DOCUMENT RESUME

ED 242 785 TM 840 203

AUTHOR	Marsh, Herbert W.; Ireland, Robert
TITLE	Multidimensional Evaluations of Writing Effectiveness.
PUB DATE	28 Mar 84
NOTE	33p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Correlation; Essay Tests; Foreign Countries; Grade 7; Grading; *Holistic Evaluation; *Interrater Reliability; Junior High Schools; *Master Teachers; Measurement Techniques; Rating Scales; Student Evaluation; *Student Teachers; Testing Problems; *Writing Evaluation
IDENTIFIERS	Australia (Sydney); Confirmatory Factor Analysis; *Multidimensional Approach; Multitrait Multimethod Techniques

ABSTRACT

To test the applicability of multidimensional ratings of writing effectiveness that are amenable to normal classroom usage, all grade 7 students (N=139) from one suburban school (Sydney, Australia) wrote a brief essay. Master and student teachers evaluated all the essays according to overall effectiveness of written expression and according to holistic ratings of specific components (mechanics, sentence structure, organization, word usage, content/ideas, and style). Ratings of writing effectiveness by master teachers and by student teachers were substantially correlated with each other and with an external validity criterion. Correlations were particularly high for the sum of ratings of specific components, but were nearly as high for overall, global ratings. The single-rater reliability (r=0.7), the average of correlations between each pair of raters, was higher than expected from previous research. The average of single-rater reliabilities for specific components (r=0.6) was also high. However, the predicted ability of teachers to discriminate among the multiple components, except perhaps the mechanics facet, was not supported in a variety of multitrait-multimethod analyses. The student teacher ratings were nearly as reliable and as valid as master teacher ratings, and student teachers were prehaps better able to differentiate among different components of writing effectiveness. (PN)

Multidimensional Evaluations of Writing Effectiveness

Herbert W. Marsh and Robert Ireland
The University of Sydney, Australia

28 March, 1984

Running Head: Writing Effectiveness

The objective of an essay test is to determine whether a student is able to write a clear and effective essay on a given topic. The assessment of an essay may focus on writing effectiveness or on the level of achievement in a content area. While these two uses of essay testing have some common features, it is important to distinguish between them. The focus of the present study is on the evaluation of effective writing, but as Foley (1971) suggests, the lack of a definition of "good writing" makes the task difficult. Most research in this area employs holistic/impressionistic ratings where raters form a single, overall impression, but there is research which uses analytic procedures or a technique which combines aspects of both approaches. In a purely analytic approach objective measures of language production (e.g., number of words, words per clause, ratio of subordinate clauses to total clauses, spelling errors, etc.) are measured or counted. A hybrid technique, representing a compromise between the two approaches, is to obtain global ratings on specific components hypothesized to underlie overall writing effectiveness (e.g., mechanics, organization, style).

Traditionally, high school English teachers evaluate writing in two ways. First, they "correct" an essay and provide varying amounts of formative feedback to the student writer. This task typically involves some form of the analytic or hybrid approach. Second, they assign an overall, summative evaluation (a mark or a grade) to the essay, which is generally a holistic rating. Harris (1977) found that content and organization, as opposed to mechanics, were more important in determining overall evaluations of writing samples, but that written, formative feedback to students emphasized mechanics. Freedman (1979) experimentally manipulated essays and found that while content and organization were most important to the determination of overall evaluations, mechanics and sentence structure also had some influence. She suggested that the relative influence of different components might vary depending upon the rater, or the type or purpose of essay, and this might contribute to the unreliability of overall impressions. Chase (1968, 1983) argued that even when raters are specifically instructed to ignore factors such as quality of handwriting and mechanical errors, they are apparently unable to do so and that their overall ratings reflect experimentally manipulated effects due to such influences.

Quellmalz, Capell and Chou (1982) argue that there are generically distinct methods of writing for particular purposes such as narrative and expository essays. They found that overall ratings were consistently lower for narrative essays than expository essays and that the rater agreement on two different essays written by the same student was higher when both essays were expository or both were narrative than when one was expository and one was narrative. However, their results also depended on the particular component of writing effectiveness that was being assessed. For example, mechanics was the component of writing effectiveness that was most clearly distinguished, and rater agreement on two different essays for Mechanics did not depend on whether both essays were written for the same purpose or for different purposes. The authors concluded that their findings challenged the assumption that writing effectiveness is a unidimensional construct, and argued for the development of specific components of writing effectiveness.

Single Rater Reliabilities With Holistic/Impressionistic Marking.

