ED 177 205 TM 009 707

AUTHOR Marsh, Herbert W.; Overall, J. U.

TITLE Validity of Students' Evaluations of Teaching: A Comparison with Instructor Self Evaluations by Teaching Assistants, Undergraduate Faculty and Graduate Faculty.

PUB DATE Apr 79

NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979); Appendices marginally legible

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *College Faculty; Course Evaluation; *Effective Teaching; Evaluation Criteria; Factor Analysis; *Faculty Evaluation; Graduate Students; Higher Education; Reliability; *Self Evaluation; *Student Evaluation of Teacher Performance; Teaching Assistants; Undergraduate Students; *Validity

IDENTIFIERS Interrater Reliability; *Multitrait Multimethod Techniques

ABSTRACT

College instructors evaluted the quality cf their own teaching and were evaluated by their students in 331 different courses. Student evaluations of teaching correlated with instructor self evaluations in courses taught by teaching assistants, in undergraduate courses taught by faculty, and in graduate level courses, demonstrating their validity at all levels of university teaching. Both student and instructor ratings were reliable, and separate factor analyses indicated that the same nine evaluation factors influenced both sets of ratings: learning/value; instructor enthusiasm; organization; group interaction; individual rapport; breadth of coverage; examinations/grading; value of assignments; and workload/difficulty. Student-instructor agreement on each factor was independent of the factor's agreement with other factors. While correlations between student and instructor ratings on the same factors were high, correlations between their ratings on different factors were low. This finding supports the distinctiveness of the different factors, as well as the use of multifactor evaluation instruments developed with the use of factor analytic techniques. These findings establish the validity and accuracy of student evaluations at all levels of university teaching, and suggest the possible usefulness of instructor self evaluations. (Questionnaires are appended.) (Author/MH)

Validity of Students' Evaluations of Teaching: A Comparison With Instructor Self Evaluations by Teaching Assistants, Undergraduate Faculty and Graduate Faculty

Herbert W. Marsh
University of Southern California

J.U. Overall
California State University,
Dominguez Hills

ABSTRACT

Instructors evaluated the quality of their own teaching and were evaluated by their students in each of 331 different courses. Student evaluations of teaching correlated with instructor self evaluations in courses taught by teaching assistants (r=.46), in undergraduate courses taught by faculty (r=.41), and even in graduate level courses (r=.39), demonstrating their validity at all levels of university teaching. Both student and instructor ratings were reliable, and separate factor analyses indicated that the same nine evaluation factors (learning/value, organization, enthusiasm, etc.) underlay both sets of ratings. Furthermore, student-instructor agreement on each factor was independent of its agreement on other factors. While correlations between student and instructor ratings on the same factors were high (median r=.45) correlations between their ratings on different factors was low (median r=.00). This argues for the distinctiveness of the different factors and for the use of multifactor evaluation instruments that have been developed with the use of factor analytic techniques. These findings establish the validity of student evaluations at all levels of university teaching, suggest the possible usefulness of instructor self evaluations, and will help reassure faculty about the accuracy of the student ratings.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge John Schutz, Donald Lewis, Joseph Kertes, and Robert Linnell, administrators and faculty at the University of Southern California who have supported the student evaluation program. Thanks are also extended to Zoe Cosgrove for her helpful comments and careful reading of this paper, and to Steven Kesler and Christopher Le Blanc for their assistance in compiling the data.

Paper presented at 1979 Annual Meeting
of American Educational Research
Association, San Francisco, April 1979.
Session 27.09

LIST OF TABLES AND APPENDICES

Validity of Students' Evaluations of Teaching: A Comparison with Instructor Self Evaluations by Teaching Assistants, Undergraduate Faculty and Graduate Faculty

Common criticisms of students' evaluations are that they are biased by variables unrelated to teaching effectiveness and that they lack validity. However, researchers have reported considerable empirical evidence indicating that most background variables, including class size, reason for taking the course, workload, and grade point average, are not substantially related to student ratings (Marsh, 1978; Marsh, Overall & Thomas, 1976; McKeachie, 1973; Remmers, 1963). In addition to this apparent lack of bias, student ratings have been validated against a variety of differert criteria. The most common criterion has been performance on a standardized examination; when different sections of the same courses are taught by different instructors, the sections that do best on the standardized examination given to all sections are also the ones who evaluate their instructors more favorably (Centra, 1977; Cohen & Berger, 1970; Frey, 1973; Marsh, Fleiner & Thomas, 1975, Overall and Marsh, 1978). Other researchers have successfully validated student ratings against the ratings of former students (Centra, 1973; Marsh, 1978; Marsh & Overall, 1979).

Validity research such as that described above has generally been limited to a specialized setting or has employed criteria that are unlikely to convince skeptics. Thus, faculty will continue to question the usefulness of student ratings until validity criteria applicable across a wide range of classes is utilized. A criterion that meets this requirement--instructor slef evaluations of their own teaching--should also be acceptable to most faculty and administrators. Instructors can be evaluated and evaluate their own teaching in any instructional context, even graduate level coureses and courses taught by teaching assistants. Furthermore, instructors can be asked to evaluate their own teaching along the same dimensions employed in the student rating form, thereby testing the specific validity of the different rating factors.

In spite of the apparent appeal of instructor self evaluations as a criterion for validating student ratings, relatively few studies have considered it. Centra (1973) found correlations of about .20 between faculty self evaluations and student ratings, but both sets of ratings were collected at midterm as part of a larger project that examined the impact of feedback from midterm evaluations. Blackburn and Clark also reported correlations of about .20, but they only asked faculty to rate their teaching in a general sense rather than to rate their teaching in the specific class being evaluated by students. In contrast, higher correlations have been reported in three other investigations. Doyle and Crichton(1978) found a median correlation of .47 between the self ratings of teaching assistants in 10 sections of a multisection course and the corresponding student ratings. Webb and Nolan(1955) reported a correlation of .62 in a military setting in which instructors were not professional teachers. Marsh, Overall, and Kesler(1979) asked regular faculty teaching undergraduate courses to evaluate themselves on the same form that was used by their students. Mean differences between faculty and student ratings were small, and separate factor analyses revealed that the same set of evaluation factors underlay both sets of ratings. The median correlation between self-ratings and student ratings was .49.

The Marsh, Overall and Kesler(1979) study served as a basis for the present one. This study, although a replication of the earlier research, differs in several important aspects. First, the evaluation instrument was expanded to include several new evaluation factors. Second, the sample size was increased to include 331 courses. Third, courses taught by teaching assistants and graduate level courses were included as well as undergraduate courses taught by faculty.

The present study has two purposes. First, it investigates the validity of student ratings for three instructional subgroups: courses taught by teaching assistants, undergraduate courses taught by regular faculty, and graduate level courses. Previous research has not considered the validity of the ratings in graduate level courses. Second, as a consequence of the large number of courses--a total of 331--this study permits a detailed application of the multitrait-multimethod procedure to test for both convergent and divergent validity. Convergent validity, which is typically considered, is based upon the correlation between student and faculty ratings on the same evaluation factor. However, even if general convergence is demonstrated, this does not argue for the usefulness of the many different evaluation factors often employed. Some users of student evaluations --faculty, administrators, and researchers--explicitly or implicitly assume that most of the useful information is contained in a single overall rating item or in a simple average across a number of specific items. This ignores the divergent or discriminant validity of the ratings. On the other hand, the demonstration that student-instructor agreement on any one dimension is independent of agreement on other dimesnions would demonstrate the utility of the distinct factors and argue for the use of factor analytic techniques in the development of evaluation instruments.

METHODOLOGY

During the academic year 1977-78 student evaluations were collected in virtually all courses offered in the Division of Social Sciences at the University of Southern California. Evaluations were administered shortly before the end of the term, generally by a designated student in the class or by staff person. Students were told that the evaluations would provide feedback to instructors and would be considered as part of personnel decisons. The surveys ware completed by an average of 76% of the students enrolled in each class.

The evaluation instrument (See Appendix I) consisted of 35 evaluation items adapted from Hildebrand, Wilson & Dienst(1971) and Marsh, Overall & Thomas (1976). The median reliability of indvidual evaluation items--intraclass correlation coefficients based upon sets of responses from 25 students per class--was .88(See Appendix II). A factor analysis (See Appendix III) of the student ratings of all undergraduate courses taught by regular faculty revealed nine separate evaluation factors. The reliability of the factors, coefficient alphas, varied from .88 to .97 (See Appendix II).

Instructor self evaluation surveys were sent to all teachers who had been evaluated by students in at least two different courses during the same term. Instuctors were asked to evaluate the effectiveness of their own teaching in both courses. These surveys were completed after the end of the term, but before summaries of the student evaluations were returned. While participation was voluntary, a cover letter from the Dean of the Division strongly encouraged cooperation and guaranteed the anonymity of each teacher's response. Instructors evaluated both courses with a set of items identical to those used by students, except that items were worded in the first person. They were specifically instructed to rate their own teaching effectiveness and not to report how students would rate them. A total of 181 (78%) surveys were returned. Since faculty had been requested to rate the effectiveness of their teaching in both classes they taught, self ratings for a total of 331 courses were completed--ratings of 183 undergraduate courses taught by faculty, 45 graduate level courses, and 103 courses taught by teaching assistants.

Eleven evaluation scores--factor scores representing the nine evaluation factors and overall ratings of the teacher and the course were used to summarize the student ratings and the instructor self ratings. Evaluation factor scores were weighted averages of standardized responses to each evaluation item. The weights, factor score coefficients, were derived from the factor analysis described in Appendix III.

In addition to actual evaluation of their own teaching, faculty were asked to express their agreement or disagreement with statements about student evaluations and other methods of evaluating the effectiveness of their teaching. Faculty also rated themselves and the course they taught on selected background variables that have been suggested as potential biases to the student ratings (e.g., their "grading leniency", their "popularity with students", their perceptions of their students' subject interest before the start of the course, etc.). Attitudes and variables faculty felt were likely to bias the student ratings are presented in Appendix VI; the relationship between both student and faculty ratings and potentially biasing variables are presented in Appendix VII.

RESULTS

Faculty Attitudes Toward Student Ratings.

As part of the study, faculty were asked to express their agreement or disagreement with statements concerning student ratings, potential biases in student ratings, and other possible methods of evaluating the quality of their teaching (See Table 1). A majority (59%) of the faculty indicated that some measure of teaching quality should be given more emphasis in promotional decisions. Faculty clearly agreed that student ratings were useful to the faculty themselves as feedback, and a majority even agreed that they should be made publicly available for students to use in course selection. However, they were more skeptical about the accuracy of the student ratings. Furthermore, faculty were even more critical about using classroom visitation by peers or faculty self evaluations in promotional decisions, though they were somewhat more favorable towards the use of colleague examination of course outlines, reading lists, classroom examinations, etc.

Insert Table 1 About Here

Faculty were also asked to indicate the items in a list of "potential biases" that they believed would actually cause a substantial bias. The most frequently mentioned were: Course Difficulty (72%), Grading Leniency--lots of A's" (68%), Instructor Popularity (63%), and Student Interest in Subject Before Taking Course (62%). It was interesting to note, however, that faculty self evaluations of their own teaching and student evalutions of the faculty were: 1) both positively related to Workload/Difficulty (harder courses were evaluated more favorably by both), 2) both positively related to faculty self ratings of their "popularity with students", 3) both positively related to student prior subject interest, and 4) both uncorrelated with faculty self ratings of their "grading leniency". These findings suggest that three of these variables --workload, prior subject interest, and instructor popularity--are variables actually related to quality of teaching, since each shows similar relationships to two different measures of teaching quality. The fourth variable, grading leniency, is apparently unrelated to either quality of teaching or student ratings of quality of teaching.

A dilemma clearly exists. Faculty are concerned about teaching effectiveness, even to the extent of wanting it to play a more important role in their own promotions. However, many expressed doubts about any of the possible measures of teaching effectiveness that were suggested-including student evaluations. In particular, faculty suggested a number of sources of potential bias in the student ratings, even though each of these potential biases showed similar relationships to student and faculty ratings of teaching effectivenss. Before the potential usefulness of student ratings can be realized, faculty and administrators have to be convinced that student ratings are valid and relatively free of bias.

Factor Analysis

Separate factor analyses were performed on student and instructor self ratings of the 35 evaluation items (See Table 2), to determine if the evaluatio factors underlying student evaluations were similar to those representing instructor self evaluations. Both confirmed the nine evaluation factors that had previously been identified. Each item, for both student and instructor ratings, loaded highest on the factor it was designed to measure. Loadings for items defining each factor were generally at least .40 and usually exceeded .50. All other loadings were less than .30 and were usually less than .20. The similarity in the two factor patterns implies that similar dimensions underlie both student and instructor ratings of effective teaching. The results of both factor analyses were also quite similar to results of a previous factor analysis performed on the student ratings of all undergraduate courses taught by faculty (See Appendix II) -- including those considered in this study.

