DOCUMENT RESUME

ED 206 635 TM 810 498

AUTHOR Marsh, Herbert W.; Hocevar, Dennis
TITLE Confirmatory Factor Analysis and Anova Analyses of Multitrait-Multimethod Matrices.
PUB DATE [80]
NOTE 48p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Comparative Analysis; Evaluation Methods; *Factor Analysis; Self Evaluation (Individuals); Statistical Significance; Student Evaluation of Teacher Performance; Teacher Effectiveness; Test Validity
IDENTIFIERS *Confirmatory Factor Analysis; *Multitrait Multimethod Techniques

ABSTRACT

Campbell and Fiske (1959) developed four criteria of construct validity when measures of more than one trait are obtained with more than one method. In this study these criteria are compared with two other procedures--an analysis of variance (ANOVA) model and confirmatory factor analysis--for analyzing multitrait-multimethod (MTMM) data. The principle advantage of the ANOVA model is a convenient summary and test of convergent, divergent and method/halo effects. However, the limitations of this approach are even more numerous than those encountered with the Campbell-Fiske criteria, and so the ANOVA approach should only be used to supplement other procedures. Confirmatory factor analysis provides a direct test of the statistical significance and importance of various trait and method factors. The size of factor loadings provide a convenient description of the magnitude of method and trait effects. By constraining various parameters the researcher may formulate and test alternative configurations of method and trait factors. Consequently, confirmatory factor analysis offers the advantages of both the other approaches without many of their limitations, and is the recommended procedure for analyzing MTMM data. (Author/BW)

Confirmatory Factor Analysis and Anova Analyses

of Multitrait - Multimethod Matrices

Herbert W. Marsh

University of Sydney, Australia

Dennis Hocevar

University of Southern California

Running Head: Multitrait-Multimethod

Abstract

Campbell and Fiske (1959) have developed four criteria of construct validity when measures of more than one trait are obtained with more than one method. In this study these criteria are compared with two other procedures--an ANOVA model and Confirmatory Factor Analysis--for analyzing multitrait-multimethod (MTMM) data. Despite important limitations of the Campbell-Fiske criteria, the usefulness of interpretations based upon the criteria, the heuristic value of their application, and the popularity of the method all dictate that it continue to be used as a preliminary inspection of MTMM matrices. The principle advantage of the ANOVA model is a convenient summary and test of convergent, divergent and method/halo effects. However, the limitations of this approach are even more numerous than those encountered with the Campbell-Fiske criteria, and so the ANOVA approach should only be used to supplement other procedures. Confirmatory factor analysis provides a direct test of the statistical significance and importance of various trait and method factors. The size of factor loadings provide a convenient description of the magnitude of method and trait effects. By constraining various parameters the researcher may formulate and test alternative configurations of method and trait factors. Consequently, confirmatory factor analysis offers the advantages of both the other approaches without many of their limitations, and is the recommended procedure for analyzing MTMM data.

Confirmatory Factor Analysis and ANOVA Analyses

of Multitrait - Multimethod Matrices

Campbell and Fiske (1959) have advocated the assessment of validity by obtaining measures of more than one trait, each of which is assessed by more than one method. In the present example the different traits are nine dimensions of evaluations of instructional effectiveness: the different methods of assessing the traits are student ratings of teaching effectiveness and instructor ratings of their own teaching effectiveness. Convergent validity, that which is most typically determined, is the agreement between measures of the same trait assessed by two different methods--student-faculty agreement on evaluations of teaching. Discriminant validity refers to the distinctiveness of each of the trait-factors.

Determination of convergent and discriminant validity is based upon inspection or analysis of a multitrait-multimethod matrix such as the one shown in Table 1 (considering only the coefficients below the main diagonal of the entire 18 x 18 matrix at this point). Correlations between different traits assessed by the same method appear in monomethod-heterotrait (the upper left and lower right) blocks of the matrix. Correlations between different traits assessed by different methods are in the heteromethod-heterotrait (lower left) blocks of the matrix. The convergent validity coefficients, correlations between the same traits assessed by different methods appear in the heteromethod-monotrait diagonal of this matrix--the values in <> in Table 1. It is also valuable to have the reliabilities of each measure in the diagonals of the heterotrait-monomethod matrices--the values in parentheses in Table 1. Campbell and Fiske (1959) proposed four criteria for assessing convergent and divergent validity:

1) The convergent validity coefficients should be statistically significant and sufficiently different from zero to warrant further examination of validity. Failure of this test indicates that the different methods are measuring different constructs and implies a lack of validity in at least one of the methods.

2) The convergent validities should be higher than the correlations between different traits assessed by different methods. The failure of this test implies that agreement on a particular trait is not independent of agreement on other traits, perhaps suggesting that the agreement can be explained in terms of a generalized agreement that encompasses more than one (or all) of the traits.

3) The convergent validities should be higher than correlations between different traits assessed by the same method. If the convergent validities are not substantially higher, there is the suggestion that the traits may be correlated, that there is a method effect, or some combination of both these possibilities. If the correlations between different traits assessed by the same method approach the reliabilities of the traits, then there is evidence of a strong halo or method bias.

4) The pattern of correlations between different traits should be similar for each of the different methods. Satisfaction of this criterion—assuming that there are significant correlations among traits—would suggest that the underlying traits are truly correlated. Failure to meet this criterion implies that the observed correlation between traits assessed by a given method is due to a method or halo bias.

Despite the intuitive appeal of the Campbell-Fiske criteria, there are numerous potential problems in their application. Although many of these were anticipated by Campbell and Fiske, solutions were not offered. Perhaps recognizing the dangers in the precise formulation of their criteria, these authors stated that the development of statistical treatments might be unnecessary or inappropriate.

An obvious problem with the Campbell-Fiske criteria is the lack of specification as to what constitutes satisfactory results. The application to be presented in this paper, for example, involves nine traits, each assessed by two methods. Testing the second and third criteria alone requires that each of the nine convergent validities be compared with 32 different correlations--a total of 288 comparisons. Besides being unwieldy, the likelihood of obtaining rejections due to sampling fluctuations alone increases geometrically with the number of traits and methods. The user is left with the task of determining either the proportion of failures or some average difference between the convergent validities and coefficients against which they are to be compared. In either case, the decision as to what constitutes a failure is arbitrary.

An even more serious ambiguity exists in the criteria used to assess discriminant validity. At least conceptually, Campbell and Fiske make clear distinctions between method variance, trait variance, and trait covariation. Method variance--the introduction of systematic variation due to a specific method of data collection--is clearly detrimental to discriminant validity, though it does not preclude the demonstration of either divergent or convergent validity. True trait variance (i.e., convergent validity)--the correlation between different methods of assessing the same trait that is independent of method variance--is obviously good, but it does not imply discriminant validity. True trait covariation--the true correlation between different traits that does not depend upon the method of data collection--will increase the likelihood of failures in the application of the second and third criteria. However, the fourth criterion specifically tests for true trait covariation, and its demonstration is taken as support for discriminant validity. A complete lack of trait covariation makes interpretation more simple, but is unlikely to exist in any but the most contrived of situations (e.g., attitudes toward cigarette smoking and capital punishment). Trait correlations approaching unity can be unambiguously interpreted as a complete lack of discriminant validity. For most applications, however, some low to moderate true trait covariation is likely, and its interpretation is left ambiguous.

The most serious problem with the Campbell-Fiske criteria is that they are based upon inspection of correlations between observed variables, but make inferences about underlying trait and method factors. The validity of any set of interpretations depends upon the behavior of the underlying constructs. This can be illustrated with the problem of systematically differing reliabilities. Application of the criteria implicitly assumes, as recognized by Campbell and Fiske, that each of the measures are equally reliable. If there are substantial differences in the reliabilities of different traits, or in the measures obtained with different methods, then failures of one or more of the criteria may be a function of the differential reliabilities alone. For example, if traits assessed by one method are systematically more reliable than those assessed by a second method, then the correlations among traits assessed with the more reliable method will be higher, and give the appearance of a method effect. Some authors have suggested that the multitrait-multimethod matrix be corrected for attenuation (Heberlein, 1969; Althauser & Heberlein, 1970).