Sources of error in evaluating essays include:

1) Student error -- chance fluctuations in the performance of the student which are not stable over time;

2) Test error -- since a writing sample can be considered a one-item test based on only a limited sample of relevant behavior, individual students may perform better or worse on different, equally appropriate essay topics;

3) Scale error -- idiosyncratic ways in which a particular rater uses the given response scale when evaluating an essay;

4) Rater error -- error due to disagreement in ratings of the relative quality of the same essay by different raters;

5) Writing Purpose Error -- alternative writing tasks (e.g., narrative and expository) may tap different components of writing effectiveness and performance on one task may not generalize to the other.

The focus of this study will be on the rater error, though it is clear that each source of error and interactions among the different sources can be substantial (Breland & Gaynor, 1979; Coffman, 1966; 1971; French, 1966; Moss, Cole & Khampalkit, 1982; Quellmalz, Capell & Chou, 1982). The reliability of essay evaluations, even when consideration is limited to rater error, varies systematically and predictably with the number of raters (e.g., Coffman, 1966; 1971). Consequently, for purposes of this study, the single-rater reliability will be defined as the correlation between two ratings of the same essay each performed

independently by two separate individuals, or the average correlation between all pairs of raters when there are more than two markers.

Hall (1972), in a review of research conducted in the US, England and Australia prior to 1972, concluded that single-rater reliabilities of about 0.60 appears to represent "the limit of the extent of agreement one can generally expect between single judges marking one essay" (p. 32). Coffman (1966) reported that the correlation between responses by two raters (i.e., the single rater reliability) to the same short essay was about 0.38, though the reliability of the sum of responses by five raters was 0.76. Huddleston (1954) reported that highly trained examiners for English compositions on the College Board Examination were able to achieve a single-rater reliability of about 0.55 for a long paper on a single topic. Diederich (1974) reported that "even after working with an English staff for some time, I have rarely been able to to boost the average correlation between pairs of readers above 0.50, and other examiners tell me that this is about what they get" (p. 33). French (1966) suggested that with extensive training and monitoring, the single-rater reliabilities could be as high as 0.70, but that when untrained raters from various academic disciplines were asked to evaluate essays according to their own judgments of what constitutes writing ability, the single-rater reliability was only 0.31.

In the same study French (1966) reported that the single-rater reliability for English teachers was 0.41, and was appreciably higher than for the group as a whole. Thompson & Bailes (1926) reported single-rater reliabilities of 0.65 for experience teachers and 0.50 for untrained students. Michael, Cooper, Shaffer and Wallis (1980) also found that the single-rater reliability for English professors was 0.64 and 0.85 on two essay topics, while corresponding values based upon ratings by faculty from other disciplines were 0.56 and 0.64. However, Phillips (1948) had essays graded by 77 practicing teachers and 373 education students who were not teachers and found that the single-rater reliability was 0.43 for teachers and 0.41 for non-teachers.

Harkin (1983) described procedures used in the corporate holistic marking of the New South Wales (Australia) English reference test which is completed by 75,000 year 10 students each year. Prior to marking, senior examiners select "range-finder" essays which are used to define each of the categories on the 15-point response scale used to evaluate essays. Examiners are brought to a single location, briefed on the use of the range finder essays, and given considerable training

and practice before the actual marking exercise is begun. During the marking, examiners each work within a group of three, and are encouraged to converse with other members of his/her team when questions arise, though the actual ratings are made by a single examiner. During the marking operation the mean, standard deviation, and reliability estimates of responses by each marker are tabulated on a daily basis, and additional consultation with senior examiners occurs when necessary. Using this system random samples of essays were each graded by multiple markers and Harkin reported a mean single-rater reliability of 0.80. However, this value is probably a somewhat inflated estimate of the correlation between two raters grading the same essay when each is working strictly independently. Even higher estimates of reliability were obtained with samples of the essays specifically selected to unambiguously represent each of the 15 scale points, although these essays are selected to monitor the marking operation and not to provide an unbiased estimate of single rater reliability.

In summary, single rater reliabilities generally vary between 0.3 and 0.8 depending upon the length and the topic of the essay, the amount of freedom students have in selecting and responding to the essay, the experience of the raters, the extent of training given the raters, and the control exercised in monitoring and standardizing the rating environment. The single rater reliability for short, in-class essays marked by classroom teachers tend to be substantially lower than estimates obtained in large, corporate marking studies.

Components of Writing Effectiveness.

French (1966) summarized an attempt to derive components of writing effectiveness from the comments from readers of essays. The comments were classified into 55 categories and submitted to factor analytic techniques. French identified five factors representing Ideas, Form, Flavor, Mechanics, and Wording. Foley (1971), using this and other research, argues that writing effectiveness can be categorized into five major components; Ideas, Organization, Style, Mechanics, Choice of Words. However, much of this research is based upon inferences based upon written comments or a logical analysis of the writing process, rather than upon the determination of whether raters are actually able to distinguish between these components. Studies by Cast (1939), Hartog (1941), Moss et al. (1982), and Smith (1979) each suggest that a general factor of writing effectiveness underlies ratings of specific components.