Insert Table 2 About Here

Several analytic techniques are available for the comparison of differen factor analyses, but none have been thoroughly developed (Levine, 1977). Target analysis, the rotation of one structure to fit the structure of another, is better suited to matching one empirical structure to a second theoretical structure. Furthermore, it forces data--while capitalizing on chance--to fit the proposed model, or as suggested by Hurley and Cattell (1962), the procedure "lends itself to the brutal feat of making almost any data fit almost any hypothesis." An alternative procedure developed by Cattell and Baggaley (1960), the salient variable similarity index, classifies loadings into those that are higher than an arbitrarilly defined substantial loading and those that are not. This procedure ignores much of the information in the loadings by converting them into dichotomies. Thus, at least in this application, careful selection of the "substantial" cutoff would result in "perfect" fits for all factor patterns. Levine(1977), amoung other procedures, suggests simply correlating the the factor loadings. In the present application, each factor pattern (See Table 2) had 315 factor loadings; each of 35 items had loadings on each of the nine factors. Factor loadings for the factor analysis of instructor self ratings correlated r = .90 with both the loadings from the factor analysis of student ratings in this study and the previous analysis of student ratings in all undergraduate courses taught by faculty; loadings from the two factor analyses of student ratings correlated .95 with each other. These results also confirm the similarity of the factor patterns resulting from student and instructor ratings.

Convergent and Divergent Validity

Campbell and Fiske (1959) advocate the assessment of validity by determining measures of more than one trait, each of which is assessed by more than one method. In the present application, the multiple traits are the nine evaluation factors, while the multiple methods refer to the two distinct groups of raters--students rating their instructor and the instructors rating themselves. Convergent validity, that which is most typically determined, is the correlation between the same evaluation factors rated by two different groups. Discriminant validity refers to the distinctiveness of each of the evaluation factors. Two different aspects of discriminant validity are particularly relevant to the present application. The first examines whether or not student-instructor agreement on each factor is independent of agreement on other factors. For example, if a single "generalized rating factor" underlies both student and instructor ratings, then agreement on any particular factor might be a function of agreement on the generalized factor and not have anything to do with the specific content of the factor being considered. As a consequence, while correlations between student and instructor ratings on the same factors would be high, so would the correlations between their ratings on different factors. The second aspect of discriminant validity considers the possibility that the relationship between different factors as rated by the same group of raters is due to the method of data collection rather than "true" relationships between the underlying dimensions being considered. The most likely source of this method variance in the present application would be a halo effect.

Convergent and discriminant validity across all courses was determined by examining the correlation matrices in Table 3. The correlations between different evaluation factors as assessed by the same group of raters are contained in the two triangular matrices--intercorrelations among instructor self evaluation factors (upper left) and student evaluation factors (lower right). The diagonals of these trianqular matrices contain the reliabilities of the factors--coefficient alphas--for each group of raters. The square matrix (lower left) contains the correlations between student evaluation factors and instructor self evaluation factors. The diagonal of the square matrix (the convergent validity coefficients) contains correlations between the same evaluation factors as assessed by the two different groups. Since there is unreliability in both the student ratings (median reliability = .94) and particularly the instructor self evaluations (median reliability = .82), the convergent validity coefficients have been corrected for unreliability. The set of matrices in Table 3, referred to as a multitrait-multimethod matrix, was based upon the combined data of all three sets of classes--those taught at the graduate and at the undergraduate levels by regular faculty and those taught by teaching assistants. Multitrait-multimethod matrices were also contructed separately for each of the three sets of classes (See Appendix IV).

Insert Table 3 About Here

Convergent validity requires that the diagonal values of the square matrix be substantially higher than zero. Inspection of Table 3 indicates that this was the case for all evaluation factors. Validity coefficients varied between .17 and .69 (median r=.[illegible]) and all were statistically different from zero. These finding demonstrate good support for the convergent validity of teacher evaluations. Convergent validity was also determined separately for each of the three sets of courses (See Appendix IV and Table 4). The median convergent validity coefficient was .41 for faculty-taught undergraduate courses, .39 for faculty-taught graduate courses, and .46 for undergraduate level courses taught by teaching assistants. Only four of these 27 validity coefficients--three of the nine validity coeficients for graduate level courses and one of nine for courses taught by teaching assistants-failed to reach statistical significance. However this--as compared to the combined data in which all validity coefficients were significant--was a function of the reduced sample size rather than lower validity coefficents; every validity coefficient in each of the sets of classes would have been statistically significant if based on the same number of cases as in the combined data.

Insert Table 4 About Here

Divergent validity is harder to assess, and Campbell and Fisk(1959) offer only general guidelines. The minimal condition is that all correlations between different factors rated by the same group (off-diagonal correlations in the triangular matrices) must be substantially lower than the reliabilities of these factors. This tests whether the different evaluation factors as judged by the same group of raters are distinctive. This condition was clearly met for instructor self evaluations, and even the moderate intercorrelations among the student ratings (median r=.33) were much lower than the reliabilities of these factors (median r = .94). These same general conclusions hold when matrices for each of the three groups of courses were considered separately (See Appendix IV).

Campbell and Fisk(1959) stated that "various statistical treatments for multitrait-multimethod matrices might be developed, but we feel that such summary statistics are neither necessary nor appropriate at this time." Instead, they suggest three general guidelines that have more intuitive appeal than quantitative rigor. While other researchers have attempted to develop more rigorous procedures, they have been only partially successful (See Alwin, 1973) and most applications of the multitrait-multimethod procedure still rely on the orginal guidelines proposed by Campbell and Fisk(1959).

The first of their guidelines is that each convergent validity coefficient (diagonals of the square matrix in Table 3) should be lower than th any other correlation in the same row or column of the square matrix. This test requires that each of the nine convergent validity coefficients be higher than any of the 16 correlations in the same row or column of the square matrix. For example, the validity coefficent for Instructor Enthusiasm was .54 (.48 if not corrected for unreliability). This was higher than any of the eight correlations between student ratings of Enthusiasm and the eight other instructor self-rating factors, and was also higher than any of the eight correlations between instructor self-ratings of Enthusiasm and the eight other student rating factors. With one minor exception--the Examinations Grading factor failed the test in one of 16 comparisons--this guideline was met in all cases. Inspection of the separate mutitrait-multimethod matrices constructed for each of the three sets of courses also indicates that this test was met with few exceptions. The only evaluation factor that did not consistently demonstrate divergent validity was Examinations/Grading; it consistently passed this test for only graduate level courses taught by faculty.

Their second guideline requires that each convergent validity coefficient be higher than correlations between that factor and any other factor assessed by the same group of raters. For example, the validity coefficient for Enthusiasm (r = .54) was higher than any correlation between student ratings of Enthusiasm and any other student rating, and was higher than the correlation between instructor ratings of Enthusiasm and any other instructor rating. This guideline is the most stringent, and has several problems when applied to this particular setting. Its application implicitly assumes that the different factors are truly uncorrelated--an assumption that seems unwarranted in this case. Thorndike (1920) suggests, for example, that there should be little or no true correlation between a teacher's intelligence and the quality of his voice, and the obtained correlation of .63 between ratings of these attributes clearly suggests a halo effect. It is not so clear that an instructor's enthusiasm in teaching a course should be unrelated to student learning in the course. Trying to separate halo effect from true relationships among the underlying dimensions was further complicated by the fact that the reliability of the student ratings was consistently higher than the reliability of the instructor self ratings. The higher reliabilities of the student ratings was a function of the fact that each student rating was the mea response from different students, while each faculty self rating was based upon the response of only one individual (See Doyle & Crichton, 1978 and Marsh & Overall, 1979). Nevertheless, if true relationships did exist between the different rating dimensions, then--as a consequence of the higher reliabilities alone--correlations among the student ratings would be higher than among instructor self ratings.

Convergent validity coefficients were higher than correlations among instructor self evaluations, even when corrected for unreliability, for all but one factor--Examinations/Grading. However, this second guideline was only partially satisfied when validity coefficients were compared to correlations among student rating factors; 23 of 72 comparisons (eight comparisons for each of the nine factors) failed this test and most of these were for comparisons involving the Examinations/Grading and Organization factors. In general, these conclusions hold when this test is applied separately to each of the three sets of classes; failures of this test were more common in comparisons between validity coefficients and student ratings than in comparisons involving instructor self evaluations. Failures for instructor self ratings were most frequent with the Examination/Grading factor; failures for student ratings were most common with the Examination/Grading and Organization factors.

Their third and final guideline is that the pattern of intercorrelations among different factors should be similar in both the triangular and square matrices. For example, there were four correlations between the factors of Enthusiasm and Learning; the correlations between instructor self evaluations o Enthusiasm and Learning (.29--upper left triangular matrix), student ratings of Enthusiasm and instructor ratings of Learning(.21--below the diagonal of the square matrix), instructor ratings of Enthusiasm and student ratings of Learnin (.10--above the diagonal of the square matrix), and student ratings of Enthusiasm and Learning(.45--lower right triangular matrix). Inspection of Table 3 reveals that these four correlations were among the highest in each of the respective sets of correlations. For the nine factors there are 36 possible pairs of different factors, and the relationship between ratings of each of these 36 pairs is represented by four correlations (e.g., the four correlations between Enthusiasm and Learning described above). To test the similarity of th pattern of correlations among the different factors, the 36 correlations among the instructor self ratings were correlated with the corresponding 36 correlations among student ratings. The pattern was quite similar ($r = .43$, $p < .01$), implying that some of the covariation among factors represents a true relationship among the underlying dimensions rather than a simple halo effect.

An alternative approach, based upon multiple regression was also used to explore the multitrait-multimethod matrices. In the first stage, multiple regression was used to predict each instructor rating using the entire set of nine student ratings, and to predict each student rating with the entire set of nine instructor rating factors. For each of these 18 regressions, two aspects were of particular interest: 1) how much variance beyond that explained by the matching variable alone could be accounted for by the entire set of nine variables, and 2) how much of the variance explained by the entire set was uniquely due to the matching variable. The unique contribution was taken to be the change in multiple R squared (adjusted for the number of variables in the equation) due to the matching variable when it was entered separately as the last variable in the regression equation. For example, instructor ratings of Enthusiasm alone were able to explain 23% (before correcting for unreliability) of the variance in student ratings of Enthusiasm (See Table 5). The entire set of 9 instructor ratings was able to explain 24% of the variance in student ratings of enthusiasm--an addition of only 1%. Furthermore, most of the variance that could be explained by the entire set of nine variables was uniquely due to the matching variable (18% of the 24%). Averaged across all nine student rating factors, the matching instructor rating variable alone could explain 14% of the variance in student ratings, while the entire set of nine instructor ratings could explain 17%--an addition of only 3%. Furthermore, 13% of this 17 was uniquely due to the matching variable alone. Consequently, little variance in student ratings that was explained by student-instructor agreement on the same factor could be predicted by any of the other eight instructor rating factors (1% of the 14%).

The results of the first stage of multiple regression analyses has implications of particular interest to this study. Most importantly, virtually none of the variance in student ratings that could be explained by instructor ratings on the same factor could be explained by any of the other instructor ratings; only 1% of the 14%. This finding offers strong support for the conclusion that student-instructor agreement on each particular factor was distinct from agreement on other factors. It also argues for the importance of using multifactor evaluation instruments that have been developed with factor analytic techniques.

The second stage in the multiple regression analysis was to predict each student rating with the eight other student ratings and the one matching instructor rating, and to predict each instructor rating with the other eight instructor ratings and the one matching student rating. For each of these 18 regressions, the unique contribution of the matching factor was determined as described in the first stage. This analysis was directed to the issue of a halo effect. Specifically, how much variance in student ratings could be explained by the remaining eight student factors, how much additional variance could be explained by the rating of the same variable by instructors, and how much of the variance in student ratings that was attributable to student-instructor agreement on the same factor could also be explained by other student factors? Averaged across all nine factors for all courses the other student rating factors explained 39% of the variance and the matching instructor self-rating uniquely accounted for an additional 8%. This suggests that there is considerable covariation among the student rating factors beyond that which can be explained by either student-instructor agreement on the same rating factors or even the relationship between each student rating factor and the entire set of instructor self rating factors (the analysis performed in stage one of the multiple regression analysis). The same conclusion does not hold for the instructor self ratings. On the average, covariation among the instructor self ratings factors accounted for only 10% of the variance within the factors, and the matching student rating factor uniquely contributed an additional

12% to the total variance that could be explained. These findings show that for the student ratings there is the possibility of a substantial halo effect, but that there is little indication of a halo effect in instructor self ratings.