Similarly, the Campbell-Fiske criteria also assume that convergent validities reflect the effect of shared trait variance. While this is true, the convergent validity coefficients can also be affected by shared method variance or a trait-method interaction. Furthermore, the existence of shared method variance or trait-method interactions may act to either artificially increase or decrease the observed validity coefficient. A more detailed discussion of the implications of these underlying inferences is presented by Alwin (1974).

Since the development of the Campbell-Fiske criteria for assessing the multitrait-multimethod matrix, a variety of specific statistical tests have been developed (Althauser & Heberlein, 1970; Alwin, 1974; Joreskog, 1974; Kavanagh, MacKinney & Wolins, 1971; Kenny, 1979; Lomax & Algina, 1974; Schmitt, 1978; Schmitt, Coyle & Saari, 1977; Werts & Linn, 1970). In the present study two of these procedures are applied, and their limitations are illustrated. The first is an analysis of variance technique that was presented by Kavanagh, et al. (1971), while the second is a variety of confirmatory factor analysis models as elaborated by Schmitt (1978).

In the present study, the multitrait-multimethod approach was used to validate students' evaluations of teaching effectiveness. Instructors in 329 college classrooms were asked to evaluate their own teaching effectiveness on the same nine-trait instrument as their students. Previous application (Marsh, in press; Marsh & Overall, 1979; Marsh, Overall & Kesler, 1979) of the Campbell-Fiske criteria left several questions unanswered. In spite of evidence for both convergent and divergent validity, there was the suggestion of a moderate method variance--particularly with the student ratings. However, confounding this suggestion were the facts that: 1) the student ratings were more reliable than the instructor ratings (perhaps explaining the higher correlations among the student ratings), and, 2) the likelihood that the correlations among the traits (instructional evaluation factors) were true correlations rather than method or halo bias. The purpose of this study is to compare the conclusions based upon Campbell-Fiske criteria with those obtained from two alternative analytic procedures, and to discuss advantages and disadvantages of the approaches.

Method

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 a staff person. The surveys were completed by an average of 76% (a range of from 54% to 100%) of the students enrolled in each class.

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. Instructors 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 confidentiality 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 instructors (78%) returned self evaluations from 331 courses; ratings of 183 undergraduate courses taught by faculty, 45 graduate level courses, and 103 courses taught by teaching assistants.

The evaluation instrument consisted of 35 items that were designed to measure 9 traits. Previous research, based upon a different sample of 511 undergraduate classes taught by regular faculty, determined the reliability of the evaluation factors (median alpha = .94), confirmed the existence of the nine evaluation dimensions, and provided weights that were used in calculating factor scores (See Marsh, in press; Marsh & Overall, 1979). The evaluation factor scores used in the present study were weighted averages, the weights having been derived from the previous factor analysis, of standardized responses to each item. The evaluation trait-factors and a brief description are as follows:

LEARNING/VALUE--The extent to which students felt they encountered a valuable learning experience that was intellectually challenging.

INSTRUCTOR ENTHUSIASM--The extent to which students perceived the instructor to display enthusiasm, energy, humor and an ability to hold interest

ORGANIZATION--The instructor's organization of the course, course materials, and class presentations.

GROUP INTERACTION--Students' perceptions of the degree to which the instructor encouraged class discussions and invited students to share their own ideas or to be critical of those presented by the instructor.

INDIVIDUAL RAPPORT--The extent to which students perceived the instructor to be friendly, interested in students, and accessible in or out of class.

BREADTH OF COVERAGE--The extent to which students perceived the instructor to present alternative approaches to the subject and to emphasize analytic ability and conceptual understanding.

EXAMINATIONS--Students' perceptions of the value and fairness of graded materials in the course.

ASSIGNMENTS--The value of class assignments (readings, homework, etc) in adding appreciation and understanding of the subject.

WORKLOAD/DIFFICULTY--Students' perceptions of the relative difficulty, workload, pace of presentations, and the number of hours required by the course.

Separate factor analyses were performed on the student and instructor self evaluations for the 329 classes included in this study (Marsh, in press; Marsh & Overall, 1979). This analysis was performed to determine if similar evaluation trait-factors underlie both the student and instructor self evaluations, and if these were similar to results previously obtained for a different sample of student ratings. Factor analyses of both student and instructor ratings confirmed the existence of the same nine trait-factors that had been previously identified. Each item, for both student and instructor evaluations, loaded highest on the factor it was designed to measure. Loadings for items defining each factor generally exceeded .50, and all other loadings were typically less than .20. Furthermore, the factor loadings from both these analyses were quite similar to those previously obtained with a different population of student evaluations. The 315 factor loadings (35 items loading on each of 9 factors) for the factor analysis of instructor ratings considered in this study correlated .90 with both the 315 factory loadings obtained for student evaluations in this study and those obtained with a previous factor analysis of a different sample of student evaluations; the two sets of 315 loadings from the two factor analyses of the student ratings correlated .95 with each other. These findings justify the assumption that similar evaluation trait-factors underlie both the student and instructor evaluations.

Results

Campbell-Fiske Criteria

Application of the Campbell-Fiske criteria discussed earlier requires a visual inspection of the multitrait-multimethod matrix presented in Table 1. One of the limitations of the use of these criteria, as indicated by Campbell & Fiske (1959), is the implicit assumption that the trait reliabilities obtained with different methods are comparable. This is clearly not the case in the present example, since student evaluations (based upon class average responses) are consistently more reliable. Coefficient alphas (see Table 1) for student ratings vary from .87 to .98 (median .94), while those for the instructor self evaluations vary from .70 to .90 (median .82). Consequently, for each of the correlations presented in Table 1, the same correlation corrected for attenuation is also presented. Interpretation of the Campbell-Fiske criteria is discussed in terms of both corrected and uncorrected correlations.

The first Campbell-Fiske criterion requires that convergent validity coefficients be statistically significant and high enough to warrant further consideration of validity. Each of the convergent validity coefficients presented in Table 1 is statistically significant, and they are substantial (median r = .45, corrected for attenuation).

Insert Table 1 About Here

The second Campbell-Fiske criterion requires that each convergent validity coefficient be higher than any other correlation in the same row or column of the same heterotrait-heteromethod block. This test requires that each of the nine convergent validity coefficients be compared to each of 16 other coefficients—a total of 144 comparisons in all. Data presented in Table 1 satisfy this criterion for 143 of the 144 comparisons (for both corrected and uncorrected correlations), providing good support for this aspect of discriminant validity.

The third criterion requires that each convergent validity be higher than correlations between that trait and any other trait assessed by the same method. Application of this criterion to the uncorrected data indicates only 4 rejections (out of 72 comparisons) for the instructor self evaluations. For the student evaluations, however, there are 30 rejections (also out of 72 comparisons). On the surface, this would seem to suggest a method or halo effect for the student ratings, though little for the instructor self evaluations. However, this interpretation is biased by the fact that the student ratings are consistently more reliable than the instructor ratings. Correlations involving only student ratings are least attenuated, while those involving only instructor ratings are most attenuated. Consequently, relative to the convergent validities, correlations between student ratings are systematically increased and correlations among instructor ratings are systematically decreased. When all correlations are corrected for attenuation, however, this criterion is still not met in 27 comparisons involving the student evaluations and only 5 with the instructor self evaluations. The correction for attenuation decreased the apparent method effect and lessened the difference in method effect between student and instructor ratings, but these changes were small.

The fourth criterion requires that the pattern of correlations among different traits should be similar for the different methods. A visual inspection of Table 1 suggests that this may be the case. To provide a more precise test, the 36 off-diagonal coefficients in the student rating block were correlated with those in the instructor rating block. The result, r = .43, was significant at the .01 level and suggests that there is a similarity in the pattern of correlations. This suggests that there is true trait covariation that is independent of method.

In summary, the data provide clear support for convergent validity, and at least two of the criteria of discriminant validity. Student-instructor agreement on any one trait was independent of their agreement on other traits. Furthermore, there was a similarity in the pattern of trait correlations for student and instructor ratings. There was an indication, however, of some halo or method effect—particularly with the student ratings.