Smith (1979) compared impressionistic/holistic ratings, ratings on six specific components (focus, development, organization, support, paragraphing and mechanics), and objective test scores designed to measure the same six components. Total scores for the holistic and rating scales were substantial, and correlated 0.65 and 0.61 respectively with the objective total score. However, ratings among the six specific scales were highly correlated, ranging from 0.69 to 0.90, and were highly correlated with the total of the specific ratings, correlations ranging from 0.82 to 0.96. While Smith concluded that it was tempting to infer that the specific rating scales actually **tapped a single unitary dimension of writing effectiveness,** she suggested that distinguishable subscales may emerge when the writing task is less structured. She also found some support for rater's ability to distinguish Mechanics from other components of writing effectiveness.

In a technically sophisticated study, Quellmalz, Capell and Chou (1982) compared ratings of general impression, ratings on four specific components of writing effectiveness, and objective test scores designed to measure three of the four specific components. The specific components were four of the six employed in the Smith (1979) study, and the scoring systems used in the two studies were similar. Quellmalz et al., however, examined writing effectiveness for expository essays, narrative essays, and for a paragraph writing task. Although a wide variety of analyses are reported, the most relevant to the present investigation was the multitrait-multimethod analysis of specific ratings of the essays (their step 1 in Table 4, p. 253). While their analysis argued for the existence of three distinguishable facets, Coherence, Support and Mechanics, correlations among these trait-factors varied from 0.63 to 0.80. Although not reported, the authors indicated that the correlations among the components were even larger when results from the paragraph writing task and/or the objective test scores were included in the analysis. As with the Smith study, the Mechanics component was most distinct. The authors argued that "further examination of the value of rating writing according to separate component features should consider both their diagnostic utility and component distinctiveness" (p. 256), and is consistent with the aim of the present investigation. The study also demonstrated the importance of confirmatory factor analysis of MTMM data in the study of multiple dimensions of writing effectiveness.

Intuitively, evaluations of effective writing seem to be

multifaceted, and the structure outlined by Foley (1971) provides a well-conceived, theoretical basis for what the different facets might be. If these components can be reliably differentiated, then the evaluation of each component separately has several possible advantages, particularly in the typical classroom setting:

1) Feedback to Students. Scores on the separate components, in addition to written comments, and perhaps, an overall mark, will provide students with more detailed feedback which is formative in nature. This is particularly important, if, as Harris suggests, formative feedback traditionally emphasizes different components than does overall, summative assessments.

2) Definition of Effective Writing. Effective writing is difficult to evaluate, partially because there is no operational definition of what constitutes effective writing. The successful application of these categories would provide a better definition of what is meant by effective writing, and how this differs for different kinds of writing.

3) Reliability. An average rating across several components may be more reliable than is an overall global assessment, particularly if part of the disagreement among raters is due to the relative emphasis placed on the different components.

4) Validity. Improving reliability may improve validity as well. Furthermore, the optimal weighting of the different components may vary, depending upon the criterion of effectiveness, but this information is lost if only an overall assessment is used.

5) Bias. Variables which may bias ratings of writing effectiveness are likely to have a larger impact on a single, ill-defined, overall assessment of writing effectiveness than on separate, more narrowly defined components.

The Present Study.

The present study is designed to test the applicability of multidimensional ratings of writing effectiveness which are amenable to normal classroom usage, rather than to determine what might be possible in an ideal setting. It is important to note that raters were specifically not given extensive training in the rating task, that the ratings were not made in a highly controlled setting, that the raters had no chance to discuss the task with each other or the researchers, and that the constraints on the task for student writers and for raters were not specifically designed for purposes of this study. The rating tasks were relatively unstructured and teachers were encouraged to use perspectives they typically employ in their own practice.

Ratings of multiple components and an overall evaluation were made by both master and student teachers. Two procedures were used in the analysis. First, overall ratings and total scores derived from the component ratings were obtained. Single rater reliabilities were determined, and the ratings were correlated with an external validity criterion. Second, multitrait-multimethod analyses were employed to determine if the teachers were able to differentiate among the hypothesized components of writing effectiveness. It was predicted that:

1) the single rater reliability of responses by master teachers would be about 0.5 for overall impressions, and somewhat higher for ratings based upon the sum of ratings of specific components;

2) validity estimates would also be somewhat higher for total scores than for overall impressions;

3) both reliability and validity estimates would be somewhat lower for student teachers;

4) master and student teachers would be able to differentiate among the different rating components, that the differentiation would be better for more objective components like mechanics, and master teacher would be better able to differentiate among the components.