The same multiple regression analyses were performed on each of the three sets of classes separately (See Table 5). The findings of each of these separate analyses were similar to those reported for all classes. In each of the three set of classes, student-instructor agreement on the same evaluation factor was reasonably distinctive from agreement on other factors, and most of the variance in the student ratings that could be explained by the entire set of instructor ratings was uniquely due to the student-instructor agreement on the same factors. Furthermore, there was evidence suggesting a halo effect in each set of student ratings, but little halo effect in any of the instructor ratings. Covariation among student ratings for courses taught by teaching assistants was somewhat larger than in other courses, but more of this covariation was explicable in terms of covariation among instructor ratings as well.

Insert Table 5 About Here

Across all nine evaluation factors student-instructor agreement was generally high, but the extent of the agreement did vary considerably. In particular, except for graduate level courses, there was lower agreement on the Examinations/Grading factor. Factor analyses of the student and instructor ratings (See Table 2) indicated that the dimension was well defined, and its reliability was comparable to the other factors(See Table 4). Examination of agreement on individual items (See Appendix VII) suggests the lack of good agreement was consistent across each of the three items designed to measure the factor, but was particularly marked for the item "methods of evaluating student work were fair and appropriate". the correlation for this item was the lowest of the 35 items and the only one that failed to reach statistical significance. Furthermore, the difference between mean instructor and mean student rating--instructor ratings were about one-quarter of a category higher on a five-point response scale--was also one of largest for any of the 35 individual items. Differences ranged from +0.37 (higher student ratings on the item "Instructor enhanced presentations with the use of humor") to -0.27 (higher instructor ratings on the item "methods of evaluating student work were fair and appropriate"); instructor self ratings were significantly higher on 6 items and significantly lower on 10 items.

The lack of good agreement on the Examinations/Grading factor is difficult to interpret. Results of the factor analysis and the factor's reliability both suggest that the factor is "real", and intuitively it would seem to be an important aspect of teaching. Perhaps, instructors just have no basis for assessing the quality of their examinations, and the student ratings might be valid even though they do not agree with instructor self ratings. In fact, other research has shown this factor to be valid when the validity criterion was objective student learning (Frey, 1973; Overall & Marsh, 1978) or student disposition towards further study and application of the course content(Overall & Marsh, 1978). However, the lack of convergent validity demonstrated in this study also has implications for the discussion of divergent validity as well. Each of the guidelines proposed by Campbell and Fisk(1959) involves a comparison between a convergent validity coefficient and some other correlation coefficient. If the convergent validity for a factor is low, the factor will automatically fail the divergent validity tests. Any correlation between this factor and other factors will appear to be halo effect. In particular, comparisons involving the Examinations/Grading factors most frequently failed the divergent validity tests and contributed to the conclusion that there was a halo effect present. This was true in spite of the fact that other sources suggest that at least the student ratings of this factor may be more valid than suggested by the lack of agreement with instructor self ratings.

In summary, several different approaches have supported both the convergent and at least one aspect of divergent validity of the teacher evaluations. The convergent validity of the teacher evaluations--agreement between student and instructor ratings on the same factors--was consistantly demonstrated for each of the nine factors considered across all courses combined and within each of the three sets of courses considered separately. Student-instructor agreement on the same factor was shown to be independent of agreement on different factors and could not be explained in terms of a generalized evaluation factor that was common to both student and instructor ratings, thus illustrating one aspect of divergent validity. The question of a halo effect, particularly among the student ratings, was more complicated. The similarity of the pattern of relationships among student ratings and the corresponding pattern among instructor ratings implies that some of the covariation among factors represents true relationships among the underlying dimensions. Part of the elevated correlations among student ratings as compared to instructor ratings was a function of the higher reliabilities of the student ratings. Furthermore, some of the apparent halo effect among student ratings may also be a function of a lack of validity in the instructor ratings; particularly with the Examination/Grading factor. However, there was still a strong indication that there might also be some halo effect underlying the student ratings, though probably not the instructor ratings. The existence of some halo effect in the student ratings, if it does exist, does not undermine either the convergent validity of the teacher evaluations or the specificity of the student-instructor agreement on different factors.

DISCUSSION

Instructors evaluated the effectivenss of their own teaching and were evaluated by their students on the same 35 item evaluation form in a total of 331 different courses. The study included undergraduate and graduate level courses taught by faculty and undergraduate courses taught by teaching assistants. In spite of faculty scepticism concerning the validity of student ratings and their belief that many sources of potential bias do substantially impact the ratings, there was good student-instructor agreement. Separate factor analyses of student and instructor self ratings both resulted in the same set of nine evaluation factors that had been previously identified. This suggests that similar dimensions underlie both student and instructor evaluations. Correlations between students and instructor on the same factors were generally high (median r = .45) and always statistically significant, while correlations between student and instructor ratings on different factors tended to be low (median r = .00) and generally did not reach statistical significance. This argues for the validity of the ratings in general, and for the distinctiveness of the different factors. While the validity coefficients were slightly lower for graduate level courses--median r = .39 as opposed to

.41 and .46 for undergraduate courses taught by faculty and teaching assistants respectively--the general conclusions based upon the entire set of courses were also true for each of the three sets of classes considered separately. This offers evidence for the validity of student ratings at all levels of university teaching.

Several alternative approaches were used to explore both the convergent and divergent validity of the teacher evaluations. Convergent validity, that which is typically determined, refers to the relationship between student and instructor ratings on the same evaluation factor. The results of the study offered clear support for the convergent validity of teacher evaluations.

Divergent or discriminant validity was assessed by seeking the answers to two related questions. First, is the student-instructor agreement on the same evaluation factors specific to that factor, or can it explained in terms of a generalized agreement common to all the different factors? Second, are the correlations between the different factors as evaluated by faculty and students indicative of a halo effect, or do they represent true relationships among the underlying dimensions? The answer to the first question was quite clear; student-instructor agreement on the same evaluation factors was specific and distinctive from other factors. While correlations between student and instructor ratings on the same factors were uniformly high, correlations between their ratings on different factors were generally low. Furthermore, virtually none of the variance in student ratings that could be explained by student-instructor agreement on the same factors could be explained by instructor ratings on any of the other eight factors.

The question of a halo effect was somewhat more complicated. Correlations among the different student factors (median r = .33) were definitely lower than the reliabilities of their ratings (median r = .94), but were higher than those among instructor self-ratings (median r = .09). Part of this could be explained in terms of the lower reliability of the instructor self ratings (median r = .82), and some of it could be explained in terms of a true relationship underlying some of the factors indpendent of the method of collection. Furthermore, for ratings of Examinations/Grading in particular, a possible lack of validity in the instructor ratings would give the appearance of an inflated halo effect in the student ratings, even though alternative criteria have supported the validity of student ratings for this factor. However, the results still suggest that there was at least some halo effect in the student ratings. There was little evidence for any halo effect in the instructor self ratings.

Three previous studies most comparable to this investigation reported convergent validity coefficients of .47 (Doyle and Crichton), .62 (Webb and Nolan, 1955) and .49 (Marsh, Overall and Kesler, 1979). Two of these studies (Doyle & Crichton, 1978; Marsh, Overall & Kesler, 1979) also consdiered the divergent validity of the teacher evaluations. Doyle and Crichton found little support for the discriminant validity of the ratings, but their study was based upon correlations among only 10 different sections, and they considered ratings of individual items rather than evaluation factors. In the Marsh, Overall, and Kesler study, there was good support for both the convergent validity and the divergent validity of the student ratings. The results of the present study provide a strong replication of this previous finding.

Many researchers (e.g., Whitely & Doyle, 1978; Marsh, 1978; Beatty & Marsh, 1975; Frey, Leonard & Beatty, 1975; Finkbeiner, Lathrop & Schulerger, 1971; Hildebrand, Wilson & Dienst, 1971; Bendig, 1954) have used factor analytic techniques to identify distinct dimensions that underlie student ratings of teaching quality. Frey (1978) recently argued for the existence of two distinct factors which he labeled as "pedagogical skill" and "rapport". He demonstrated that his skill factor was more reliable and more closely related to objective student learning, while his rapport factor was correlated with class size and expected grade. While Frey's study (1978) did not demonstrate that there were only two factors (his factor factor analysis was based upon only seven items and several of these had substantial loadings on both his factors), it convincingly showed that different components of the student ratings have quite different meanings. Overall and Marsh (1978) also found that some evaluation components (e.g., Tnstuctor Enthusiasm and Overall Instructor Rating) were more closely related to objective student learning, while others (e.g., Learning/Value & Overall Course Rating) were more closely related to student disposition towards further study and application of the course content. Other findings presented in Appendix VII of the present study show that for both student and instructor ratings, student prior subject interest was more highly correlated with Learning/Value than with other factors. Similarly, course enrollment was highly corrolated with quality of Group Interaction, but not with other factors (also see Marsh, Overall & Kesler, 1979b).

The studies above each argue for distinctive interpretations of the meaning of different student evaluation factors. Yet, in spite of this growing evidence to the contrary, some users of student evaluations--students, faculty, administrators, and even researchers--continue to assume that all the useful information can be obtained from a single rating or simple average of ratings. The findings of this study offer dramatic evidence that this is not so, and argue for the distinctiveness of the different evaluation factors. Student-instructor agreement on each of the nine evaluation factors was independent of their agreement on the other factors. While there was some evidence for a generalized factor within the student ratings, perhaps indicative of some halo effect, it did not contribute to the specific student-instructor agreement on the same factors. In fact, correlations between student and instructor ratings on different factors were generally quite low. This conclusion argues for the use of multifactor evaluation instruments that have been carefully constructed with the use of factor analytic procedures.

Students' evaluations of teaching effectiveness will not be useful unless faculty and administrators are convinced of their worth. While researchers have demonstrated their reliability, validity, and relative lack of bias, many faculty remain sceptical. This scepticism, whether justified or not, will continue to undermine the value of the student ratings until they have beer validated against criteria that are acceptable to most faculty. In the present investigation, student ratings were validated against instructor self evaluations of their own teaching effectiveness. This criterion, besides being acceptable to most faculty, has two distinct advantages. First, it can be applied to all levels of instruction; student ratings were successfully validated against instructor self evaluations in graduate level courses and courses taught by teaching assistants as well as undergraduate courses taught by faculty. Second, instructors can be asked to evaluate their teaching along the same dimensions employed on the student rating form; in the present study it was shown that student-instructor agreement on any one factor was independent of agreement on other factors. In summary, the findings of this investigation establish the validity of student ratings at all levels of postsecondary education, demonstrate the importance of the distinctiv evaluation factors, and should also be helpful in overcoming faculty reservations about the usefulness of student ratings.

BIBLIOGRAPHY

Alwin, D. F. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (ed.), Sociological Methodology (1973-1974). San Francisco: Jossey-Bass Publishers, 1974.

Beatty, P. & Marsh, H. W. Students' Evaluations of Instructional Effectiveness: Research and a Survey Instrument, Los Angeles: Evaluation of Instruction Program, University of California, Los Angales, 1974.

Bendig, A. W. A factor analysis of student ratings of psychology instructors on the Purdue Scale. Journal of Educational Psychology, 1954, 45, 385-393.

Blackburn, R. T. & Clark, M. J. An assessment of faculty performance: Some correlates between administrators, colleagues, students, and self-ratings. Sociology of Education, 1975, 18, 242-256.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cattell, R. B., & Baggaley, A. R. The salient variable similarity index for factor matching. British Journal of Statistical Psychology, 1960, 13, 33-46.

Centra, J. A. The Student Instructional Report: Item Reliabilities (Student Instructional Report #3) Princeton, N. J.:Educational Testing Service, 1973.

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

Centra, J. A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Cohen, S. A. & Berger, W. G. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. Proceedings of the 78th Annual Convention of the American Psychological Association, 1970, 5, 605-606. (Summary)

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

Finkbeiner, C. T., Lanthrop, J. S. & Schuerger, J. M. Course and instructor evaluation: Some dimensions of a questionnaire. Journal of Educational Psychology, 1973, 64, 159-163.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Frey, P. W. A two-dimensional analysis of student ratings of instruction. Research in Higher Education, 1978, 9, 69-91.

Frey, P. W., Leonard, D. W. & Beatty, W. W. Student ratings of instruction: Validation research. American Educational Research Journal, 1975, 12, 327-336.

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

Hurley, J. R. & Cattell, R. B. The procrustes program: Producing direct rotation to test a hypothesized factor structure. Behavioral Science, 1962, 7,258-26

Lovine, M. S. Canonical Analysis and Factor Comparison. Sage University Paper Series on Quantitive Applications in the Social Sciences, Series No. 07-001. Beverly Hills, Calif: Sage Publications, 1977.

Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14, 441-447.