The ANOVA Approach

Based upon recent citations in the literature, this technique appears to have been popularized by Kavanagh, MacKinney, and Wolins (1971). Stanley (1961) demonstrated how multitrait-multimethod data could be analyzed with a three-factor unreplicated analysis of variance; when repeated measurements of cases--ratings of college classes in the present application--are measured over all levels of two other variables--traits and methods in this case--three orthogonal sources of variation can be estimated. The main effect due to classes is a test of how well ratings in general discriminate between classes, and is suggested to be analogous to convergent validity. It should be noted that this is NOT the same use of convergent validity as that discussed by Campbell and Fiske (1959). The interaction between classes and traits tests whether the differentiation between classes depends upon traits. If it does not, then the traits have no differential validity (i.e., each class is ranked the same regardless of the trait). This is taken to be a measure of discriminant validity. The interaction between classes and methods tests whether the differentiation between classes depends upon methods. If it does, then the different methods introduce a source of systematic (undesirable) variance. This is taken to be a measure of method or halo effect. The class by trait by method interaction is assumed to measure only random error (i.e., the differentiation between classes is assumed not to depend upon any specific trait-method combination). Stanley (1961) recommends that the measures be replicated for each subject within a given study, thus providing independent estimates of the three way interaction and the error term (also see King, Hunter & Schmidt, 1980). However, his recommendation does not seem to ever been followed. In this model main effects due to traits and methods can also be calculated, but these are generally of less interest.

Boruch, Larkin, Wolins and MacKinney (1970) and Kavanagh, MacKinney and Wolins (1971) have described computational procedures whereby the mean squares and the variance component estimates for the analysis of variance model could be computed directly from the correlations contained in the multitrait-multimethod matrix. The computational equations for computing these effects are presented in Table 2. The systematic differences in the reliabilities of student and instructor ratings, as previously discussed, will produce biased estimates of the discriminant validity and method/halo effects (Boruch, Larkin, Wolins & MacKinney, 1970; Schmitt, et al., 1977). Consequently, the ANOVA procedure was also applied to the correlations that were corrected for unreliability (see Table 1).

Each of the ANOVA effects--Convergent Validity, Divergent Validity, and Method/Halo bias--and their variance components are presented in Table 2. All three effects are statistically significant for analyses based upon both the corrected and uncorrected correlation coefficients. The size of the discriminant validity effect (the variance component) was approximately twice that of the method/halo effect. When the correlation coefficients were corrected for attenuation, each of the effects--except the error term--increased. However, the largest increase occurred for the discriminant validity effect. As was observed with the Campbell-Fiske analysis, the correction for attenuation improved the discriminant validity, but did not eliminate the method/halo bias.

Insert Table 2 About Here

The principal advantages of the ANOVA model are its ease of application and the convenient descriptive statistics summarizing the relative magnitude of the effects of convergent validity, divergent validity, and method/halo bias. However, the model also has major shortcomings. The problem of differing reliabilities, which this approach shares with the Campbell-Fiske analysis, has already been discussed. The assumption that the class by method by trait interaction contains only error variance is not normally testable, and its violations may have varying influences on the estimation of the other effects. The model makes no provision for the possibility of true trait covariation or correlated method effects, and provides no test for their existence. Finally, many of the heuristic inferences that are likely to result from the application of the Campbell-Fiske criteria will be lost with application of only the ANOVA analysis. Many of the disadvantages of the ANOVA model are shared with the Campbell-Fiske analysis, but the misleading precision and simplicity of the ANOVA approach are less likely to reveal these potential problems.

There is no clear equivalence between the effects estimated by the ANOVA model and the Campbell-Fiske criteria. Inspection of the computational equation for the convergent validity effect (see Table 2), indicates that it is a function of the average correlation in the entire multitrait-multimethod matrix. This is clearly different from the Campbell-Fiske criterion that is based upon just the convergent validity diagonal. In particular, even if all the convergent validity coefficients approached unity, the average correlation in the entire matrix generally would not. Similarly, the ANOVA model might indicate a moderate degree of convergent validity even if the average convergent validity coefficient were close to zero.

The similarity of the divergent and method/halo effects in the ANOVA model and the Campbell-Fiske criteria is harder to assess. Inspection of the computational equations for the ANOVA effects (Table 2) indicates that the discriminant validity and method/halo effects are a function of the difference between the average of specified correlations and the average correlation in the entire MTMM matrix. The comparisons in the Campbell-Fiske criteria are more specific. Furthermore, the proportion of variance accounted for by the four effects in the ANOVA model—the convergent, divergent, method/halo, and error effects—must sum to 1.0. This means that an increase in the convergent effect will cause a decrease in the divergent effect so long as the method/halo and error effects remain constant. This is quite different from the Campbell-Fiske approach where an increase in convergent validity will lead to an increase in discriminant validity. Similarly, when correlations in the present application were corrected for attenuation, the Campbell-Fiske analysis indicated that the Method effect was reduced (i.e., fewer rejections of criterion 3), but that the method effect in the ANOVA analysis actually increased—though the increase was less than the increase in the divergent validity effect. The ANOVA model has no term that is comparable to the fourth Campbell-Fiske criterion. In fact the ANOVA model is based upon the assumption that traits are uncorrelated (see King, et al., 1980) but provides no test of this assumption. These observations indicate that comparisons between the ANOVA and Campbell-Fiske analyses should be made cautiously.

In summary, application of the ANOVA model indicates significant effects of convergent, divergent and method/halo effects. The size of the discriminant validity effect (the variance component) was more than twice the size of either of the other two effects. The variance component for this effect was also increased the most by the correction for attenuation.

Confirmatory Factor Analysis

The confirmatory factor analysis approach is described under a variety of different labels in the literature: restricted factor analysis (Boruch & Wolins, 1970), confirmatory factor analysis (Werts, Joreskog & Linn, 1972), path analysis (Schmitt, Coyle & Saari, 1977; Schmitt, 1978), and exploratory factor analysis (Lomax & Algina, 1979). This plethora of labels, and particularly the emphasis on path analysis (and structural equations) is unfortunate. The analysis of the MTMM can be viewed as a straightforward application of confirmatory factor analysis with a priori factors corresponding to specific traits and methods, and the major findings can be interpreted in much the same way as can any other factor analysis.

The Confirmatory Factor Analysis Model. In this study the notation, the specification of the model, and the actual analysis are performed with the commercially available LISREL IV program (Joreskog & Sorbom, 1978). This program embodies Joreskog's maximum-likelihood approach to confirmatory factor analysis. The model used in this analysis requires the specification of three different matrices.[1] These are the LAMBDA matrix that contains the factor loadings, the PSI matrix that contains the correlations between the factors, and the THETA matrix that contains the error/uniqueness of each measured variable. These are conceptually similar to the rotated factor matrix, the matrix of correlations between factors, and the communalities (actually one minus the communalities) that result from common factor analysis. In confirmatory factor analysis, however, the researcher is able to constrain various parameters in the different matrices in order to test alternative models. On the basis of these three matrices, a reproduced correlation matrix is determined that provides a "best fit" to the original correlation matrix within the constraints that are imposed by the proposed model. Using matrix notation SIGMA, the reproduced correlation matrix is defined as:

SIGMA = [LAMBDA * PSI * LAMBDA '] + THETA EPSILON

In the present example, the configuration for the factor loading (LAMBDA) matrix and the matrix of correlations between factors (the PSI matrix) is presented in Table 3.

Insert Table 3 About Here

In the LAMBDA matrix, each of the 11 factors (Eta 1 - Eta 11) represents either a Method factor (Eta 1 & Eta 2), or a Trait factor (Eta 3 - Eta 11). The first method factor is defined by the nine instructor self evaluations (Ilrn, Ient,...,Iwrk), while the second method factor is defined by the nine student ratings (Slrn, Sent,..., Swrk). Each of the nine trait factors is defined by the one instructor and one student rating of the same trait. For example, the first trait factor (Eta 3) is the learning trait factor and is defined by the instructor and student ratings of Learning. Each of the "0" elements in the matrix represents a fixed parameter, while the other 36 elements are free and will be estimated.