Method

Sample and Procedures.

Students consisted of all 139 seventh grade students attending one public, coeducational high school in suburban Sydney, Australia. Virtually all students were native English speakers and were born in Australia. The students were somewhat brighter than average, as indicated by the mean IQ of 106 obtained from their school records. The socioeconomic status of the geographic areas serviced by this school varied from working class to upper class, though the majority of the students came from middle class backgrounds.

Early in the academic school year, all seventh grade students were asked to write a brief story of one or two pages about one of three possible subjects -- a chase, an animal, or a game. The choice of the subjects was up to the student. (Wiseman & Wrigley, 1958, demonstrated that allowing children to select a topic had little impact on errors in marking.) Instructions were read aloud to all students, but once they actually began writing, they were given no help or assistance. Hence, the task which is the focus of this study is similar to the school performance test described below. The completed essays in this study varied in length from about 100 words to about 500 words.

specific components, and also gave an overall evaluation. They were given the following descriptions of the components:

1) MECHANICS (e.g., spelling, capitalization, punctuation, grammar, tense, subject-verb agreement, etc.)

2) SENTENCE STRUCTURE (e.g., use of complete sentences, appropriate use of phrases/clauses, word order, variations in structure, etc.)

3) WORD USAGE (e.g., fluency, appropriateness, selection, range of usage, level of usage, etc.)

4) ORGANIZATION (e.g., adequate introduction & ending, logical order, paragraph/theme structure, coherence, emphasis, transition, etc.)

5) CONTENT/IDEAS (e.g., relevance to topic, comprehensibility, logic, clarity, appropriate explanation and summarization, relevant arguments/examples, etc.)

6) QUALITY OF STYLE (e.g., originality, creativity, flavor, interest value, freshness, individuality, etc.)

7) OVERALL EVALUATION (This judgment should be made according to your own criteria and should represent your own subjective impression. It may or may not reflect the first six criteria, and may also represent other characteristics that you feel are important.)

Teachers were asked to make each of their ratings on a nine-point response scale which varied from "1-Very Poor" to "9-Very Good"; and to adhere to standards of quality that they felt were appropriate for year seven. The teachers were asked to make all ratings for each essay after a single reading (i.e., they were not asked to reread the set of essays separately in order to make each rating).

Three university students, who were in the process of completing a degree in Education which would qualify them to teach English in secondary schools, were also asked to evaluate the essays. The student-teachers were selected by a university lecturer as being good, responsible students in the teacher education program. However, except for practice teaching, these student-teachers had had no actual classroom teaching experience. The student teachers were given exactly the same set of instructions as the master teachers and were requested to evaluate the essays according to the specific components of writing effectiveness and to provide an overall evaluation, but they had not made early ratings 10 months prior to this task as had the master teachers.

The following set of scores, derived from the procedures described above, was computed for each of the 139 students who completed essays for this study:

Validity Criterion -- 1 score based upon school performance on the essay test administered by the school.

Early Ratings -- 3 scores, one from each Master teacher, which represent global, holistic impressions of essays in this study.

Component Ratings -- 36 scores, six from each of the three student-

teachers and six from each of the three master-teachers, representing scores on the specific components used in evaluating the essays in this study.

Overall Ratings -- 6 scores, one from each of the teachers, representing global, holistic impressions of the essays used in this study at the time of the second rating.

Total Ratings -- 6 scores, one from each of the teachers, representing the sum of scores on the six component ratings (but not the overall rating).

In addition to the 52 scores described above, nine scores for each essay were obtained by summing across the responses by the three master teachers for the early ratings, the six component ratings, the overall ratings, and the total ratings. Eight corresponding scores were obtained by summing across responses by the three student-teachers for all but the early ratings (student-teachers did not make early ratings).

Statistical and Multitrait-multimethod Analyses.

Correlations among various sets of scores were used to determine the single rater reliability and the validity of the overall and total scores. However, an important aspect of this study was to determine the extent to which teachers can differentiate among the various components of writing effectiveness described above. Multitrait-multimethod (MTMM) analyses, where responses by different teachers correspond to methods of evaluation and the specific components of writing effectiveness correspond to different traits, is ideally suited to this purpose. In MTMM analyses the distinction is made between convergent validity, the agreement between different raters on the same component, and divergent validity, the ability of the raters to differentiate among the different components. Hence, the convergent validities in multitrait-multimethod analyses, are really single rater reliabilities in this particular application. This distinction is important in the interpretation of the findings, but in no way affects the actual procedures in conducting MTMM analyses (for further discussion of this distinction see Marsh, Smith, Barnes & Butler, 1983; Marsh, Barnes, & Hocevar, in press). Three approaches to MTMM analyses are briefly summarized below, but an extensive review of the procedures is beyond the scope of this paper and the interested reader is referred to Marsh and Hocevar (1983, in press; also see Kenny, 1979; Schmitt, Coyle, & Sarri, 1977).