Marsh, H. W. Students' Evaluations of Instructional Effectiveness: Relationship to Student, Course, and Instructor Characteristics. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, March 1978. (ERIC Document Reproduction Service No. ED 155 217)

Marsh, H. W., Fleiner, H., & Thomas, C. S. Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 1975, 67, 833-839.

Marsh, H. W. & Overall, J. U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". Research in Higher Education, 1979, in press.

Marsh, H. W., Overall, J. U. & Kesler, S. P. Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their student. Journal of Educational Psychology, 1979, 71, in press.

Marsh, H. W., Overall, J. U., & Thomas, C. S. The Relationship Between Students' Evaluation of Instruction and Expected Grade. Paper presented at the Annual Meeting of the American Educational Research Association, San Fransisco, April 1976. (ERIC Document Reproduction Service No. ED 126140)

McKeachie, W. J. Correlates of student ratings. In Sockloff, A. L. (Ed.), Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia: Measurement and Research Center, Temple University, 1973.

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. Journal of Educational Psychology, 1956, 47, 79-88.

Overall, J. U. & Marsh, H. W. Relationship Between Student Evaluations of Teaching, Faculty Self-evaluations, and Student/Instructor Course Characteristics. Los Angeles: Office of Institutional Studies, University of Southern California, 1978.

Remmers, H. H. Teaching methods in research on teaching. In Gage (Ed.). Handbook on Teaching. Chicago: Rand McNally, 1963.

Rodin, M. & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 22, 415-430.

Webb, W. B. & Nolan, C. Y. Student, supervisor, and self-ratings of instructional Proficiency. Journal of Educational Psychology, 1955, 46, 42-46.

Whitely, S. E., & Doyle, K. O. Dimensions of effective teaching: Factors or artifacts. Educational and Psychological Measurement, 1978, 38, 107-117.

TABLE 1

Faculty Attitudes Toward Students' Evaluations of Teaching Effectiveness

	% RESPONDING DISAGREE (1-3)	% RESPONDING NEUTRAL (4-6)	% RESPONDING AGREE (7-9)	MEAN RESPONSE
Quality of teaching, whether determined by students' evaluations or other methods, should be given more emphasis in making promotional decisions.	8%	33%	59%	6.5
Students' evaluations represent accurate assessments of instructional quality.	25%	37%	38%	5.2
Students' evaluations provide information which is potentially useful for the improvement of the course and/or quality of teaching.	4%	16%	80%	7.0
Students' evaluations actually have been useful to you for the improvement of a course and/or quality of teaching.	11%	30%	59%	6.3
Students' evaluations should be made available to students for use in course selection.	13%	35%	52%	6.2
Colleague evaluation of course materials, as a measure of quality teaching, should be given careful consideration in promotional decisions.	19%	37%	44%	5.7
Classroom visitation evaluations by colleagues, as one measure of quality teaching, should be given careful consideration in promotional decisions.	33%	37%	30%	4.9
Instructor Self-Evaluation, as one measure of quality teaching, should be given careful consideration in promotional decisions.	32%	41%	27%	4.7

POTENTIAL BIASES IN STUDENTS' EVALUATIONS: Critics of students' evaluations suggest that some variables unrelated to quality of teaching may have a significant influence on the ratings. Below is a list of some potential biases and the percentage of instructors who believed that each influenced ratings.

68%	Grading Leniency (Lots of "A's")	62%	Student Interest In Subject Before Course	28%	Instructor's Appearance
55%	Class Size/Enrollment	23%	Course Level (upper division vs. lower)	15%	Instructor's sex
55%	Required vs. Elective	53%	Students' Scholastic Ability Measured by GPA	8%	Instructor's rank
60%	Course Workload	15%	% Frosh & Soph Students in the Class	63%	Instructor's Popularity
72%	Course Difficulty	20%	Instructor's Age	28%	% of Students Majoring in A Department
16%	Instructor's Academic/ Research Prestige	35%	Student's Prior Knowledge of Course Content		

NOTE: Only faculty responses were included in this table. Attitudes expressed by teaching assistants to the first 9 items were similar to those of faculty except that they expressed even stronger agreement with the statement endorsing the importance of some measure of effective teaching being given more emphasis in promotional decisions.

TABLE 2

Factor Analyses of Students' Evaluations of Teaching Effectiveness and the Corresponding Faculty Self Evaluations of Their Own Teaching in All 131 Courses

Evaluation Items (paraphrased)	Factor Pattern Loadings I	II	III	IV	V	VI	VII	VIII	IX
I LEARNING/VALUE									
Course Challenging/Stimulating	42(50)	23(25)	09(-10)	04(04)	00(-03)	15(27)	09(05)	16(23)	29(20)
Learned something valuable	53(77)	15(02)	10(-02)	09(04)	01(-01)	10(00)	10(-04)	17(09)	16(06)
Increased Subject Interest	57(70)	12(05)	08(07)	08(07)	02(-03)	14(04)	33(-04)	19(05)	14(-02)
Learned/Understood Subject Matter	55(52)	12(12)	13(12)	05(03)	03(11)	02(-01)	19(07)	14(-04)	-23(-11)
OVERALL COURSE RATING	36(33)	25(29)	16(09)	12(05)	09(02)	12(16)	13(-04)	14(27)	08(16)
II ENTHUSIASM									
Enthusiastic about teaching	15(29)	55(82)	16(00)	07(02)	21(15)	10(00)	05(16)	01(09)	05(06)
Dynamic & Energetic	08(03)	70(70)	15(01)	11(06)	08(05)	06(05)	07(16)	01(05)	06(03)
Enhanced Presentations with Humor	10(04)	66(58)	-04(04)	05(01)	13(02)	12(02)	14(07)	02(-14)	-07(-10)
Teaching Style Held Your Interest	09(12)	59(64)	23(20)	16(06)	06(00)	03(14)	10(05)	06(03)	-02(-03)
OVERALL INSTRUCTOR RATING	12(27)	40(54)	23(09)	14(08)	23(02)	11(16)	10(-08)	05(27)	05(16)
III ORGANIZATION									
Instructor Explanations Clear	12(00)	07(24)	55(82)	20(09)	05(-04)	10(06)	13(01)	06(23)	-08(-03)
Course Materials Prepared & Clear	06(06)	03(-02)	73(69)	09(01)	10(-02)	09(04)	06(03)	10(03)	01(12)
Objectives Stated & Pursued	19(12)	-05(-04)	49(41)	03(05)	08(05)	14(08)	25(27)	06(05)	06(06)
Lectures Facilitated Note Taking	-03(02)	20(09)	58(53)	-17(07)	-02(05)	14(04)	15(06)	08(01)	-04(-05)
IV GROUP INTERACTION									
Encouraged Class Discussions	04(06)	10(02)	-01(03)	88(86)	03(00)	00(00)	06(00)	06(-05)	00(-03)
Students Shared Ideas/Knowledge	02(08)	06(-07)	-04(-03)	85(88)	05(13)	05(01)	08(-02)	08(-10)	-02(01)
Encouraged Questions & Answers	03(-04)	06(09)	14(06)	62(69)	16(-02)	15(03)	07(11)	08(21)	00(01)
Encouraged Expression of Ideas	07(01)	02(06)	01(-11)	73(75)	20(09)	05(07)	09(12)	05(09)	00(-02)
V INDIVIDUAL RAPPORT									
Friendly Towards Students	-08(10)	17(06)	00(-06)	13(12)	68(78)	-01(-05)	13(02)	10(-05)	-07(01)
Welcomed Seeking Help/Advice	04(-10)	05(02)	02(07)	06(03)	85(75)	-04(-04)	12(06)	05(20)	03(-04)
Interested in Individual Students	07(10)	13(09)	00(01)	14(07)	69(77)	-01(-09)	14(03)	08(-09)	03(09)
Accessible to Individual Students	02(-13)	-11(-11)	16(09)	09(-02)	62(43)	20(25)	08(13)	00(14)	04(07)
VI BREADTH OF COVERAGE									
Contrasted Implications	-05(02)	12(01)	05(03)	08(01)	-03(01)	72(84)	08(-03)	14(02)	08(-06)
Gave Background of Ideas/Concepts	08(03)	08(10)	16(07)	-03(-02)	02(-02)	71(78)	01(04)	11(-03)	03(03)
Gave Different Points of View	04(-06)	04(09)	11(11)	08(16)	06(01)	72(55)	07(17)	01(-06)	04(08)
Discussed Current Developments	23(29)	08(-04)	-04(-04)	05(12)	09(00)	50(48)	06(05)	16(10)	-01(-02)
VII EXAMINATIONS/GRADING									
Examination Feedback Valuable	-03(01)	08(-09)	06(-11)	09(05)	08(12)	-04(03)	72(62)	05(-03)	09(03)
Eval Methods Fair/Appropriate	06(02)	00(-03)	03(14)	07(06)	14(00)	10(17)	69(64)	11(11)	-08(04)
Tested Emphasized Course Content	08(00)	-01(04)	11(21)	01(01)	06(03)	11(-04)	70(58)	07(10)	-02(-03)
VIII ASSIGNMENTS									
Readings/Texts Valuable	-06(09)	-03(-03)	03(07)	-01(-06)	03(01)	07(-07)	31(11)	91(70)	02(04)
Added to Course Understanding	12(01)	-01(-12)	01(04)	09(21)	01(17)	-02(08)	07(05)	81(56)	06(10)
IX WORKLOAD/DIFFICULTY									
Course Difficulty (Easy-Hard)	-06(00)	06(-01)	04(-05)	-04(02)	-01(00)	08(00)	-34(08)	10(04)	85(79)
Course Workload (Light-Heavy)	14(-04)	-09(-01)	03(03)	07(05)	00(04)	06(01)	00(01)	00(-04)	88(86)
Course Pace (Too Slow - Too Fast)	-20(07)	12(00)	04(18)	-12(-09)	06(02)	-03(-07)	04(09)	05(-04)	62(32)
Hours/week Outside of Class	14(00)	07(00)	-11(10)	07(02)	00(02)	-04(03)	03(-08)	05(21)	73(46)

NOTE: Factor loadings in boxes are the loadings for items designed to measure each factor. All loadings are presented without decimal points. Factor analyses of student ratings and instructor self ratings (loadings in parentheses) consisted of a principal-components analysis, Kaiser normalization, and rotation to a direct oblimin criterion. The first nine unrotated factors for the instructor self ratings had eigenvalues of 9.5, 2.9, 2.5, 2.2, 2.0, 1.4, 1.3, 1.1 & 1.0, and accounted for 68% of the variance. For the student ratings the first nine eigenvalues were 14.9, 3.3, 2.3, 1.5, 1.2, 0.9, 0.7, 0.6 & 0.5, and accounted for 88% of the variance. The analyses were performed with the commercially available SPSS routine (See Nie, et. al., 1975).

TABLE 3

Multitrait-Multimethod Matrix: Correlations Between Student and Faculty Self Evaluations In All 331 Courses

INSTRUCTOR SELF EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS								
	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(83)								
ENTHUSIASM	29	(82)							
ORGANIZATION	12	01	(74)						
GROUP INTERACTION	01	03	-15	(90)					
INDIVIDUAL RAPPORT	-07	-01	07	02	(82)				
BREADTH	13	12	13	11	-01	(84)			
EXAMINATIONS	-01	08	26	09	15	20	(76)		
ASSIGNMENTS	24	-01	17	05	22	09	22	(70)	
WORKLD/DIFFICULTY	03	-01	12	-09	06	-04	09	21	(70)

STUDENT EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS									STUDENT EVALUATION FACTORS								
	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(46)	10	-01	08	-12	09	-04	08	02	(95)								
ENTHUSIASM	21	(54)	-04	-01	-07	-01	-03	-09	-09	45	(96)							
ORGANIZATION	17	13	(30)	-03	04	07	09	00	-05	52	49	(93)						
GROUP INTERACTION	19	05	-20	(52)	00	-02	-14	-04	-08	37	30	21	(98)					
INDIVIDUAL RAPPORT	03	03	-05	13	(28)	-19	-03	-02	00	22	35	33	42	(96)				
BREADTH	26	15	09	00	-14	(42)	00	09	02	49	34	56	17	15	(94)			
EXAMINATIONS	18	09	01	-01	06	-09	(17)	-02	-06	48	42	57	34	50	33	(93)		
ASSIGNMENTS	20	03	02	09	-01	04	-01	(45)	12	52	21	34	30	29	40	42	(92)	
WORKLD/DIFFICULTY	-06	-03	04	00	03	-03	12	22	(69)	06	02	-05	-05	08	18	-02	20	(87)

NOTE: Values in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (See Nie, et. al., 1977). Values in the diagonal of lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with Learning would be .41, .48, .25, .46, .25, .37, .13, .36, & .54. All correlation coefficients are presented without decimal point. Correlations greater than .10 are statistically significant.