In most of the models to be discussed—with some notable exceptions, the factors are oblique (correlated). The correlations among the 11 factors appear in the PSI matrix (see Table 3). Each of the elements in the PSI matrix represents a correlation between two factors; for example, r10.11 represents the correlation between the Assignment factor (Eta 10) and the Workload/Difficulty factor (Eta 11). Elements of the matrix that begin with r were free and estimated by the program; the "0" elements were fixed to be zero; and the diagonals were fixed to be 1.0.

The LISREL program attempts to minimize a maximum-likelihood loss function that is based upon differences between the original and reproduced correlation matrices, and provides an overall chi-square test of the goodness-of-fit of the proposed model. As described by Joreskog (Joreskog & Sorbom, 1978), it also determines a test of identification, asymptomatically efficient estimates of each free parameter in the proposed model under the assumptions of multivariate normality, estimates of the standard error of each fitted parameter--allowing a statistical test of its difference from zero, and additional information that is helpful in determining what changes in the proposed model would provide a better fit to the data (see Maruyama & McGarvey, 1980, for further discussion).

The minimum condition for fitting the complete model (Alwin, 1974; Werts, et al., 1972) is that there be at least three traits and three methods. This means that, without making any further assumptions (i.e., constraining more parameters to a fixed value), the most unrestricted form of the model is not identified and cannot be tested. On the basis of both substantive (Boruch & Wolins, 1970) and practical (Althauser & Herberlein, 1970) considerations, the correlations between traits and methods were set to zero. However, the model was still not identified.[2] In order to obtain a testable model, the reliability of the student and instructor ratings (coefficient alphas based upon the items that define each of the factors) were computed and used as a basis for determining the values of THETA (error/uniqueness components). Preliminary analysis indicated that this resulted in a very poor fit to the data, suggesting that each factor may have a unique component as well as error. Consequently, the 18 variables were entered into a standard factor analysis procedure (Nie, et al., 1975) and an 11 factor solution was determined. The Communalities resulting from this analysis (see Nie, et al., 1975, pp. 475-477) were then used to determine an estimate of the THETA elements. This procedure, which provides an estimate of the combined uniqueness and unreliability, provided a much better fit to the data. Consequently in order to circumvent the identification problem, all the THETA elements were set at a value of 1 minus the communality of the variable. This same set of values was used for each of the models to be discussed. Consequently, the most general model to be considered in this study is one in which correlations between methods and traits are fixed to be zero, and the values of THETA (error/uniqueness components) are predetermined.

<u>The Goodness of Fit of the Model</u>. The LISREL program provides a chi-square test of the overall goodness-of-fit, but the test is dependent upon the sample size. A reasonably good fit to the data will produce a statistically significant chi-square value if the sample size is large, while a poor fit based upon a small sample size may not result in a statistically significant chi-square value. Alternative indices of fit (Schmitt, 1978) include the ratio of the chi-square to the degrees of freedom, the average difference between the reproduced and original correlation matrix, and a reliability coefficient developed by Tucker and Lewis (1973). The reliability coefficient is defined as:

r = (Co - Cm)/ (Co - 1) Where:

Co = the chi-square/df ratio for a null model,

Cm = the chi-square/df ratio for the tested model,

1 = the expected value of the chi-square/df ratio

This coefficient scales the chi-square goodness-of-fit value along a scale that varies from zero (the null model) to 1.0, though values greater than 1.0 are possible. The null model generally consists of specifying SIGMA to be a diagonal matrix, testing the assumption that the measured variables are uncorrelated. Tucker and Lewis suggest a value of .90 or higher provides an adequate fit to the data. Their coefficient provides an index of the proportion of the variance that is explained by the model rather than a statistical test of its goodness-of-fit. For example, a model that is tested with a small number of cases (e.g., less than 50 cases) may result in statistically insignificant differences from the observed data (based upon the chi-square test) and yet only have a Tucker-Lewis reliability coefficient of .50. This suggests that while the proposed model fits the data in a statistical significance sense, the test was a very weak one and there may be many possible models that would do as well. Alternatively, a model that is tested with a large number of cases may have a Tucker-Lewise reliability of .99 and still have a significant chi-square value (see Bentler & Bonett, 1980, for further discussion).

The estimated parameters for the general model (Model I) described in this section are presented in Table 4. The chi-square value for this model is statistically significant, but the chi-square/df ratio was only 2.38 and the Tucker-Lewis reliability coefficient is .98. This indicates a good fit to the data.

Insert Table 4 About Here

Inspection of the values suggest that each of the nine trait factors is well defined, that there is substantial method variance associated with the student ratings and some associated with instructor self-evaluations, and that the traits are moderately correlated.

Testing Alternative Models. Comparisons of two tested models can be made by taking the difference in their two chi-square values and testing this against the difference in the degrees of freedom (Bentler & Bonett, 1980; Kenny, 1976; Schmitt, et al., 1977). For example, one of the alternative formulations of Model I postulated that the 36 correlations between the nine trait factors (in the PSI matrix) are really zero (Model V in Table 5). Analysis of this model produced a chi-square value (543.6 with 134 degrees of freedom--see Table 5) that was necessarily larger than the value obtained with Model I (233.6 with 98 degrees of freedom); the two chi-squares would only be equal if the estimated parameters in Model I were exactly equal to zero. Since the difference in the two chi-square values (310.0) assessed against the difference in degrees of freedom (36) is statistically significant and substantial, the analysis argues for Model I.

In order to make more precise tests of the data, a series of alternative models were derived and their ability to fit the data (using the Tucker-Lewis coefficient as an index) was examined. These models are summarized in Table 5--including the general and null models--along with their chi-squares, degrees of freedom, chi-square/df ratios, and Tucker-Lewis reliabilities. Alternative models considered the consequences of eliminating one or more of the trait factors, eliminating one or both of the method factors, or constraining some of the correlations between these factors to be zero. For example, the student method factor was eliminated (Model III in Table 5) by setting all the factor loadings for this factor (the Eta 2 factor in the LAMBDA matrix) equal to zero and setting all the correlations (in the PSI matrix) involving this factor--including the diagonal element--equal to zero. However, this model provides a poorer fit to the data than Model I. Similarly, the elimination of the instructor method factor also produces a poorer fit than does the general model, but a better fit than when the student method factor was eliminated. This shows that the student method factor is more important than the instructor method factor.

Insert Table 5 About Here

In summary the analyses of these alternative models indicates that:

(1) Substantial portions of the variance in the data were accounted for by both the different traits and the different methods. However, exclusion of the trait factors was far more detrimental to the fit of the model than was exclusion of the method factors.

(2) The elimination of correlations among the traits produced a poorer fit to the data, indicating that the underlying traits considered in this study are truly correlated.

(3) While there was substantial method variance in both the student and the instructor ratings, elimination of the student method factor was more detrimental than was elimination of the instructor method factor. This indicates that there is more method variance in the student ratings than in the instructor self evaluations.

A classic problem in factor analysis is the determination of the number of factors. Researchers typically resort to heuristic guidelines. In the present application, a precise statistical test is used to explore the consequences of combining two or more factors (see Joreskog, 1974). The Organization and Breadth of Coverage trait-factors were consistently among the most highly correlated in each of the different models (e.g., see

PSI matrix in Table 4). Furthermore, these two factors seem conceptually related as well. Consequently, an eight-trait solution was tested that combined these two factors. This was accomplished by eliminating the Organization trait-factor, and allowing the Organization items to load on the Breadth of Coverage factor. However, the results of this model (Model X--see Table 5) produced a substantially poorer fit to the data than did the nine-trait model. This implies that the best description of the data requires all nine trait factors, or at least that these two should not be combined. The ability to test the statistical and practical impact of combining traits offers an important advantage for the confirmatory factor analysis approach, particularly when research does not begin with a well established factor structure.

Descriptive Statistics. The values in Table 4 can also be used to derive descriptive statistics similar to those obtained with the ANOVA model, and to assess the adequacy of each of the measures separately. Loadings in the LAMBDA matrix can be interpreted in much the same way as with common factor analysis; high loadings of items on a trait or method factor supports the existence of the factor. Trait and method variance components for the general model (as depicted in Table 4) can be estimated by squaring the factor loadings in the LAMBDA matrix (Joreskog, 1974), and are presented in Table 6.