Campbell and Fiske (1959) argue that the demonstration of

construct validity requires both convergent and discriminant validity; that is, multiple indicators of the same component of writing should be substantially correlated with each other, but less correlated with indicators of other components. Convergent validity is inferred from agreement between measures of the same component of writing effectiveness assessed by different teachers. Discriminant validity or divergent validity refers to the distinctiveness of the different traits, and in this case is inferred from the relative lack of correlation between different components of writing effectiveness. Campbell and Fiske proposed four guidelines for evaluating MTMM matrices. These guidelines have been criticized, but they are still represent the most frequently employed strategy, are useful, and are recommended as the first step in analysis of MTMM data (Marsh & Hocevar, 1983; in press).

An ANOVA model (Kavanagh, et al., 1971) provides a more analytic approach to MTMM analysis. When repeated measures of cases -- the essays in this application -- are measured over all levels of traits (the rating components) and methods (the teachers), three sources of variation can be identified. The main effect of essays is a test of how well the ratings discriminate between essays, and is taken to be an indication of convergent validity. The essay-by-trait interaction tests whether differentiation among the essays depends upon the specific components of writing effectiveness; if it does not then the components have no discriminant validity. The essay-by-teacher interaction tests whether the differentiation depends upon teachers; if it does the the different teachers introduce a source of systematic (undesirable) variance which is taken to be an indication of method/halo effect. Kavanagh, et al. (1971; also see Marsh & Hocevar, 1983) describe procedures whereby these effects and corresponding variance components can be obtained directly from the MTMM matrix, and these are employed in the present application. However, despite the convenience of statistical tests and summary statistics, this procedure has important limitations, the effects tested with this model bear no straightforward correspondence to the interpretation of convergent and discriminant validity as used in other MTMM analytic strategies, and it is recommended only to supplement the application of other approaches (Marsh & Hocevar, 1983;.

Confirmatory factor analysis (CFA) has more recently been applied to the analysis of MTMM matrices. MTMM matrices, like any other correlation matrix, can be used to infer the underlying dimensions that

are being measured. In the present application, factors defined by the measures of the same component of writing effectiveness support their construct validity, while factors defined by different components rated by the same teacher argue for method/halo effects.

Conventional/exploratory factor analysis, because of the indeterminancy of the solution and the researcher's relative lack of ability to define a model, is generally inappropriate for analyzing MTMM matrices. With confirmatory factor analysis, the researcher is able to specify different models and to determine how well these various models fit the data. Hence, the analysis of the MTMM matrix can be viewed as a straightforward application of confirmatory factor analysis with a priori factors corresponding to specific methods and traits, and the findings can be interpreted in the same way as can other confirmatory factor analyses.

In the present application, the CFA was conducted with the commercially available LISREL V program (Joreskog & Sorbom, 1981). The most general MTMM model employed in this study consisted of 12 factors representing the six components of effective writing (traits) and the six teachers (methods). Each of the 36 measured variables was used to define one method factor and one trait factor while loadings on the other 10 factors were fixed to be zero. For example, ratings by the first teacher of the Mechanics component was used to define the method factor for the first teacher (along with the other five ratings by the same teacher) and the Mechanics trait-factor (along with the other five ratings of Mechanics by each of the other teachers). Hence, the 36 measured variables were used to define 72 factor loadings, and all the other factor loadings are defined to be zero. The 15 correlations among the six method factors and the 15 correlations among the six trait factors were estimated in the analysis, but correlations between method and trait factors were fixed to be zero. The 36 error/uniquenesses, one for each measured variable, were defined so as to form a diagonal matrix so that the error terms were uncorrelated. This pattern of loadings represents the standard model used in the analysis of MTMM matrices (see Marsh & Hocevar, 1983; in press). The fit of this CFA model to the data was assessed by the magnitude of the parameter estimates, the ratio of the chi-square to the degrees of freedom in the analysis, the root mean square of the residual differences between the observed and reproduced correlation matrices, and coefficient d which scales the observed chi-square along a scale of zero-to-one where the end-points represent a null fit and a perfect

fit (see Bentler & Bonett, 1980; Joreskog & Sorbom, 1981; Marsh & Hocevar, 1983; in press; Maruyama & McGarvey, 1980). As yet there are no universally accepted measures of goodness of fit in CFA (Marsh & Hocevar, 1984), but the most widely applied indication is the chi-square/df ratio where values of less than 2.0 are taken as an indication of a good fit (despite the relationship between this indicator and sample size), while the coefficient d provides an index analogous to measures of the proportion of variance explained in ANOVA procedures.