TABLE 4

RELIABILITY AND CONVERGENT VALIDITY OF STUDENT AND INSTRUCTOR SELF RATINGS: SEPARATE ANALYSES FOR UNDERGRADUATE COURSES TAUGHT BY FACULTY (UF--183 CLASSES), GRADUATE LEVEL COURSES TAUGHT BY FACULTY (GF--45 CLASSES), UNDERGRADUATE COURSES TAUGHT BY TEACHING ASSISTANTS (TA--103 CLASSES), AND COMBINED DATA FOR ALL COURSES (COMB--331 CLASSES)

	RELIABILITY COEFFICIENTS								VALIDITY COEFFICIENTS			
	INSTRUCTOR SELF-RATINGS				STUDENT RATINGS							
EVALUATION FACTORS	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB
LEARNING/VALUE	.80	.87	.82	.83	.95	.96	.94	.95	.41	.20	.46	.46
INSTR ENTHUSIASM	.83	.83	.82	.82	.97	.97	.97	.96	.48	.60	.62	.54
ORGANIZATION	.79	.78	.59	.74	.93	.95	.93	.93	.28	.41	.31	.30
GROUP INTERACTION	.88	.83	.94	.90	.98	.98	.97	.98	.54	.46	.39	.52
INDIVIDUAL RAPPORT	.82	.81	.83	.82	.96	.97	.95	.96	.17	.31	.52	.28
BREADTH OF COVERAGE	.79	.72	.87	.84	.91	.96	.94	.94	.43	.06	.37	.42
EXAMS/GRADING	.77	.74	.76	.76	.93	.90	.94	.93	.15	.39	.15	.17
VALUE OF ASSIGNMENTS	.77	.64	.50	.70	.92	.94	.88	.92	.33	.20	.74	.45
WORKLOAD/DIFFICULTY	.67	.71	.72	.70	.87	.89	.88	.87	.69	.63	.69	.69
MEDIAN VALUE ACROSS ALL FACTORS	.79	.78	.82	.82	.93	.96	.94	.94	.41	.39	.46	.45
OVERALL RATINGS (SINGLE ITEMS)												
OVERALL COURSE	--	--	--	--	--	--	--	--	.27	.17	.17	.26
OVERALL INSTRUCTOR	--	--	--	--	--	--	--	--	.36	.20	.24	.33

NOTE: RELIABILITY ESTIMATES, COEFFICIENT ALPHAS (SEE NIE, ET. AL., 1977) WERE BASED UPON THE CORRELATIONS AMONG ITEMS IN THE SAME FACTOR AND COULD NOT BE COMPUTED FOR THE SINGLE ITEMS. VALIDITY COEFFICIENTS FOR THE FACTOR SCORES WERE CORRECTED FOR UNRELIABILITY WITH THE SPEARMAN BROWN EQUATION. VALIDITY COEFFICIENTS FOR THE TWO SINGLE ITEMS WERE NOT CORRECTED SINCE NO RELIABILITY ESTIMATES WERE AVAILABLE. IT SHOULD BE NOTED THAT THE VALIDITY OF THE TWO SINGLE ITEMS, THE OVERALL RATINGS, WERE LOWER THAN THE MEDIAN VALIDITY COEFICIENTS OF THE FACTORS EVEN WHEN NOT CORRECTED FOR UNRELIABILITY. THIS IS PROBABLY DUE TO THE FACT THAT INDIVIDUAL ITEMS TEND TO HAVE LOWER RELIABILITIES THAN DO FACTOR SCORES THAT ARE BASED UPON SEVERAL ITEMS.

TABLE 5

Multiple Regression Analysis of Convergent and Divergent Validity: Separate Analyses For Undergraduate Courses Taught By Faculty (UF--183 Classes), Graduate Level Courses Taught By Faculty(GF--45 Classes), Undergraduate Courses Taught By Teaching Assistants (TA--103 classes) and Combined Data For All Courses (Comb--331 Classes)

	RELIABILITY Coefficients[a]				r^2 with Matching INSTR Self Rating[b]				Mult R^2 with :[c] --Matching INSTR --All Other INSTR (Unique Var Due To Matching Rating)				Mult R^2 with :[c] --Matching INSTR --All Other STDNT (Unique Var Due To Matching Rating)				Mult R^2 WITH :[c] --Matching INSTR --All Other INSTR --All Other STDNT (Unique Var Due To Matching Rating)			
STUDENT EVALUATION FACTORS	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB
LEARNING/VALUE	95	96	94	95	12%	03%	16%	16%	12% (09%	02% 00%	17% 17%	16% 13%)	52% (04%	45% 00%	55% 01%	53% 03%)	53% (05%	61% 00%	61% 03%	54% 05%)
INSTR ENTHUSIASM	97	93	97	96	18%	29%	30%	23%	23% (14%	26% 26%	38% 17%	24% 18%)	46% (13%	59% 20%	58% 18%	50% 16%)	52% (17%	61% 21%	64% 15%	53% 17%)
ORGANIZATION	93	95	93	93	06%	13%	06%	06%	05% (03%	17% 17%	09% 03%	08% 05%)	60% (03%	55% 08%	65% 04%	59% 05%)	52% (03%	53% 12%	66% 04%	61% 05%)
GROUP INTERACTION	98	98	97	98	25%	17%	14%	20%	30% (21%	32% 15%	24% 09%	27% 20%)	45% (20%	29% 06%	43% 09%	42% 15%)	45% (18%	37% 03%	49% 10%	45% 15%)
INDIVIDUAL RAPPORT	[illegible]	97	95	96	02%	07%	21%	06%	07% (04%	00% 00%	27% 15%	11% 07%)	39% (03%	29% 03%	50% 04%	40% 04%)	41% (04%	25% 02%	52% 05%	41% 05%)
BREADTH OF COVERAGE	91	96	94	94	13%	00%	10%	14%	15% (13%	00% 00%	11% 04%	18% 11%)	52% (10%	45% 00%	46% 10%	50% 10%)	41% (11%	58% 00%	49% 10%	53% 10%)
EXAMS/GRADING	93	90	94	93	02%	10%	02%	02%	03% (02%	11% 06%	15% 03%	07% 03%)	52% (03%	50% 21%	55% 00%	50% 02%)	54% (04%	46% 24%	58% 00%	53% 04%)
VALUE OF ASSIGNMENTS	92	94	88	92	08%	02%	24%	13%	11% (06%	00% 00%	29% 15%	14% 10%)	41% (05%	48% 04%	46% 18%	45% 08%)	42% (07%	50% 01%	46% 14%	46% 10%)
WORKLOAD/DIFFICULTY	87	89	88	87	28%	25%	30%	29%	28% (26%	16% 16%	44% 18%	30% 25%)	33% (25%	49% 35%	51% 17%	35% 25%)	35% (20%	58% 57%	53% 17%	38% 13%)
MEAN ALL 9 FACTORS	94	94	93	94	13%	12%	17%	14%	15% (10%	12% 09%	24% 12%	17% 13%)	47% (09%	45% 11%	52% 09%	47% 08%)	46% (10%	50% 13%	55% 09%	49% 09%)

	RELIABILITY Coefficients[a]				r^2 with Matching STDNT Rating				Mult R^2 with : --Matching STDNT --All Other STDNT (Unique Var Due To Matching Rating)				Mult R^2 with : --Matching STDNT --All Other INSTR (Unique Var Due To Matching Rating)				Mult R^2 WITH : --Matching STDNT --All Other STDNT --All Other INSTR (UniqUE VAR Due To Matching Rating)			
INSTRUCTOR SELF EVALUATION FACTORS	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB	UF	GF	TA	COMB				
LEARNING/VALUE	80	87	82	83	13%	03%	16%	17%	11% (07%	00% 00%	22% 01%	17% 06%)	25% (08%	36% 00%	34% 15%	27% 12%)	25% (08%	33% 00%	35% 04%	31% 08%)
INSTR ENTHUSIASM	83	83	82	82	18%	29%	30%	23%	18% (18%	28% 28%	33% 27%	25% 25%)	23% (14%	33% 25%	30% 20%	26% 18%)	26% (19%	34% 33%	41% 23%	31% 24%)
ORGANIZATION	79	78	59	73	06%	13%	06%	06%	13% (07%	14% 14%	06% 06%	14% 10%)	20% (02%	48% 17%	06% 04%	16% 05%)	22% (06%	52% 12%	14% 08%	21% 10%)
GROUP INTERACTION	88	83	94	90	25%	17%	14%	21%	33% (23%	10% 08%	16% 12%	23% 20%)	26% (22%	34% 14%	26% 08%	24% 20%)	35% (22%	32% 07%	16% 10%	28% 20%)
INDIVIDUAL RAPPORT	82	81	83	82	02%	07%	21%	06%	08% (04%	15% 04%	22% 07%	10% 05%)	11% (04%	21% 04%	17% 17%	12% 07%)	20% (06%	27% 02%	22% 07%	16% 06%)
BREADTH OF COVERAGE	79	72	87	84	13%	00%	11%	14%	23% (15%	00% 00%	29% 13%	21% 16%)	22% (11%	02% 00%	26% 03%	17% 11%)	29% (17%	06% 00%	38% 12%	27% 16%)
EXAMS/GRADING	77	73	76	76	02%	10%	02%	02%	17% (05%	30% 29%	13% 01%	07% 04%)	18% (02%	10% 06%	26% 02%	16% 03%)	28% (07%	38% 28%	41% 00%	23% 07%)
VALUE OF ASSIGNMENTS	77	64	50	70	08%	02%	24%	13%	12% (08%	20% 06%	32% 22%	18% 12%)	33% (05%	27% 00%	20% 17%	26% 08%)	35% (08%	38% 01%	34% 22%	32% 12%)
WORKLOAD/DIFFICULTY	67	71	72	70	28%	25%	30%	29%	27% (27%	45% 38%	31% 24%	29% 26%)	31% (24%	23% 20%	25% 25%	30% 25%)	28% (23%	61% 53%	27% 27%	30% 25%)
MEAN ALL 9 FACTORS	79	77	76	79	13%	10%	16%	14%	18% (12%	18% 14%	23% 13%	18% 14%)	23% (10%	26% 10%	23% 12%	22% 12%)	27% (13%	36% 15%	32% 13%	27% 14%)

a--Reliability estimates, coefficient alphas (see Nie, et. al., 1977) were based upon the correlations among items within the same factor.

b--These are simple bivariate correlations(uncorrected for unreliability) that have been squared

c--Multiple correlation coefficients were computed by entering each set of items simultaneously--correcting for the number of variables in the regression equation--and then entering the one "matching variable" on the last step tc determine the proportion of variance that can be uniquely explained by it. If R squared or the change in R squared was negative (due to the adjustment for the number of variables in the equation) it was considered to be zero, and the change in R squared on the next step was the difference from zero. In some instances there was evidence of suppression in that the change in R squared was larger than the contribution of a variable by itself, and the standardized beta weights were larger than the simple correlations.