The trait variance in every measure, both student and faculty ratings, was substantial and statistically significant. The average trait variance across all measures was approximately twice that of the average method variance. The trait variance in the student ratings was somewhat higher than for the faculty self evaluations. However, the faculty self evaluations had little method variance (except for the Learning/Value factor), while that observed with the student ratings was substantial. One factor, Learning/Value, had substantial method variance for both student and instructor ratings. For instructor ratings of Learning/Value, there was substantially more method variance that trait variance. Similarly, there was more method variance in the student ratings of Examinations than there was trait variance.

Insert Table 6 About Here

It must be emphasized that evidence for the existence of a particular trait or method should be based upon the size of the factor loadings in the LAMBDA matrix (e.g., Table 4) or the variance components based upon these loadings (Table 6). Some researchers (e.g., Schmitt, et al., 1977) have incorrectly suggested that support for the discriminant validity should be based upon the correlations among the trait-factors (in the PSI matrix) rather than the factor loadings. However, significant correlations in the PSI matrix merely means that the underlying trait-factors are correlated in a manner that is independent of the method of data collection. This situation is actually related to the fourth Campbell-Fiske criterion (that the pattern of correlations among traits is similar for each of the different methods), and they interpret this as evidence supporting the discriminant validity of the measures. As with the interpretation of other oblique factor analyses, it is only when correlations between traits become extreme that the researcher need be concerned about the distinctiveness of the different factors. As in the present application, the correlations among factors may be quite consistent with the substantive nature of the data.

Application of the matrix equation (equation 1) or the equivalent tracing rule (Schmitt, 1978; Kenny, 1979) allows the decomposition of each reproduced correlation into components that are due to trait variation, method variation, and trait-method interactions. As previously discussed, one of the limitations of both the Campbell-Fiske and ANOVA techniques is that they make inferences about latent or unobserved variables that are based upon observed relationships. For example, the true trait variation in the convergent validities may be systematically increased or decreased, depending upon the influence of the method or trait-method interactions. A computational equation for decomposing each reproduced correlation into distinct components is presented in Table 7. Application of this decomposition for each of the reproduced correlations indicated that there was very little method variation in any correlations other than the correlations among the student ratings.

Insert Table 7 About Here

Summary of the Confirmatory Factor Analysis Approach. The analysis of MTMM matrices can be viewed as an application of confirmatory factor analysis. The matrices upon which this analysis based--except for the constraints used to define various models--are familiar to users of factor analysis, and the interpretation of the results is similar to the interpretation of common factor analyses. However, the ability to constrain various parameters allows the formulation and testing of various descriptions of the latent trait and method factors. The "goodness of fit" of the various models and their parameter estimates (e.g., factor loadings) provide a direct test of the existence of various trait and method factors.

Discussion

The purpose of this study was to compare different techniques for analyzing multitrait-multimethod matrices. In particular, the conclusions based upon the Campbell-Fiske criteria were compared with those generated by the ANOVA model and the set of confirmatory factor analysis models. At the most general level each of the different approaches showed good support for both the convergent and divergent validity, but also indicated some method or halo bias. The Campbell-Fiske criteria, through inspection, showed that agreement on any one trait was relatively independent of agreement on other traits (criterion 2), that the method variance was more pronounced in the student ratings (criterion 3), and that there was evidence of trait covariation that was independent of method (criterion 4). The ANOVA model indicated that the variance component for the divergent validity effect was approximately twice that for the method/halo effect. Confirmatory factor analysis provided précise tests of each of the observations generated by the Campbell-Fiske criteria, provided a statistical summary similar to that generated by the ANOVA model, and also estimated separate method and trait variance components for each of the different measures. Confirmatory factor analysis also provided tests of additional hypotheses that were not testable with either the Campbell-Fiske or the ANOVA approaches.

As previously discussed, there are several important limitations of the Campbell-Fiske approach to analysis of multitrait-multimethod matrices. The most important are: 1) the informal nature of criteria and the lack of clear statements of what constitutes satisfactory results; 2) the inability to provide and incorporate information about the reliability of of the measures (particularly if reliability estimates are not available); 3) the cumbersome and unwieldy number of comparisons that must be made for large problems; 4) the ambiguity between trait variance, trait covariance, and method variance; 5) the reliance on observed variables for making speculations about latent factors; and 6) the lack of any meaningful summary statistics that describe the data.

Despite these problems, the Campbell-Fiske criteria performed well in the present application. Each of the descriptive speculations based upon this analysis were confirmed with the more rigorous tests of alternative LISREL models. The approach, while lacking rigor, does provide an important initial assessment of convergent and discriminant validity, and method/halo biases. The popularity of the method, the ease of its application, the heuristic appeal of the criteria, and the usefulness of interpretations all dictate that these criteria continue to be used for the preliminary inspection of any multitrait-multimethod matrix.

The limitations with the ANOVA model, though perhaps less apparent, are more numerous than those encountered with the Campbell-Fiske analysis. The principal advantage in the use of this approach is that it provides a convenient summary of the relative magnitude of trait and method effects and a test of their statistical significance. However, the appropriateness of the test and the summary depend upon many of the same underlying assumptions that were discussed with the Campbell-Fiske analysis, and the detailed inspection of the multitrait-multimethod matrix required by the Campbell-Fiske approach will often provide an indication of problems that may be overlooked in the deceptively simple summary statistics resulting from the ANOVA analysis. Finally, many of the heuristic speculations that result in the application of the Campbell-Fiske criteria will be lost if only the ANOVA model is used. For example, application of the Campbell-Fiske criteria indicated that there was considerably more method/halo effect in the student ratings than in the instructor ratings, that there was true trait covariation among the different traits that was independent of method, and that the correction for unreliability reduced the method/halo effect in the student ratings. None of these findings could have been identified by the ANOVA model to analyze multitrait-multimethod matrices. It does, however, provide useful summary statistics that can supplement the Campbell-Fiske criteria.

The limitations in the application of the LISREL models stem primarily from the difficulty of use. Paul Lohnes (1979, p. 334), an influential researcher and textbook author in the application of quantitative analysis, recently stated that "LISREL is a complex and expensive fitting and testing machine to which the author does not have access." The key points seem to be the complexity, the expense, and the lack of availability. The LISREL program is commercially available for a rather nominal charge, so availability is not a critical problem. Complexity represents a large initial hurdle that must be overcome, in much the same way that the complexity of multiple regression was a limitation of its of its application before the publication of the Draper & Smith (1966) text. Similarly, the complexity of LISREL will become less of a problem as the technique becomes more widely known and applied. The expense--in terms of computer time--is an important limitation that probably will not be easily resolved. While many finite problems--the kind that are likely to appear in textbooks--can be solved with small amounts of computer time, exploration of large scope problems quickly become very expensive. This will be a particularly important limitation to the novice user who may be forced to use considerable amounts of computer time in formulating the problem.

Beyond these general difficulties in using LISREL, its application to analysis of multitrait-multimethod data also imposes other limitations. In order to test a model with free parameters for all of the off-diagonal values in the PSI matrix (correlations between the factors) and the THETA matrix (the uniqueness/error variances) a minimum of three traits and three methods are needed. However, as demonstrated in this study, a variety of constraints can be imposed that allow testing of an alternative models. Even when there are an adequate number of traits and methods, it is necessary to have a large number of cases in order to provide strong tests of alternative models and to obtain high Tucker-Lewis reliability coefficients. This is particularly important when the researcher sequentially develops alternative models on the basis of prior analysis of the same data. This problem, taking advantage of chance variation that may be specific to the particular data being considered, is not unique to this analysis, and the best control for the problem is to cross-validate the findings.