RESULTS

Overall and Total Ratings.

The first purpose of this study is to determine the ability of master and student teachers to assess overall writing effectiveness. Single rater reliabilities, correlations among the overall ratings and among the total ratings, and the validity coefficients (see Table 1) are consistently high and remarkably uniform for both student and master teachers. Correlations among the six total scores vary between 0.68 and 0.78 (mean r = 0.72); correlations among student-teachers (mean r = 0.69), among master-teachers (mean r = 0.72), and between student and master teachers (mean r = 0.73) are nearly the same. A similar pattern of slightly smaller correlations (mean r = 0.67) exists among the overall ratings, and among the early ratings by the three master teachers (mean r = 0.71). Hence, the correlation between ratings by any two teachers, whether student or master teachers, is approximately 0.70 whether based upon total scores, on the overall rating, or on the early ratings which were available only for master teachers.

Marks on the school performance essay examination provides one external criterion of validity against which to assess the ratings. Correlations between Master teacher ratings and the criterion are again close to 0.7 whether based upon total scores (mean r = 0.71), overall ratings (mean r = 0.69) or the early ratings (mean r = 0.68), while correlations between the criterion and student-teacher ratings are nearly as high (mean r's = 0.66 & 0.65 for total and overall ratings).

Correlations between overall and total ratings by the same person (e.g., O1 & T1) are quite high (mean r = 0.96), indicating that the sum of the component ratings is measuring a construct which is nearly the same as the overall rating. Correlations between early ratings and subsequent ratings by the same master teacher are also high for both overall ratings (mean r = 0.80) and total ratings (mean r = 0.82),

indicating that the ratings are stable over time.

The focus here, as well as in subsequent analyses, is on the relative agreement between different teachers based upon correlations among their ratings. However, the means of the different ratings in Table 1 also provide a basis for looking at absolute differences. Master teachers, based upon overall ratings and total scores, assign somewhat lower marks than do student teachers. It is interesting to note, however, that the early ratings by the group of master teachers are also somewhat lower than are the marks assigned on the school performance test by other experienced teachers, though the two sets of marks are for different tasks and may not be strictly comparable (see footnote 1).

In summary, correlations between the ratings by any two teachers, whether they be student or master teachers, and correlations between any teacher's rating and the validity criterion are all approximately 0.70. Correlations based upon the total scores are slightly higher in each of the various comparisons, but the differences are small. Correlations between ratings by the same teacher at two different times are higher, suggesting that there is a small systematic method/halo effect in the ratings by each teacher which generalizes over time. The similarity in correlations between ratings by different teachers in our study, and between their ratings and the validity criterion, apparently reflects two countervailing effects; the validity correlations should be lower since they are based upon ratings of a different essay, but should be higher in that the validity criterion, based upon ratings by two teachers, is probably more reliable than ratings of essays by any one teacher in this study.

Multitrait-Multimethod (MTMM) Analyses.

The second purpose of this study is to determine if teachers are able to distinguish among the different components of writing effectiveness. This is examined in various analyses based upon the MTMM matrix (Table 2) where correlations in the triangular (heterotrait-monomethod) blocks represent correlations among the component ratings by the same teacher, correlations in the square (heterotrait-heteromethod) blocks represent correlations based upon ratings by different teachers, and convergent validities (the diagonals of the square blocks which are underlined in Table 2) represent agreement between two different teachers on the same component.

Campbell-Fiske Guidelines. The application of the four Campbell-Fiske guidelines indicates:

1) the convergent validities, ranging from 0.32 to 0.75 (median r = 0.60), are all statistically significant, though those based upon master-teacher ratings (median r = 0.63) are slightly higher than for student-teacher ratings (median r = 0.55).

2) the convergent validities are higher than other correlations in the same row and column of the square blocks for only 70% of the comparisons, and the percentages are similar when ratings by student-teachers (74%) and master teachers (70%) are considered separately. None of the components satisfies this test for all the comparisons and the convergent validities (median r = 0.60) are only slightly higher than the correlations with which they are compared (median r = 0.55).

3) the convergent validities (median r = 0.60) are higher than the correlations in the same row and column of the corresponding triangular blocks (median r = 0.73) in only 12% of the comparisons based upon the entire matrix, and in 11% and 3% respectively when ratings by student and master teachers are considered separately. The median of correlations against which the convergent validities are compared is higher here than those for comparison in guideline 2 (0.73 vs. 0.55), suggesting a halo/method effect in the ratings of different teachers.