USC EVALUATION SERVICES

AS A DESCRIPTION OF THIS COURSE/INSTRUCTOR, THIS STATEMENT IS:

(SELECT THE BEST RESPONSE FOR EACH OF THE FOLLOWING STATEMENTS LEAVING A RESPONSE BLANK ONLY IF IT IS CLEARLY NOT RELEVANT)

		VERY POOR	POOR	MODERATE	GOOD	VERY GOOD
1	**LEARNING:** YOU FOUND THE COURSE INTELLECTUALLY CHALLENGING AND STIMULATING	1	2	3	4	5
2	YOU HAVE LEARNED SOMETHING WHICH YOU CONSIDER VALUABLE	1	2	3	4	5
3	YOUR INTEREST IN THE SUBJECT HAS INCREASED AS A CONSEQUENCE OF THIS COURSE	1	2	3	4	5
4	YOU HAVE LEARNED AND UNDERSTOOD THE SUBJECT MATERIALS IN THIS COURSE	1	2	3	4	5
5	**ENTHUSIASM:** INSTRUCTOR WAS ENTHUSIASTIC ABOUT TEACHING THE COURSE	1	2	3	4	5
6	INSTRUCTOR WAS DYNAMIC AND ENERGETIC IN CONDUCTING THE COURSE	1	2	3	4	5
7	INSTRUCTOR ENHANCED PRESENTATIONS WITH THE USE OF HUMOR	1	2	3	4	5
8	INSTRUCTOR'S STYLE OF PRESENTATION HELD YOUR INTEREST DURING CLASS	1	2	3	4	5
9	**ORGANIZATION:** INSTRUCTOR'S EXPLANATIONS WERE CLEAR	1	2	3	4	5
10	COURSE MATERIALS WERE WELL PREPARED AND CAREFULLY EXPLAINED	1	2	3	4	5
11	PROPOSED OBJECTIVES AGREED WITH THOSE ACTUALLY TAUGHT SO YOU KNEW WHERE COURSE WAS GOING	1	2	3	4	5
12	INSTRUCTOR GAVE LECTURES THAT FACILITATED TAKING NOTES	1	2	3	4	5
13	**GROUP INTERACTION:** STUDENTS WERE ENCOURAGED TO PARTICIPATE IN CLASS DISCUSSIONS	1	2	3	4	5
14	STUDENTS WERE INVITED TO SHARE THEIR IDEAS AND KNOWLEDGE	1	2	3	4	5
15	STUDENTS WERE ENCOURAGED TO ASK QUESTIONS & WERE GIVEN MEANINGFUL ANSWERS	1	2	3	4	5
16	STUDENTS WERE ENCOURAGED TO EXPRESS THEIR OWN IDEAS AND/OR QUESTION THE INSTRUCTOR	1	2	3	4	5
17	**INDIVIDUAL RAPPORT:** INSTRUCTOR WAS FRIENDLY TOWARDS INDIVIDUAL STUDENTS	1	2	3	4	5
18	INSTRUCTOR MADE STUDENTS FEEL WELCOME IN SEEKING HELP/ADVICE IN OR OUTSIDE OF CLASS	1	2	3	4	5
19	INSTRUCTOR HAD A GENUINE INTEREST IN INDIVIDUAL STUDENTS	1	2	3	4	5
20	INSTRUCTOR WAS ADEQUATELY ACCESSIBLE TO STUDENTS DURING OFFICE HOURS OR AFTER CLASS	1	2	3	4	5
21	**BREADTH:** INSTRUCTOR CONTRASTED THE IMPLICATIONS OF VARIOUS THEORIES	1	2	3	4	5
22	INSTRUCTOR PRESENTED THE BACKGROUND OR ORIGIN OF IDEAS/CONCEPTS DEVELOPED IN CLASS	1	2	3	4	5
23	INSTRUCTOR PRESENTED POINTS OF VIEW OTHER THAN HIS/HER OWN WHEN APPROPRIATE	1	2	3	4	5
24	INSTRUCTOR ADEQUATELY DISCUSSED CURRENT DEVELOPMENTS IN THE FIELD	1	2	3	4	5
25	**EXAMINATIONS:** FEEDBACK ON EXAMINATIONS/GRADED MATERIALS WAS VALUABLE	1	2	3	4	5
26	METHODS OF EVALUATING STUDENT WORK WERE FAIR AND APPROPRIATE	1	2	3	4	5
27	EXAMINATIONS/GRADED MATERIALS TESTED COURSE CONTENT AS EMPHASIZED BY THE INSTRUCTOR	1	2	3	4	5
28	**ASSIGNMENTS:** REQUIRED READINGS/TEXTS WERE VALUABLE	1	2	3	4	5
29	READINGS, HOMEWORK, ETC. CONTRIBUTED TO APPRECIATION AND UNDERSTANDING OF SUBJECT	1	2	3	4	5
30	**OVERALL:** HOW DOES THIS COURSE COMPARE WITH OTHER COURSES YOU HAVE HAD AT USC?	1	2	3	4	5
31	HOW DOES THIS INSTRUCTOR COMPARE WITH OTHER INSTRUCTORS YOU HAVE HAD AT USC?	1	2	3	4	5

STUDENT AND COURSE CHARACTERISTICS (LEAVE BLANK IF NO RESPONSE APPLIES)

32	COURSE DIFFICULTY, RELATIVE TO OTHER COURSES, WAS (1-VERY EASY... 3-MEDIUM ... 5-VERY HARD)	1	2	3	4	5
33	COURSE WORKLOAD, RELATIVE TO OTHER COURSES, WAS (1-VERY LIGHT... 3-MEDIUM . 5-VERY HEAVY)	1	2	3	4	5
34	COURSE PACE WAS (1-TOO SLOW... 3-ABOUT RIGHT... 5-TOO FAST)	1	2	3	4	5
35	HOURS/WEEK REQUIRED OUTSIDE OF CLASS 1) 0 TO 2. 2) 2 TO 5. 3) 5 TO 7. 4) 8 TO 12. 5) OVER 12	1	2	3	4	5
36	LEVEL OF INTEREST IN THE SUBJECT PRIOR TO THIS COURSE (1-VERY LOW... 3-MEDIUM ... 5-VERY HIGH)	1	2	3	4	5
37	OVERALL GPA AT USC 1) BELOW 2 5 2) 2 5 TO 3 0 3) 3 0 TO 3 4. 4) 3 4 TO 3 7. 5) ABOVE 3 7 LEAVE BLANK IF NOT YET ESTABLISHED AT USC	1	2	3	4	5
38	EXPECTED GRADE IN THE COURSE (1-F, 2-D, 3-C, 4-B, 5-A)	F	D	C	B	A
39	REASON FOR TAKING THE COURSE (1-MAJOR REQUIRE 2-MAJOR ELECTIVE 3-GENERAL ED REQUIRE. 4-MINOR/RELATED FIELD 5-GENERAL INTEREST ONLY) - SELECT THE ONE WHICH IS BEST	1	2	3	4	5
40	YEAR IN SCHOOL 1) FRESH. 2) SOPH. 3) JR. 4) SR. 5) GRAD.	1	2	3	4	5
41	MAJOR DEPARTMENT 1) SOC SCI/COMM. 2) NAT SCI/MATH. 3) HUMANITIES. 4) BUSINESS. 5) EDUCATION. 6) ENGINEERING. 7) PERF ARTS. 8) PUB AFFAIRS. 9)OTHER. 10) UNDECLARED/UNDECIDED	1 6	2 7	3 8	4 9	5 10

SUPPLEMENTAL QUESTIONS (USE RESPONSES BELOW FOR INSTRUCTOR'S QUESTIONS)

42	1	2	3	4	5	47	1	2	3	4	5	52	1	2	3	4	5	57	1	2	3	4	5
43	1	2	3	4	5	48	1	2	3	4	5	53	1	2	3	4	5	58	1	2	3	4	5
44	1	2	3	4	5	49	1	2	3	4	5	54	1	2	3	4	5	59	1	2	3	4	5
45	1	2	3	4	5	50	1	2	3	4	5	55	1	2	3	4	5	60	1	2	3	4	5
46	1	2	3	4	5	51	1	2	3	4	5	56	1	2	3	4	5	61	1	2	3	4	5

APPENDIX II.

Factor Analysis of Student Evaluation Instrument (N=511 Class Average Responses)

Factor Pattern Loadings

Evaluation Items (paraphrased)	Mean	Standard Deviation	I	II	III	IV	V	VI	VII	VIII	IX
I. LEARNING/VALUE											
Increased Interest as Course Consequence	3.91	0.56	69	14	-04	06	04	09	06	17	08
Learned Something Valuable	4.15	0.40	59	06	11	12	00	04	11	15	10
Learned & Understood Subject Matter	4.01	0.41	53	11	17	09	06	-09	10	12	-20
OVERALL COURSE RATING	3.83	0.66	44	23	12	07	05	07	20	17	10
Intellectually Challenging/Stimulating	3.98	0.54	43	17	03	08	-01	18	10	13	31
II. ENTHUSIASM											
Dynamic & Energetic	3.90	0.65	08	67	15	07	04	09	09	11	07
Enhanced with humor	3.85	0.65	01	67	16	00	08	06	07	10	08
Held your Interest	3.64	0.67	14	65	26	06	02	03	03	10	01
Enthusiastic about Teaching	4.10	0.57	10	48	19	07	13	14	13	09	06
OVERALL INSTRUCTOR RATING	3.97	0.65	14	43	25	10	17	11	12	08	05
III. ORGANIZATION											
Materials Prepared & Explained	3.90	0.56	12	-06	70	03	07	14	14	10	04
Instructor Explanations Clear	3.90	0.56	10	12	57	14	03	08	12	08	-07
Lectures Facilitated Note Taking	3.77	0.62	08	-02	51	-19	06	27	08	11	-03
Objectives stated & pursued	3.94	0.53	20	-10	49	06	08	12	24	11	04
IV. GROUP INTERACTION											
Students shared Ideas/Knowledge	4.07	.59	08	10	-01	81	07	04	07	08	08
Encouraged to Participate	4.05	.60	11	12	03	80	07	00	08	07	08
Encouraged to Express Own Ideas	4.09	.55	06	12	04	73	16	07	11	04	08
Encouraged to Question & Given Answers	4.08	.55	09	13	17	62	16	04	12	07	-02
V. INDIVIDUAL RAPPORT											
Welcomed Seeking Help/Advice	4.13	.54	08	10	05	06	82	-02	10	03	-01
Interested in Individual Students	4.02	.57	06	10	06	17	69	-06	14	03	00
Accessible to Students	3.91	.56	-02	08	03	01	65	24	11	11	07
Friendly towards Students	4.20	.49	00	-25	12	10	61	-08	10	06	-09
VI. BREADTH OF COVERAGE											
Presented Background of Concepts	3.97	.48	12	05	12	02	05	68	07	12	-03
Contrasted Implications	3.94	.49	06	10	03	07	08	67	01	20	04
Presented Different Points of View	4.03	.46	03	08	12	12	03	60	12	12	-03
Discussed Current Developments	4.14	.49	19	12	15	16	02	32	17	12	-06
VII. EXAMINATIONS											
Eval Methods Fair/Appropriate	3.80	.54	03	04	02	05	16	05	72	15	-06
Tested Actual Content	3.88	.55	09	02	10	02	06	09	67	14	-04
Exam Feedback Valuable	3.67	.59	03	05	09	10	16	-02	66	07	09
VIII. ASSIGNMENTS											
Readings/Text Valuable	3.72	.59	-02	-05	02	00	04	11	-01	91	02
Contributed to Understanding	3.88	.54	09	01	06	10	02	01	16	70	03
IX. WORKLOAD/DIFFICULTY											
Workload (Light-Heavy)	3.37	.61	10	02	02	08	00	00	00	08	89
Difficulty (Easy-Hard)	3.45	.52	-02	02	00	-01	-01	11	07	08	86
Hours Out of Class	2.61	.61	13	03	-10	01	10	07	-09	12	76
Pace (Too Slow-Too Fast)	3.09	.39	-09	01	11	-10	-05	12	14	04	60

1--Factor Analysis was Oblique (correlated) with the Delta Factor=-2.0 (Nie, et. al, 1975)
2--First nine eigenvalues were 10.8, 3.3, 2.3, 1.5, 1.2, 1.0, .76, .60, .50
3--Correlations between Factors ranged from r=-.01 to r=.49 (Median r=.22)
4--All items except Workload/Difficulty were answered along 5-point response scale (1-Very Poor, 3-Moderate, 5-Very Good). Workload/Difficulty items varied on 5-point response scale with end-points above, except for Hours (1-0 to 2, 2-2 to 5, 3-5 to 7, 4-8 to 12, 5-Over 12).