Despite these limitations, confirmatory factor analysis is clearly the superior method to use in the analysis of multitrait-multimethod data. In summary, some of its advantages are:

1) it tests inferences that are based upon the underlying latent variables rather than relationships between observed variables;
2) it distinguishes variance due to traits and methods;
3) it allows comparison of a variety of alternative formulations of the basic model and an overall test of the goodness-of-fit

for each proposed model;

4) it provides a separate statistical test of each estimated parameter against the null hypothesis of a zero coefficient;

5) it provides convenient summary statistics of the amount of trait and method variance in each separate measure, in each set of measures, and for all the data combined;

6) it allows the decomposition of each reproduced correlation in components that are attributable to trait and method effects;

7) it provides estimates of the reliability of each measure that are incorporated into the analysis;

8) it provides an empirical test for the existence of correlations among traits, among methods, and between traits and methods;

9) it provides an empirical test of the number of trait-factors and method-factors that provide the best fit to the data.

These advantages, particularly when compared to those of alternative techniques, demonstrate the importance of using LISREL modeling in the analysis of multitrait-multimethod data.

Reference Notes

Heberlin, T.A. The Correction for Attenuation and the Multitrait-Multi-
method Matrix. Unpublished Master's thesis, University of Wisconsin, 1969.

References

Althauser, R.P. & Heberlein, T.A. Validity and the multitrait-multimethod matrix. In E.F. Borgatta & G.W. Bohrnstedt (Eds.). Sociological Methodology 1970. San Francisco: Jossey-Bass, 1970.

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, 1974.

Bentler, P.M. & Bonett, D.G. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 1980, 88, 588-606.

Boruch, R.F., Larkin, J.D., Wolins, L. & MacKinney, A.C. Alternative methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 1970, 30, 833-853.

Boruch, R.F. & Wolins, L. A procedure for estimation of trait, method, and error variance attributable to a measure. Educational and Psychological Measurement, 1970, 30, 547-574.

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

Draper, N., & Smith, H. Applied Multiple Regression Analysis. New York: Wiley, 1966.

Joreskog, K.G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Joreskog, K.G. Structural analysis of covariance and correlation matrices. Psychometrika, 1978, 43, 443-477.

Joreskog, K.G. A general method of analysis of covariance structures. Biometrika, 1970, 57, 409-426.

Joreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In R.C. Atkinson, D.H. Krantz, R.D. Suppes (Eds.) Contemporary Developments in Mathematical Psychology, Volume 1. San Francisco: W.H. Freeman & Co. 1974, 1-56.

Joreskog, K.G. & Sorbom, D. LISREL IV: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood. Chicago: International Educational Services, 1978.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 1971, 75, 34-49.

Kenny, D.A. An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. Journal of Experimental Social Psychology. 1976, 12, 247-252.

Kenny, D.A. Correlation and Causality. New York: Wiley, 1979.

King, L.M., Hunter, J.E. & Schmidt, F.L. Halo in a multidimensional forced-choice performance evaluation scale. Journal of Applied Psychology, 1980, 65, 507-516.

Lohnes, P.E. Factorial modeling in support of causal inference. American Educational Research Journal, 1979, 16, 323-340.

Lomax, R.G. & Algina, J. Comparison of two procedures for analyzing multitrait multimethod matrices. Journal of Educational Measurement, 1975, 16, 177-186.

Marsh, H.W. Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology, (in press)

Marsh, H.W. & Overall, J.U. Validity of students' evaluations of teaching effectiveness: A comparison with instructor self evaluations by teaching assistants, undergraduate faculty and graduate faculty. Paper presented at the Annual Meeting of the American Educational Research Association, April, 1979.

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 students. Journal of Educational Psychology, 1979, 71, 149-160.

Maruyama, G. & McGarvey, B. Evaluating causal models: An application of maximum-likelihood analysis of structural equations. Psychological Bulletin, 1980, 87, 502-512.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. & Bent, D.H. Update to Statistical Package for the Social Sciences. New York: McGraw-Hill, 1977.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. & Bent, D.H. Statistical Package for the Social Sciences. New York: McGraw Hill, 1975.

Stanley, J.C. Analysis of unreplicated three way classifications with application to rater bias and trait independence. Psychometrika, 1961, 26, 205-219.

Schmitt, N. Path analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 1978, 2, 157-173.

Schmitt, N., Coyle, B.W. & Saari, B.B. A review and critique of analyses of multitrait-multimethod matrices. Multivariate Behavioral Research, 1977, 12, 447-478.

Tucker, L.R. & Lewis, C.A. A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 1973, 38, 1-10.

Werts, C.E., Joreskog, K.G. & Linn, R.L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement. 1972, 32, 655-678.

Werts, C.E. & Linn, R.L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Footnotes

The authors wish to acknowledge William McGarvey and Robert Cudeck for their comments on an earlier draft of this paper, and for their help in the application of LISREL.

1--The most general model and each of the alternative models could also be specified in terms of x-variables instead of y-variables. Other specifications of the most general model (e.g., permitting correlated errors, etc.) are also possib . The particular specification used in this study is the one most generally used by other researchers.

2--A necessary, but not sufficient, condition of identification is that there are at least as many observed correlations as free parameters. This is not a sufficient condition, since there may be overriding constraints (Kenny, 1979). The LISREL program, however, checks for identification (See Joreskog, 1978; Joreskog & Sorbom, 1978; for further discussion) and generates an error message when the proposed model is not identified.

TABLE I

Multitrait-Multimethod Matrix: Correlations Between Student and Faculty Self Evaluations in 329 Courses

INSTRUCTOR SELF EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS: LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	STUDENT EVALUATION FACTORS: LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	(830)	286	117	14	-70	127	-8	243	30	<405>	214	171	192	29	256	183	207	-58
ENTHUSIASM	347	(820)	10	30	-19	124	80	-8	-10	99	<476>	132	46	31	149	89	26	-29
ORGANIZATION	149	13	(740)	-147	72	129	262	167	116	-12	-41	<254>	-203	-53	87	17	18	40
GROUP INTERACT	17	35	-180	(900)	20	107	85	46	-92	82	-13	-34	<454>	131	-3	-7	86	1
INDIVID RAPPRT	-85	-23	93	24	(820)	-14	147	218	59	-121	-16	36	-4	<250>	-142	55	-6	32
BREADTH	152	149	163	123	-17	(840)	203	85	-41	93	-7	67	-24	-188	<367>	-88	44	-31
EXAMINATIONS	-9	101	349	102	186	254	(760)	218	91	-35	-32	86	-137	-36	2	<135>	-16	123
ASSIGNMENTS	319	-11	232	58	288	111	299	(700)	214	77	-90	-3	-47	-24	95	-19	<356>	225
WRKLD/DIFFCLTY	39	-13	161	-116	78	-53	125	306	(700)	16	-98	-48	-81	3	21	-56	122	<539>

STUDENT EVALUATION FACTORS	INSTRUCTOR SELF-EVALUATION FACTORS: LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD	STUDENT EVALUATION FACTORS: LEARN	ENTHU	ORGAN	GROUP	INDIV	BRDTH	EXAMS	ASIGN	WRKLD
LEARNING/VALUE	<456>	112	-15	89	-137	104	-41	94	20	(950)	455	528	369	222	494	481	521	58
ENTHUSIASM	240	<537>	-48	-14	-19	-8	-32	-109	-120	476	(960)	497	305	350	339	419	248	17
ORGANIZATION	195	151	<306>	-37	41	76	102	-4	-59	562	526	(930)	215	334	562	571	345	-46
GROUP INTERACT	213	52	-239	<484>	-4	-27	-159	-57	-98	382	314	225	(980)	420	165	341	305	-54
INDIVID RAPPRT	33	34	-63	141	<282>	-209	-42	-30	3	233	364	353	433	(960)	156	504	288	80
BREADTH	290	169	104	-4	-162	<413>	2	117	26	522	357	601	172	164	(940)	334	403	178
EXAMINATIONS	208	132	20	-7	63	-100	<166>	-23	-69	512	443	615	357	534	357	(930)	423	-23
ASSIGNMENTS	237	30	21	94	-7	50	-19	<443>	151	557	226	373	321	307	433	457	(920)	204
WRKLD/DIFFCLTY	-68	-35	50	1	37	-36	151	289	<691>	64	18	-51	-59	87	196	-26	228	(870)

NOTE: Values enclosed in ( ) in the diagonals of the upper left and lower right matrices (the heterotrait-monomethod matrices) are reliability (coefficient alpha) coefficients. Values enclosed in < > in the diagonals of lower left and upper right matricies (the heterotrait-heteromethod matrices) are the convergent validity coefficients. All coefficients below the main diagonal of the entire 18 x 18 matrix have been corrected for unreliability. Correlations (presented without decimal points) greater than 100 (i.e., .10) are statistically significant.