4) The pattern of correlations among different components is somewhat similar for each of the different teachers; the highest correlations generally occur between ratings of Mechanics and Sentence Structure, and between ratings of Content/Ideas and Quality of Style, and the lowest correlations generally occur between ratings of Content/Ideas and ratings of either Mechanics or Sentence Structure.

In summary, the application of the Campbell-Fiske guidelines provide strong support for the convergent validity of the ratings, but not for their divergent validity. These findings suggest that while there is good agreement between the ratings of different teachers in a general sense, as was observed with the global and total ratings, teachers are not able to distinguish clearly between specific components of writing effectiveness. Surprisingly, better, albeit weak, support for the divergent validity of the ratings came from responses by student teachers than by the master teachers. Also, inspection of Table 3 indicates that there was better support for the divergent validity of some components (e.g., Mechanics and Word usage) than for that of others (e.g., Content/Ideas and Organization).

ANOVA Analysis of the MTMM Matrix. The results of the ANOVA model applied to the entire MTMM matrix, and separately to student and master teacher ratings (see Table 4) are generally consistent with the results

of the Campbell-Fiske analysis. In each of the analyses:
1) the effect of the essays (the convergent validity effect) is large and statistically significant;
2) the effect of the essay-by-teacher interaction (the method/halo effect) is moderate and statistically significant; and
3) the effect of the essay-by-component interaction (the divergent validity effect) is small and does not even reach statistical significance when the master teacher ratings are considered separately. Hence, these analyses also suggest good convergent validity, but the relative inability of teachers -- particularly the master teachers -- to distinguish among the different components of writing effectiveness.

Confirmatory Factor Analysis (CFA) of the MTMM Matrix. Recently, analysis of MTMM data with the Campbell-Fiske guidelines or the ANOVA model have been criticized, and the use of CFA has been recommended (see Marsh & Hocevar, 1983 for an overview). When this approach is used in the most general model, separate factors representing traits and methods are hypothesized, and the ability of such a model to fit the data is quite good (i.e., model 1 in Table 5 has a chi-square/df ratio of 1.6, and has a coefficient d = 0.868). However, much of the variance explicable by this model can be explained by model 2, a model which contains only a single, general factor (coefficient d = 0.621). Models 3 and 4 test the ability of trait-factors without any method factors (model 3) and method-factors without any trait-factors (model 4) to explain the data. Model 3, hypothesizing six trait-factors does little better than model 2 where a single general factor is hypothesized (0.642 vs. 0.621), while model 4, hypothesizing six method-factors, does nearly as well as model 1 (0.80 vs. 0.868). The addition of a general factor to models 3 and 4 improves their ability to fit the data (models 5 and 6).

The ability of the alternative models to fit the MTMM data supports the general findings of earlier analyses of the same data. Much of the variance can be explained by a single, general factor which incorporates all the component ratings by all the teachers (model 2). The method-only model (model 4) explains more of the variance than does the trait-only model, suggesting a method/halo effect but weaker support for divergent validity. The finding that one general factor can explain nearly as much variance as the six trait factors suggests that there is almost no discriminant validity at all.

The traditional interpretation of the CFA model suggests that method-factors are indicative of a bias, while trait-factors are

indicative of validity. In order to test this interpretation in the present application, a 13th factor, representing the convergent validity criterion, was added to model 1, and the parameter estimates are shown in Table 6. As in model 1 (whose parameter estimates are nearly the same for the first 12 factors) the measured variables loaded substantially on the method factors and less substantially on the trait factors. Furthermore, correlations among the trait factors are generally quite high and in some cases approach 1.0. However, of particular interest here are the correlations between the 13th factor (the validity criterion is labelled "V" in Table 6) and the other factors. Correlations between the validity factor and the method factors are large (median r = 0.67), while correlations between the validity factor and the trait-factors are much smaller (median r = 0.24). Thus, at least in this application, the interpretation of the method-factors as indicating bias seems unwarranted. Instead, the so-called method factors appear to represent a general component from the ratings by each teacher which is highly correlated with an external validity criterion. The high correlations among the different method factors (median r = 0.72) are also inconsistent with an interpretation that each of these factors represents a method/halo effect which is idiosyncratic to the ratings by each teacher.

Summed Student and Master Teacher Ratings. Responses by the three student teachers were summed to form summed ratings of each of the six components of writing effectiveness, as were the responses by the three master teachers. A new MTMM matrix (see Table 7) was formed where the six writing components represented traits and the two types of teacher represented methods. It was hoped that these summed ratings, since they are more reliable, would provide stronger support for the discriminant validity of the ratings. As expected, the convergent validities are quite substantial (median r = 0.84). There is modest support for the divergent validity of ratings of Mechanics, Sentence Structure, and Word Usage in that the convergent validities are higher than other correlations in the square block (the second Campbell-Fiske guideline) and higher than correlations among the different student-teacher ratings (the third Campbell-Fiske criterion), even though they are generally lower than the correlations among master-teacher ratings. Nevertheless, even here, there is only modest support for the ability of ratings to differentiate among the different components of effective writing.