APPENDIX III

RELIABILITY

Evaluation Items	ANOVA Reliability Estimates For Class[1] Averages Based Upon Different Numbers of Responses						Coefficient Alpha[2] Reliability Estimates of Factor Scores
	5	10	15	25	50	100	
I. LEARNING/VALUE							.95
Increased Interest as Course Consequence	.52	.69	.77	.83	.91	.96	
Learned Something Valuable	.55	.71	.78	.86	.92	.96	
Learned & Understood Subject Matter	.50	.67	.78	.85	.92	.95	
OVERALL COURSE RATING	.62	.76	.83	.89	.94	.96	
Intellectually Challenging/Stimulating	.64	.78	.84	.90	.95	.97	
II. ENTHUSIASM							.97
Dynamic & Energetic	.70	.83	.88	.92	.96	.98	
Enhanced with Humor	.69	.81	.87	.92	.96	.98	
Held Your Interest	.67	.80	.86	.91	.96	.97	
Enthusiastic About Teaching	.66	.79	.85	.91	.95	.97	
OVERALL INSTRUCTOR RATING	.66	.80	.85	.91	.95	.97	
III. ORGANIZATION							.93
Materials Prepared & Explained	.58	.74	.81	.88	.93	.97	
Instructor Explanations Clear	.60	.75	.82	.88	.94	.97	
Lectures Facilitated Note Taking	.60	.75	.82	.88	.94	.97	
Objectives Stated and Pursued	.51	.68	.76	.84	.91	.97	
IV. GROUP INTERACTION							.98
Students Shared Ideas/Knowledge	.64	.78	.84	.90	.95	.97	
Encouraged to Participate	.65	.79	.85	.90	.95	.97	
Encouraged to Express Own Ideas	.61	.76	.82	.89	.94	.97	
Encouraged to Question & Given Answers	.60	.75	.82	.88	.94	.97	
V. INDIVIDUAL RAPPORT							.95
Welcomed Seeking Help/Advice	.57	.72	.80	.87	.93	.96	
Interested in Individual Students	.57	.73	.80	.87	.93	.96	
Accessible to Students	.52	.69	.77	.85	.92	.96	
Friendly Toward Students	.57	.73	.80	.87	.93	.96	
VI. BREADTH OF COVERAGE							.93
Presented Background of Concepts	.55	.71	.78	.86	.92	.96	
Contrasted Implications	.52	.69	.77	.85	.92	.96	
Presented Different Points of View	.50	.67	.75	.83	.91	.95	
Discussed Current Developments	.56	.71	.79	.86	.94	.97	
VII. EXAMINATIONS							.94
Evaluation Methods Fair/Appropriate	.58	.74	.81	.88	.93	.97	
Tested Actual Content	.58	.74	.81	.88	.93	.97	
Exam Feedback Valuable	.59	.74	.81	.88	.94	.97	
VIII. ASSIGNMENTS							.90
Readings/Text Valuable	.63	.77	.84	.90	.94	.97	
Contributed to Understanding	.50	.67	.75	.83	.91	.95	
IX. WORKLOAD/DIFFICULTY							.88
Workload (Light-Heavy)	.60	.75	.82	.88	.94	.97	
Difficulty (Easy-Hard)	.52	.69	.77	.85	.92	.97	
Hours Out of Class	.55	.71	.78	.86	.92	.96	
Pace (Too Slow-Too Fast)	.36	.52	.62	.73	.85	.92	
MEDIAN RELIABILITY	.58	.74	.81	.88	.93	.97	.94

[1] --Anova Reliability estimates were obtained by taking 10 responses from each of 387 courses in which at least 15 students responded. A one-way Anova was performed in which the courses served as levels. The reliability estimate for 10 responses was computed by subtracting the reciprocal of the F-Ratio from 1.0. The other estimates were generated with the Spearman-Brown equation. This procedure is described in Winer (1971), Marsh (1976) and Centra (1973).

[2] --Coefficient Alphas were computed with Method 2 described by Nie, et. al. (1977).

Two types of reliability are presented above. The Anova reliability estimates measure the relative consistency within each class relative to the differences between different classes. The principle source of error measured by this technique is the diversity of student opinion within the courses. It should be noted that this is a more stringent criteria than would be measured by assessing the reliability of individual responses. Using the Spearman-Brown equation, the median reliability for a sample size of one would be r = .22. However, using a test-retest procedure over a three year interval, Overall and Marsh (1978) found that reliabilities of the responses of individual students were generally over .50.

The coefficient alpha reliability is based upon the degree of intercorrelation among the items defining each factor. This value will also vary with the number of responses. The average number of responses in the 511 courses used in this analysis was 26.7. (Avg. Enrollment was 34.56, Avg. Response Rate was 77%). The median reliability of the factor scores is substantially higher than the median reliability of individual items based upon a comparable number of responses. This is due, at least in part, to the greater reliability of an average.

Appendix IVa

CONVERGENT AND DISCRIMINANT VALIDITY: CORRELATIONS BETWEEN INSTRUCTOR SELF-EVALUATIONS AND STUDENT EVALUATIONS FOR ALL UNDERGRADUATE LEVEL COURSES TAUGHT BY FACULTY (N=183 COURSES)

INSTRUCTOR SELF EVALUATION FACTORS	LEARN	ENTHU	ORGAN	GROUP	INDIV	BROTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(80)								
ENTHUSIASM	32	(83)							
ORGANIZATION	18	03	(79)						
GROUP INTERACTION	04	-02	-21	(88)					
INDIVIDUAL RAPPORT	03	-09	16	-02	(82)				
BREADTH	-08	05	26	-05	07	(79)			
EXAMINATIONS	03	00	23	-04	20	20	(77)		
ASSIGNMENTS	23	00	29	09	31	18	.8	(77)	
WORKLD/DIFFICULTY	07	03	15	-09	10	-06	21	21	(67)

STUDENT EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS									STUDENT EVALUATION FACTORS								
	LEARN	ENTHU	ORGAN	GROUP	INDIV	BROTH	EXAMS	ASIGN	WRKLD	LEARN	ENTHU	ORGAN	GROUP	INDIV	BROTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(41)	12	02	10	-11	-08	-08	07	08	(95.								
ENTHUSIASM	23	(48)	-07	-09	-07	-09	-23	-11	01	51	(97)							
ORGANIZATION	17	11	(28)	-16	03	06	04	07	01	51	46	(93)						
GROUP INTERACTION	17	08	-23	(54)	-12	-19	-20	-06	-03	38	33	15	(98)					
INDIVIDUAL RAPPORT	07	02	-08	10	(17)	-20	-08	-12	04	23	35	35	39	(96)				
BREADTH	10	09	13	-08	-15	(43)	04	04	02	40	33	61	14	13	(91)			
EXAMINATIONS	14	09	-03	-05	-05	-15	(15)	-01	05	50	36	56	30	53	32	(93)		
ASSIGNMENTS	17	03	08	10	-14	05	06	(33)	12	49	25	43	28	24	45	46	(92)	
WORKLD/DIFFICULTY	-03	-04	03	03	00	00	21	15	(69)	20	08	01	04	06	18	04	23	(87)

NOTE: Values in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (See Nie, et. al., 1977). Values in the diagonal of lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with Learning would be .36, .43, .24, .50, .15, .36, .13, .28, & .53. All correlation coefficients are presented without decimal point. Correlations greater than .145 are statistically significant.

Appendix IVb

CONVERGENT AND DISCRIMINANT VALIDITY: CORRELATIONS BETWEEN INSTRUCTOR SELF-EVALUATIONS AND STUDENT EVALUATIONS FOR ALL GRADUATE LEVEL COURSES TAUGHT BY FACULTY (N=45 COURSES)

INSTRUCTOR SELF EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(87)								
ENTHUSIASM	10	(83)							
ORGANIZATION	40	08	(78)						
GROUP INTERACTION	-15	-30	-34	(53)					
INDIVIDUAL RAPPORT	-32	22	01	16	(81)				
BREADTH	-20	-02	-15	34	26	(72)			
EXAMINATIONS	10	20	37	-11	09	13	(74)		
ASSIGNMENTS	46	23	19	09	16	04	16	(64)	
WORKLD/DIFFICULTY	08	00	33	-11	03	04	07	26	(71)

STUDENT EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS									STUDENT EVALUATION FACTORS								
	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(20)	07	08	00	-17	05	-07	01	-30	(96)								
ENTHUSIASM	10	(60)	07	-02	16	-08	09	14	-27	55	(97)							
ORGANIZATION	11	12	(41)	02	05	09	01	-01	-22	57	56	(95)						
GROUP INTERACTION	-10	-23	-30	(46)	02	-04	-34	-36	-29	11	-01	10	(98)					
INDIVIDUAL RAPPORT	-19	05	-08	08	(31)	-09	-16	-16	-14	29	35	22	45	(97)				
BREADTH	20	-01	05	14	-03	(06)	-17	19	-20	50	38	59	24	36	(96)			
EXAMINATIONS	10	02	20	20	23	-04	(39)	11	-12	43	42	54	19	40	48	(90)		
ASSIGNMENTS	-02	12	-03	24	20	02	-13	(20)	-12	65	56	47	24	36	36	48	(94)	
WORKLD/DIFFICULTY	15	05	11	-09	05	14	17	26	(63)	11	21	17	-21	09	40	21	07	(89)

NOTE: Values in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (See Nie, et. al., 1977). Values in the diagonal of lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with Learning would be .18, .54, .35, .41, .27, .05, .32, .15, & .50. All correlation coefficients are presented without decimal point. Correlations greater than .29 are statistically significant.

Appendix IVc

CONVERGENT AND DISCRIMINANT VALIDITY: CORRELATIONS BETWEEN INSTRUCTOR SELF-EVALUATIONS AND STUDENT EVALUATIONS FOR ALL UNDERGRADUATE LEVEL COURSES TAUGHT BY TEACHING ASSISTANTS (N=103 COURSES)

	INSTRUCTOR SELF-EVALUATION FACTORS								
INSTRUCTOR SELF EVALUATION FACTORS	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(82)								
ENTHUSIASM	27	(82)							
ORGANIZATION	-04	-06	(59)						
GROUP INTERACTION	02	22	07	(94)					
INDIVIDUAL RAPPORT	-06	05	-09	05	(83)				
BREADTH	35	21	06	31	-16	(87)			
EXAMINATIONS	-11	19	27	43	09	23	(76)		
ASSIGNMENTS	06	-15	-04	-06	15	-18	-06	(50)	
WORKLD/DIFFICULTY	-20	-14	-03	-11	06	-18	-10	11	(72)

	INSTRUCTOR SELF-EVALUATION FACTORS									STUDENT EVALUATION FACTORS								
STUDENT EVALUATION FACTORS	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(46)	-02	-04	08	-03	14	07	-11	-16	(94)								
ENTHUSIASM	24	(62)	00	11	02	13	36	-16	-25	37	(97)							
ORGANIZATION	22	17	(31)	14	04	14	22	-14	-07	63	54	(93)						
GROUP INTERACTION	28	08	-02	(39)	25	23	17	-02	-18	44	37	41	(97)					
INDIVIDUAL RAPPORT	06	07	08	18	(52)	-19	18	22	-01	23	36	38	45	(95)				
BREADTH	28	24	12	08	-08	(37)	-02	-07	-12	43	39	60	12	18	(94)			
EXAMINATIONS	29	13	05	-04	15	00	(15)	-13	-22	56	52	63	45	49	38	(94)		
ASSIGNMENTS	15	-13	06	-05	32	-17	-08	(74)	04	32	-07	17	19	39	11	31	(88)	
WORKLD/DIFFICULTY	-41	-11	09	-04	15	-32	-10	28	(69)	47	-26	30	-30	11	-16	36	04	(84)

NOTE: Values in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (See Nie, et. al., 1977). Values in the diagonal of lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with learning would be .40, .55, .23, .37, .46, .33, .13, .49, & .55. All correlation coefficients are presented without decimal point. Correlations greater than .19 are statistically significant.

Appendix V

Absolute and Relative Agreement Between Student Evaluations of Teaching(STD) and the corresponding Instructor Self Evaluations(INS); N= 331 classes--183 undergraduate courses taught by faculty, 45 graduate level courses taught by faculty, and 103 undergraduate courses taught by teaching assistants