TABLE VI

Computational Equations and Results of the ANOVA
Analysis of a Multitrait - Multimethod Matrix

Computations For Sums of Squares and Estimated Variance Components

SOURCE	S S	DF	Variance Component
Class (C) (Convergent Validity)	Nnm (rt)	N-1	(MSc-MSctm)/nm
Class x Traits (Discriminant Validity)	Nnm (rv - rt)	(N-1) (n-1)	(MSct-MSctm)/m
Class X Methods (Method/Halo Effect)	Nnm (rf - rt)	(N-1) (m-1)	(MScm-MSctm)/n
C x T x M (error)	Nnm(1-rv-rf+rt)	(N-1) (n-1) (m-1)	MSctm

NOTE: N = Total Number of Cases (classes)
n = Number of different traits
m = Number of different methods
rt = average correlation coefficient in the entire MTMM matrix (including coefficients both above and below the diagonal and values of 1.0 in the diagonals
rv = average correlation between sources within traits , computed by:
(2 (sum of validity diagonals) + mn) / n m**2
rf = average between trait correlation in the monomethod blocks, computed by:
(2 (sum of monomethod-heterotrait triangle coefficients) + nm) / m n**2
F-Ratios for each of the three effects (convergent validity, divergent validity and method/halo effect) are obtained by dividing the Mean Squares for the effect -- the Sum of Squares divided by the degrees of freedom -- by the Mean Square of the Error term.

Results for Uncorrected and Corrected Corrolation Matricies

SOURCE	df	UNCORRECTED CORRELATIONS SS	MS	F	VARCP	CORRELATINS CORRECTED FOR ATTENUATION SS	MS	F	VARCP
Class (Convergent)	328	1009.55	3.078	6.54**	0.145	1085.89	3.311	8.17**	0.162
C x Trait (Divergent)	2624	3016.25	1.149	2.44**	0.340	3101.05	1.182	2.97**	0.392
C x Method (Method/Halo)	328	661.77	2.011	4.27**	0.171	690.54	2.099	5.27**	0.189
C x T x M (Error)	2624	1234.43	0.470	----	0.470	1044.51	0.398	----	0.398

NOTE: Rt=0.170; Rv=0.680; Rf=0.282
N=329 classes; n=9 traits; m=2 methods

NOTE: Rt=0.183; Rv=0.707; Rf=0.300
N=329 Classes; n=9 traits; m=2 methods

Table III

Configuration of the LAMBDA and PSI Matrices in the GENERAL MODEL

LAMBDA (Factor Loading Matrix)

		Inst Method Factor	Stdt Method Factor	Lrn Trait Factor	Ent Trait Factor	Org Trait Factor	Grp Trait Factor	Ind Trait Factor	Brd Trait Factor	Exm Trait Factor	Asg Trait Factor	Work Trait Factor
		Eta 1	Eta 2	Eta 3	Eta 4	Eta 5	Eta 6	Eta 7	Eta 8	Eta 9	Eta 10	Eta 11
Instructor	Learning/Value	Ilrn	0	Ilrn	0	0	0	0	0	0	0	0
	Enthusiasm	Ient	0	0	Ient	0	0	0	0	0	0	0
	Organization	Iorg	0	0	0	Iorg	0	0	0	0	0	0
	Group Interaction	Igrp	0	0	0	0	Igrp	0	0	0	0	0
	Individual Rapport	Iind	0	0	0	0	0	Iind	0	0	0	0
	Breadth	Ibrd	0	0	0	0	0	0	Ibrd	0	0	0
	Examinations	Iexm	0	0	0	0	0	0	0	Iexm	0	0
	Assignments	Iasg	0	0	0	0	0	0	0	0	Iasg	0
	Workload/Difficulty	Iwrk	0	0	0	0	0	0	0	0	0	Iwrk
Student	Learning/Value	0	Slrn	Slrn	0	0	0	0	0	0	0	0
	Enthusiasm	0	Sent	0	Sent	0	0	0	0	0	0	0
	Organization	0	Sorg	0	0	Sorg	0	0	0	0	0	0
	Group Interaction	0	Sgrp	0	0	0	Sgrp	0	0	0	0	0
	Individual Rapport	0	Sind	0	0	0	0	Sind	0	0	0	0
	Breadth	0	Sbrd	0	0	0	0	0	Sbrd	0	0	0
	Examinations	0	Sexm	0	0	0	0	0	0	Sexm	0	0
	Assignments	0	Sasg	0	0	0	0	0	0	0	Sasg	0
	Workload/Difficulty	0	Swrk	0	0	0	0	0	0	0	0	Swrk

PSI (Correlations Between Factors)

	Inst Method Factor	Stdt Method Factor	Lrn Trait Factor	Ent Trait Factor	Org Trait Factor	Crp Trait Factor	Ind Trait Factor	Brd Trait Factor	Exm Trait Factor	Asg Trait Factor	Work Trait Factor
	Eta 1	Eta 2	Eta 3	Eta 4	Eta 5	Eta 6	Eta 7	Eta 8	Eta 9	Eta 10	Eta 11
Instructor Method	1.0										
Student Method	r1.2	1.0									
Learning/Value	0	0	1.0								
Enthusiasm	0	0	r4.3	1.0							
Organization	0	0	r5.3	r5.4	1.0						
Group Interaction	0	0	r6.3	r6.4	r6.5	1.0					
Individual Rapport	0	0	r7.3	r7.4	r7.5	r7.6	1.0				
Breadth	0	0	r8.3	r8.4	r8.5	r8.6	r8.7	1.0			
Examinations	0	0	r9.3	r9.4	r9.5	r9.6	r9.7	r9.8	1.0		
Assignments	0	0	r10.3	r10.4	r10.5	r10.6	r10.7	r10.8	r10.9	1.0	
Workload/Difficulty	0	0	r11.3	r11.4	r11.5	r11.6	r11.7	r11.8	r11.9	r11.10	1.0

__Note__: All elements with the value of 0 or 1.0 represent fixed values, while all other values are estimated by the LISREL program. The third matrix, the theta matrix, is an 18 x 18 diagonal matrix in which the diagonal values represent variance attributable to random error and/or reliable uniqueness. In the present application, these values were estimated independently and fixed in this analysis. In other applications, these can also be estimated by the LISREL program.

Table IV

Configuration of the LAMBDA and PSI Matrices in the GENERAL MODEL

LAMBDA (Factor Loading Matrix)

	Inst Method Factor	Stdt Method Factor	Lrn Trait Factor	Ent Trait Factor	Org Trait Factor	Grp Trait Factor	Ind Trait Factor	Brd Trait Factor	Exm Trait Factor	Asg Trait Factor	Work Trait Factor
	Eta 1	Eta 2	Eta 3	Eta 4	Eta 5	Eta 6	Eta 7	Eta 8	Eta 9	Eta 10	Eta 11
Learning/Value	-0.666	0.0	0.525	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Enthusiasm	-0.245	0.0	0.0	0.638	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Organization	-0.067	0.0	0.0	0.0	0.523	0.0	0.0	0.0	0.0	0.0	0.0
Group Interaction	0.174	0.0	0.0	0.0	0.0	0.732	0.0	0.0	0.0	0.0	0.0
Individual Rapport	0.055	0.0	0.0	0.0	0.0	0.0	0.653	0.0	0.0	0.0	0.0
Breadth	0.093	0.0	0.0	0.0	0.0	0.0	0.0	0.587	0.0	0.0	0.0
Examinations	0.236	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.716	0.0	0.0
Assignments	-0.156	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.839	0.0
Workload/Difficulty	-0.097	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.643
Learning/Value	0.0	0.697	0.719	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Enthusiasm	0.0	0.571	0.0	0.730	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Organization	0.0	0.729	0.0	0.0	0.735	0.0	0.0	0.0	0.0	0.0	0.0
Group Interaction	0.0	0.480	0.0	0.0	0.0	0.648	0.0	0.0	0.0	0.0	0.0
Individual Rapport	0.0	0.612	0.0	0.0	0.0	0.0	0.515	0.0	0.0	0.0	0.0
Breadth	0.0	0.567	0.0	0.0	0.0	0.0	0.0	0.821	0.0	0.0	0.0
Examinations	0.0	0.829	0.0	0.0	0.0	0.0	0.0	0.0	0.527	0.0	0.0
Assignments	0.0	0.615	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.579	0.0
Workload/Difficulty	0.0	0.101	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.879

