

The relation between critical comments and ratings on Factor I are shown in Table 4. The significant chi-square (p 0.0001) supports the conclusion that the increased frequency of positive comments (from 36% for the classes rated "low" to 81% for those rated "high") is not a chance phenomenon. The corresponding data for Factor II are shown in Table 5. The nonsignificant chi-square supports the hypothesis that there was no relation between Factor II ratings and critical comments. Since only one of the core courses was rated "high" on Factor II, that cell was combined with the "average" category.

Conclusions

The attempt to evaluate courses in a school of public health by a rapid, standardized, reliable, and valid method proved successful. A high degree of student cooperation was enlisted, and responses to the scales were summarized by two independent factors. The analysis of the evaluation forms was completed three days after the final forms were re-

Table 4—Relation between type of written comment and evaluation of 11 core courses on Factor I (proficiency)

Comments:	Rating on Factor I: Low	Average	High	Total
Positive	33 (36)*	80 (55)	69 (81)	182
Negative	58 (74)	65 (45)	16 (19)	139
Total	91	145	85	321

$\chi^2=36.37$, $p<.0001$

* Parentheses indicate percentages.

turned. Within ten days the students and faculty had been informed in detail of the results of the evaluation.

Kerlinger[8] reported that primary and secondary school teachers ranked personal qualities (e.g., *positive person orientation*) above other characteristics. Pfieffer and Rosbach[9] reported that college students and teachers did not agree on what was of primary importance for teachers. Students felt that *knowledge dissemination* was the most important characteristic, while teachers chose *teacher dynamism* as primary. *Knowledge dissemination* referred to activities related to "analysis and synthesis, communication of knowledge, or production of new knowledge," while *teacher dynamism* referred to activities related to "personal warmth, involvement, and vigor." It is important to distinguish between teacher characteristics considered important by students and those considered important by teachers. The present evaluation, carried out at the graduate school level, showed that ratings given a teacher on *proficiency* were not related to ratings given him on *student-teacher interaction.*

The ultimate aim of course evaluations is improvement in the teaching of the courses. One faculty member suggested that the good teachers are more likely to be interested in the results of a curriculum evaluation, while ineffective teachers are less likely to react to such feedback. While this may be the case, there is always the chance that an objective evaluation may be more convincing than those often encountered.

The school of public health courses discussed in this paper received a generally positive evaluation. On this account, the low ratings given some courses would deserve special attention from the teachers responsible for those courses. A consideration of the ratings given on Factor I (*proficiency*) shows that 12 courses received mean standard scores that were lower than −0.55. If this were taken as an arbitrary cutoff point, these 12 courses would be regarded as in need of serious revision before being presented in the future. Another 11 courses received scores between −0.54 and −0.20, suggesting

Table 5—Relation between type of written comment and evaluaion of 11 core courses on Factor II (interaction)

Comments:	Rating on Factor II: Low	Average or high	Total
Positive	74 (52)*	108 (61)	182
Negative	69 (48)	70 (39)	139
Total	143	178	321

$\chi^2=2.22$, n.s.

* Parentheses indicate percentages.

that some inadequacies were detected in them. The 61 remaining courses (72.6%) received ratings that would indicate that they were satisfactory. While it is not advisable to make administrative decisions on the basis of a student evaluation of courses, those responsible for individual courses should regard student opinions as an unbiased, valuable source of information.

Summary

Student evaluation of teacher performance has been criticized as a "teacher popularity poll." The present report summarizes an effort to use the statistical technique of factor analysis to avoid that criticism in the evaluation of the curriculum in a school of public health. A 20-item scale was used to evaluate 89 different courses taught during the school year. Eighty-six per cent of the students completed the evaluation forms. A principal components factor analysis showed that the responses to the 20 items could be summarized by two independent dimensions. These were termed: *proficiency* or teaching ability (I) and *student-teacher interaction* or rapport (II). Scores on Factor I were related to the number of positive comments made by the students for 11 core courses, but the scores on Factor II were not related to the type of comment made. The results show that an evaluation of the quality of teacher performance (Factor I) can be made independently of the "popularity" of the teacher (Factor II).

REFERENCES

1. Tuckman, B. W., and Oliver, W. F. Effectiveness of Feedback to Teachers as a Function of Course. J. Educ. Psychol. 59: 297-301 (June), 1968.
2. Warren, R. To Publish or to Pipe. New England J. Med. 280:165-166 (Jan.), 1969.
3. Dunphy, J. E., and Freeman, D. L. Teachers, Students and Pipers. Ibid. 280:621 (Mar.), 1969.
4. Caffrey, B. Teachers and Pipers. Ibid. 280: 782 (Apr.), 1969.
5. Isaacson, R. L.; McKeachie, W. J.; Milholland, J. E.; Lin, Yi G.; Hofeller, M.; Bearwaldt, J. W.; and Zinn, K. L. Dimensions of Student Evaluation of Teaching. J. Educ. Psychol. 55:344-351 (Aug.), 1964.
6. Caffrey, B. Lack of Bias in Student Evaluations of Teachers. Proc., 77th Annual Convention, American Psychological Association 4:641-642 (Sept.), 1969.
7. Dixon, W. J. (ed.). Biomedical Computer Programs. Los Angeles: University of California Press, 1968.
8. Kerlinger, F. N. The Factor Structure and Content of Perceptions of Desirable Characteristics of Teachers. Educ. & Psychol. Meas. 27:643-656 (Nov.), 1967.
9. Pfieffer, M. G., and Rosbach, L. A. Teaching Performance Criterion Development Through Scaling: Teacher-Student Points-of-View Analysis. Perc. Mot. Skills 28: 755-766 (June), 1969.

Dr. Caffrey is Associate Professor of Psychology, Department of Social Sciences, Clemson University, Clemson, S. C. 29631. Dr. Kost is at the University of Texas School of Public Health, Houston, Texas.

This paper was submitted for publication in November, 1969.