Evaluation Items (paraphrased)	Undergrad Faculty				Graduate Faculty				Undergrad TA*s				All Courses Comb			
	STD	FAC	DIFF	CORR	STD	FAC	DIFF	CORR	STD	FAC	DIFF	CORR	STD	FAC	DIFF	CORR
I LEARNING/VALUE																
Course Challenging/Stimulating	4.1	3.9	.19**	.32**	4.2	4.0	.27**	.31**	3.7	3.3	.44**	.31**	4.0	3.7	.27**	.39**
Learned something valuable	4.2	4.0	.16**	.18**	4.4	4.2	.17	.10	3.9	3.6	.30**	.28**	4.1	3.9	.20**	.26**
Increased Subject Interest	4.0	3.9	.05	.23**	4.2	4.1	.16	.13	3.6	3.6	.08	.35**	3.9	3.8	.07	.30**
Learned/Understood Subject Matter	4.0	3.7	.27**	.32**	4.2	3.8	.38**	.11	3.8	3.5	.36**	.43**	4.0	3.6	.31**	.35**
OVERALL COURSE RATING	3.9	3.9	.02	.27**	4.2	3.9	.25	.17	3.6	3.5	.10	.17*	3.8	3.8	.08	.26**
II ENTHUSIASM																
Enthusiastic about teaching	4.2	4.2	.05	.26**	4.3	4.2	.11	.09	4.1	4.0	.12	.29**	4.2	4.1	.08	.27**
Dynamic & Energetic	3.9	3.9	.13	.22**	4.1	3.8	.32**	.45**	3.9	3.7	.17*	.55**	3.9	3.8	.11*	.35**
Enhanced Presentations with Humor	3.8	3.4	.44**	.31**	4.1	3.5	.65**	.67**	3.7	3.5	.13	.49**	3.8	3.4	.37**	.39**
Teaching Style Held Your Interest	3.7	3.7	.00	.25**	3.9	3.7	.23	.16	3.6	3.5	.11	.29**	3.7	3.6	.06	.25**
OVERALL INSTRUCTOR RATING	4.0	4.0	.05	.36**	4.3	4.0	.26*	.19	3.9	3.7	.20*	.24*	4.0	3.9	.13*	.33**
III ORGANIZATION																
Instructor Explanations Clear	3.9	3.9	.01	.22**	4.1	3.9	.19	.14	3.9	3.9	.00	.00	3.9	3.9	.03	.13**
Course Materials Prepared & Clear	4.0	4.0	-.08	.19**	4.1	3.8	.30*	.22	3.9	3.9	-.03	.38**	3.9	3.9	-.01	.24**
Objectives Stated & Pursued	4.0	4.1	-.09	.02	4.1	4.0	.09	.27*	3.9	4.0	-.11	.14	4.0	4.0	-.07	.10*
Lectures Facilitated Note Taking	3.8	3.6	.18*	.21**	4.5	4.4	.08	.28*	3.8	3.8	.05	.40**	3.8	3.6	.19**	.24**
IV GROUP INTERACTION																
Encouraged Class Discussions	4.1	4.2	-.09	.39**	4.4	4.3	.08	.29*	4.0	4.2	-.16	.15	4.1	4.2	-.09	.32**
Students Shared Ideas/Knowledge	4.1	4.0	.06	.39**	4.5	4.2	.23	.37**	4.0	4.0	.00	.43	4.1	4.1	.06	.40**
Encouraged Questions & Answers	4.1	4.2	-.09	.23**	1.4	4.4	.02	.13	4.0	4.2	-.15	.25**	4.1	4.2	-.09	.24**
Encouraged Expression of Ideas	4.1	4.1	-.02	.26**	4.4	4.2	.18	.27*	4.1	4.1	-.02	.39**	4.1	4.1	.00	.30**
V INDIVIDUAL RAPPORT																
Friendly Towards Students	4.2	4.3	-.11	.22**	4.4	4.4	.02	.27*	4.3	4.3	.02	.35**	4.3	4.3	-.05	.25**
Welcomed Seeking Help/Advice	4.1	4.3	-.16*	.05	4.3	4.3	.03	.19	4.2	4.4	-.23**	.48**	4.2	4.3	-.15**	.18**
Interested in Individual Students	4.0	4.3	-.27**	.26**	4.3	4.4	-.14	.34*	4.0	4.3	-.33**	.38**	4.0	4.3	-.27**	.31**
Accessible to Individual Students	4.0	4.1	-.06**	.21**	4.3	4.1	.19	.01	4.0	4.3	-.29**	.32**	4.1	4.2	-.10**	.19**
VI BREADTH OF COVERAGE																
Contrasted Implications	4.1	3.9	.12	.35**	4.2	4.0	.20	.06	3.8	3.4	.34**	.27**	4.0	3.8	.20**	.35**
Gave Background of Ideas/Concepts	4.1	4.0	.09	.19**	4.1	4.1	.03	.20	3.7	3.4	.35**	.32**	4.0	3.8	.16**	.31**
Gave Different Points of View	4.1	4.1	.03	.14*	4.2	3.9	.32	.05	3.9	3.8	.10	.16	4.1	4.0	.09*	.16**
Discussed Current Developments	4.1	4.2	-.06	.18**	4.3	4.2	.09	.32*	3.7	3.4	.26**	.36**	4.1	4.0	.06	.39**
VII EXAMINATIONS/GRADING																
Examination Feedback Valuable	3.6	3.7	-.06	.18**	3.9	3.7	.19	.35**	3.8	3.8	-.04	.04	3.7	3.7	-.02	.18**
Eval Methods Fair/Appropriate	3.8	4.2	-.38**	.01	4.2	4.1	.05	.03	3.8	4.0	-.21**	.14	3.9	4.1	-.27**	.05
Tested Emphasized Course Content	3.9	4.2	-.27**	.14*	4.1	3.9	.19	.24	3.9	4.1	-.18*	.11	3.9	4.1	-.18**	.14**
VIII ASSIGNMENTS																
Readings/Texts Valuable	3.8	4.0	-.12	.27**	4.2	4.2	.04	.11	3.7	3.7	-.04	.51**	3.8	3.9	-.07	.34**
Added to Course Understanding	3.9	4.0	-.08	.28**	4.3	4.2	.13	.24	3.7	3.7	.02	.27**	3.9	3.9	-.02	.32**
IX WORKLOAD/DIFFICULTY																
Course Difficulty (Easy-Hard)	3.5	3.6	-.03	.35**	3.6	3.6	.02	.38	3.1	3.3	-.18*	.40**	3.5	3.4	-.04	.41**
Course Workload (Light-Heavy)	3.4	3.4	.00	.50**	3.6	3.6	-.05	.45**	3.2	3.1	-.07	.48**	3.4	3.4	.03	.50**
Course Pace (Too Slow - Too Fast)	3.1	3.0	.09	.13*	3.1	3.1	.02	.01	3.2	3.1	.14*	.02	3.0	3.1	.09	.10*
Hours/week Outside of Class	2.7	2.9	-.25**	.34**	3.2	3.7	.50**	.35**	2.5	2.6	-.10	.24**	2.7	2.9	-.24**	.41**
MEDIAN CORRELATION FOR 35 ITEMS				.23				.22				.31				.30

----

* p ≤ .05, ** p ≤ .01

NOTE: Two-tailed statistical tests were used in determining absolute agreement (mean differences that appear under the columns labeled "DIFF") since it was assumed that student ratings might be either higher or lower than the instructor self evaluations. One tailed statistical tests were used to test relative agreement (the correlations under the columns labeled "CORR") since it was assumed that the correlations would all be positve.

APPENDIX VI

CORRELATIONS BETWEEN STUDENT/INSTRUCTOR/COURSE CHARACTERISTICS AND STUDENT EVALUATIONS OF TEACHING EFFECTIVENESS (THE VALUES NOT IN PARENTHESES), AND FACULTY SELF-EVALUATIONS OF THEIR OWN TEACHING (THE VALUES IN PARENTHESES). N=1[illegible] UNDERGRADUATE COURSES.

STUDENT/COURSE/INSTRUCTOR BACKGROUND VARIABLES	EVALUATION FACTORS AND OVERALL SUMMARY ITEMS										
	LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	OVER CRSE	OVER INSTR
STUDENTS RATING "STUDENT'S PRIOR SUBJECT INTEREST" (1-LOW.............5-HIGH)	41 (22)	25 (20)	08 (-05)	36 (23)	16 (-12)	07 (-10)	15 (-19)	22 (-06)	20 (00)	34 (16)	28 (15)
FACULTY RATING "STUDENT'S PRIOR SUBJECT INTEREST" (1-LOW.............5-HIGH)	25 (38)	21 (28)	13 (-01)	16 (08)	14 (00)	04 (-24)	09 (-02)	11 (11)	-04 (03)	25 (29)	22 (12)
STUDENTS RATING "COURSE WORKLOAD/DIFFICULTY" (HIGHER SCORES DENOTE MORE DIFFICULT COURSES)	20 (-03)	08 (-04)	01 (03)	04 (03)	06 (00)	18 (00)	04 (21)	23 (15)	100 (53)	26 (17)	16 (09)
FACULTY RATING "COURSE WORKLOAD/DIFFICULTY" (HIGHER SCORES DENOTE MORE DIFFICULT COURSES)	06 (07)	02 (03)	01 (15)	-03 (-09)	04 (10)	02 (-06)	05 (21)	12 (21)	53 (100)	15 (29)	08 (16)
STUDENTS RATING "EXPECTED COURSE GRADE" (1-F...............5-A)	28 (11)	20 (-03)	05 (-07)	38 (17)	16 (-10)	01 (-11)	28 (-11)	24 (02)	-25 (-19)	26 (-01)	27 (00)
FACULTY RATING THEIR "GRADING LENIENCY" (1-EASY/LENIENT GRADER....5-HARD/STRICT GRADER)	-04 (00)	-16 (04)	-06 (06)	06 (16)	-08 (14)	-05 (08)	-05 (32)	-02 (19)	26 (28)	-06 (14)	-10 (03)
STUDENTS "% INDICATING INTEREST AS REASON FOR TAKING CRSE" (ACTUAL PERCENTAGE)	06 (09)	10 (06)	10 (12)	-10 (-13)	-10 (-07)	21 (10)	03 (10)	14 (-08)	18 (-12)	09 (18)	06 (09)
"COURSE ENROLLMENT" (ACTUAL NUMBER OF STUDENTS ENROLLED)	-24 (-02)	-04 (03)	-13 (10)	-36 (-43)	-21 (-17)	-09 (-03)	-22 (-03)	-09 (-11)	-07 (-04)	-18 (-04)	-20 (-09)
"PERCENT OF FRESHMEN & SOPHMORES IN CLASS" (ACTUAL PERCENTAGE)	-21 (-12)	-11 (-03)	-05 (15)	-36 (-27)	-19 (05)	-05 (04)	-13 (06)	-10 (-01)	-10 (04)	-17 (-05)	-19 (00)
FACULTY "NUMBER OF TIMES HAVE TAUGHT THIS COURSE" (ACTUAL NUMBER OF TIMES)	-04 (05)	06 (09)	10 (20)	-15 (-19)	00 (15)	05 (11)	-09 (-03)	-11 (11)	00 (-04)	-02 (03)	03 (17)
FACULTY "YEARS TEACHING IN HIGHER EDUCATION" (ACTUAL NUMBER OF YEARS)	-08 (09)	-04 (-10)	-06 (12)	00 (-04)	13 (12)	04 (08)	-10 (04)	04 (17)	-02 (00)	-08 (05)	-01 (05)
FACULTY RATING THEIR OWN "POPULARITY WITH STUDENTS" (1-EXTREMELY UNPOPULAR....5-EXTREMELY POPULAR)	29 (34)	37 (37)	31 (13)	17 (07)	09 (-01)	17 (03)	17 (-07)	17 (05)	01 (02)	35 (32)	38 (48)
FACULTY RATING SELF AS "TEACHER IN UNDERGRADUATE CLASSES" (1-WELL BELOW AVG .....5-WELL ABOVE AVG)	31 (30)	42 (42)	30 (40)	-03 (-12)	00 (08)	19 (04)	16 (16)	06 (13)	10 (-05)	31 (25)	37 (48)
FACULTY RATING "ENJOY TEACHING RELATIVE TO OTHER DUTIES" (1-EXTRMLY UNENJOYABLE...5-EXTRMLY ENJOYABLE)	25 (24)	34 (39)	18 (01)	22 (10)	33 (12)	00 (-21)	20 (-20)	09 (03)	03 (-03)	29 (15)	32 (22)
FACULTY RATING " EASE OF TEACHING THIS PARTICULAR COURSE" (1-VERY EASY.....5-VERY DIFFICULT)	07 (-12)	-01 (-16)	10 (-07)	11 (17)	06 (12)	09 (06)	09 (05)	01 (04)	05 (17)	03 (-14)	08 (-10)
FACULTY RATING "SCHOLARLY PRODUCTION IN THEIR DISCIPLINE" (1-WELL BELOW AVG....5-WELL ABOVE AVG)	12 (28)	02 (20)	18 (40)	04 (09)	06 (11)	21 (26)	04 (25)	17 (25)	11 (10)	14 (40)	16 (41)
FACULTY RATING "STUDENT EVALS ARE ACCURATE ASSESSMENT OF TEACHING" (1-STRONGLY DISAGREE.....9-STRONGLY AGREE)	41 (17)	38 (26)	28 (04)	27 (00)	27 (00)	16 (-09)	24 (-12)	16 (-02)	09 (04)	(16)	42 (26)
FACULTY RATING "STUDENT EVALS POTENTIALLY USEFUL FEEDBACK TO FACULTY" (1-STRONGLY DISAGREE....9-STRONGLY AGREE)	27 (17)	34 (22)	14 (02)	29 (-01)	27 (-07)	05 (-23)	18 (-15)	08 (-12)	-01 (07)	33 (11)	31 (15)
CORRELATIONS BETWEEN STUDENT AND FACULTY RATINGS OF THE SAME EVALUATION SCORES	41	48	26	48	17	43	15	33	69	27	36

NOTE: EACH OF THE SET OF 18 STUDENT/COURSE/INSTRUCTOR CHARACTERISTICS WERE OBTAINED FROM THE STUDENT EVALUATION SURVEY FORM, THE FACULTY SELF-EVALUATION FORM, OR THE REGISTRAR'S LISTING OF CLASSES. EACH OF THESE VARIABLES WAS THEN CORRELATED WITH THE 11 EVALUATIONS OF TEACHING EFFECTIVENESS (9 FACTOR SCORES AND THE TWO OVERALL SUMMARY ITEMS). SEPARATE SETS OF CORREFATIONS WERE COMPUTED FOR CLASS-AVERAGE STUDENT EVALUATIONS AND FACULTY SELF-EVALUATIONS (VALUES IN PARENTHESES).

NOTE: CORRELATION COEFFICIENTS ARE PRESENTED WITHOUT DECIMAL POINTS. CORRELATIONS GREATER THAN .15 ARE STATISTICALLY SIGNIFICANT