PSI (Correlations Between Factors)

	Inst Method Factor	Stdt Method Factor	Lrn Trait Factor	Ent Trait Factor	Org Trait Factor	Grp Trait Factor	Ind Trait Factor	Brd Trait Factor	Exm Trait Factor	Asg Trait Factor	Work Trait Factor
	Eta 1	Eta 2	Eta 3	Eta 4	Eta 5	Eta 6	Eta 7	Eta 8	Eta 9	Eta 10	Eta 11
Instructor Method	1.0										
Student Method	-0.271	1.0									
Learning/Value	0.0	0.0	1.0								
Enthusiasm	0.0	0.0	0.273	1.0							
Organization	0.0	0.0	0.335	0.324	1.0						
Group Interaction	0.0	0.0	0.162	0.057	-0.143	1.0					
Individual Rapport	0.0	0.0	-0.210	0.085	0.127	0.211	1.0				
Breadth	0.0	0.0	0.437	0.227	0.521	-0.017	-0.176	1.0			
Examinations	0.0	0.0	0.202	0.189	0.466	0.014	0.350	0.179	1.0		
Assignments	0.0	0.0	0.409	-0.012	0.231	0.101	0.263	0.320	0.332	1.0	
Workload/Difficulty	0.0	0.0	0.097	-0.003	-0.022	-0.083	0.117	0.198	0.074	0.377	1.0

THETA EPS: Matrix of Uniqueness/Error Variances (values are the diagonals of an 18 x 18 square matrix)

Instructor Self Evaluations of

Learn	Enthus	Organ	Group	Individ	Breadth	Exams	Asignmnt	Workld
0.343	0.515	0.662	0.433	0.515	0.542	0.423	0.277	0.560

Student Evaluations of

Learn	Enthus	Organ	Group	Individ	Breadth	Exams	Asignment	Workld
0.115	0.195	0.129	0.337	0.409	0.159	0.193	0.447	0.236

TABLE V

Summary of Tested Models

	Model Description	ChiSq	DF	ChiSq/DF	Rel
I	9 correlated trait factors, 2 correlated method factors, no trait-method correlations (the general model depicted in Table III)	233.6	98	2.38	.977
II	9 correlated trait factors, 1 "general" factor with loadings on both student & instructor ratings, no trait - "general" correlations	330.7	99	3.34	.962
III	9 correlated traits, 1 student method factor (no faculty method factor), no trait-method correlations	387.7	108	3.59	.952
IV	9 correlated trait factors, 1 faculty method factor (no student method factor), no trait - method correlations	550.6	108	5.10	.933
V	9 UNcorrelated trait factors, 2 correlated method factors, no trait - method correlations	543.6	134	3.99	.951
VI	9 UN correlated traits, 2 UNcorrelated methods, no trait-method correlations	544.3	135	4.03	.950
VII	9 correlated traits, NO method factors	1126.9	117	9.63	.858
VIII	NO trait factors, 2 correlated method factors	4213.7	152	27.7	.561
IX	Null Model (a diagonal SIGMA matrix for which values were determined only by the values in the THETA --error/uniqueness-- matrix)	10564.8	171	61.8	.000
X	8 Correlated traits (the Organization & Breadth trait factors were combined), 2 Correlated Method factors, no trait-method correlations	466.0	106	4.4	.944

NOTE: The values under the column headed "Rel" are Tucker-Lewis reliability coefficients; a measure of the proportion of variance that is explained by the model being tested. Model XI, the least restricted model and the model with the best fit was not identified without further constraints of the model. However, by fixing several near-zero coefficients in the PSI matrix to be zero, the model was identified and could be tested.

TABLE VI

Trait and Method Variance Components For Model I (the General Model)

	Instructor Ratings			Student Ratings		
	Trait	Method	Error	Trait	Method	Error
LEARNING/VALUE	.275	.444	.343	.517	.486	.115
ENTHUSIASM	.407	.060	.515	.533	.326	.19[illegible]
ORGANIZATION	.274	.004	.662	.540	.536	.123
GROUP INTERACT	.536	.030	.433	.420	.230	.337
INDIVID RAPPRT	.426	.003	.519	.265	.374	.409
BREADTH	.345	.009	.542	.674	.321	.159
EXAMINATIONS	.513	.056	.423	.278	.687	.193
ASSIGNMENTS	.704	.024	.277	.335	.378	.447
WRKLD/DIFFCLTY	.413	.009	.560	.773	.010	.236
Mean Across All 9 Evaluations	.432	.071	.475	.481	.372	.246

NOTE: Variance components were derived by squaring the Trait and Method factor loadings from Table V, and using the unsquared value from the Theta Epsilon matrix.

Table VII

Decomposition of Reproduced Correlations

General Equation for the Decomposition of any Reproduced Correlation

The Correlation Between any Measure (X) and any other Measure (Y)

Trait Component		Method Component		Trait-Method Interaction Component
R X Y = <(TX)X(TY)X(RTXTY)>	+	<(MX)X(MY)X(RMXMY)>	+	<(MX)(TY)X(RMXTY)> + <(MY)X(TX)X(RMYTX)>

Where:

TX : The Trait Loading (in Lambda Matrix) for X-Variable
TY : The Trait Loading (in Lambda Matrix) for Y-Variable
RTXTY : The Correlation (in PSI Matrix) Between Trait of X-Variable and Trait of Y-Variable
MX : The Method-Factor Loading (in Lambda Matrix) of X-Variable
MY : The Method-Factor Loading (in Lambda Matrix) of Y-Variable
RMXMY : The Correlation (in PSI Matrix) Between Method of X-Variable and Trait of Y-Variable
RMXTY : The Correlation (in PSI Matrix) Between Method of X-Variable and Trait of Y-Variable
RTXMY : The Correlation (in PSI Matrix) Between Trait of X-Variable and Method of Y-Variable

Decomposition of a Convergent Validity Coefficient: Correlation Between Instructor Ratings of Breadth (X) and Students Ratings of Breadth (Y)

= <(.587)X(.821)X(1.0)> + <(.093)X(.729)X(-.271)> + <(.093)(.821)X(0.0)> + <(.729)X(.587)X(0.0)>

Decomposition of a Heterotrait - Heteromethod Correlation: Correlation Between Instructor Ratings of Organization (X) and Student Ratings of Breadth (Y)

= <(.523)X(.821)X(.466)> + <(-.067)X(.567)X(-.271)> + <(-.067)(.821)X(0.0)> + <(.567)X(.523)X(0.0)>

Decomposition of Two Monotrait - Heteromethod Correlations: Correlations Between Instructor Ratings of Organization (X) and Instructor Ratings of Breadth (Y)

= <(.523)X(.587)X(.466)> + <(-.067)X(.567)X(1.0)> + <(-.067)(.587)X(0.0)> + <(.567)X(.523)X(0.0)>

Correlation Between Student Ratings of Organization (X) and Student Ratings of Breadth

= <(.735)X(.821)X(.466)> + <(.729)X(.567)X(1.0)> + <(.729)(.821)X(0.0)> + <(.567)X(.735)X(0.0)>

Note: For this particular application (see Table IV) all the trait-method interactions were fixed to be zero. So RTXMY and RTYMX are automatically zero. When the X-variable and Y-variable share a common trait (method) the correlation between the traits (methods) is 1.0.