The validity criterion and summed responses to the overall

ratings, total ratings, and early ratings also appear in Table 7. Agreement between student and master teachers is particularly high for the total ratings (r = 0.91) and somewhat higher than correlations involving the overall and early ratings. The total scores by student and master teachers are also somewhat more highly correlated with the validity criterion (r's = 0.76 & 0.79) than the overall ratings or the early ratings, though all correlations are high and differences are small. Total ratings, overall ratings, and early ratings tend to be more highly correlated with ratings of Quality of Style and Word Usage, than with other specific components, but again, all the correlations are large. These findings offer further support for the reliability and validity of the ratings by master and student teachers, and limited support for their ability to distinguish among some components of writing effectiveness.

DISCUSSION

A variety of different analyses have demonstrated that ratings of writing effectiveness by master teachers and by student teachers are substantially correlated with each other and with an external validity criterion representing actual school performance. Agreement among ratings by different teachers, and between these ratings and the validity criterion were particularly high for the sum of ratings to specific components of writing effectiveness, but were nearly as high for overall, global ratings. Student-teacher ratings, using a variety of different comparisons, were nearly as reliable and valid as master-teacher ratings, and student-teachers seemed better able to differentiate among the components of writing effectiveness.

The results of the study provide a number of surprises, particularly when compared with the results which were predicted. On the positive side, single rater-reliabilities and validity coefficients were substantially higher than expected. As expected, the total ratings did somewhat better than did overall, holistic responses, but the differences were small. Of surprise was the finding that student teachers did nearly as well as master teachers on most comparisons, and perhaps were better able to differentiate among the different components of writing effectiveness. On the negative side, the predicted ability of teachers to differentiate among the components of writing effectiveness was so weak as to be of little practical value.

The size of the single-rater reliabilities and validity coefficients are larger than typically found, even when raters receive extensive training, when essays are marked in highly controlled

situations, and when essays are much longer. This demonstration is important and may reflect the fact that essay testing is more common in Australia so that both the students and the raters are more familiar with the task. This finding is also important because it demonstrates that raters were able to maintain a high level of concentration throughout the task so that this cannot account for their apparent difficulty in distinguishing among traits.

The apparent difficulty that readers have in distinguishing among components of writing effectiveness is consistent with other research. We know of no other research where raters have been able to clearly differentiate among multiple component of writing effectiveness, though there is relatively little research which has employed rigorous tests of this conclusion. Here, as in the studies by Smith (1979) and Quellmalz et al. (1982) where components were more explicitly defined and raters received considerable training, the most distinguishable component was mechanics. Alternative strategies might provide better differentiation among the components of writing effectiveness, but only at the expense of the applicability. This is important since the goal of the present investigation is to devise a procedure which is likely to be employed by classroom teachers. Teachers could be asked to perform multiple sorts of the essays into separate piles, once for each specific component. However, such a procedure would require much more time than a holistic strategy or the one employed here, and this might not be acceptable in many settings. Also, teachers could be asked to judge four or five subcategories within each of the components of effective writing, and these ratings could then be factor analyzed to test the hypothesized factor structure. While this would probably improve the differentiation among the components, it would also require considerable more time and might be unacceptable in many settings. Teachers could be asked to participate in extensive training programs where the rating categories are more explicitly defined and feedback is provided on practice essay marking, but previous research has not shown even this to produce clear differentiation among multiple components of effective writing. We believe that further research such as suggested here will demonstrate the multidimensionality of writing effectiveness, and that the goal of this research should be to demonstrate how this can be best accomplished. The use of MTMM and CFA as demonstrated here provide an important tool for such research on writing effectiveness.

FOOTNOTES

1 -- The statistical significance of differences between student and master teachers was based upon comparisons of their summed overall ratings and total scores; in each case master teacher responses were significantly lower ($t(138) = 9.28$ & 6.21 respectively, $p < .001$). A similar comparison between the summed early ratings by the Master teachers and the validity criterion based upon school performance also showed that the master teacher ratings were significantly lower ($t(138) = 9.52$, $p < .001$). However, since the essays evaluated in this final comparisons were not the same, the significant effect may reflect differences in grading standards or differences in the quality of the essays being evaluated.

REFERENCES

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

Breland, H. & Gaynor, J. I. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

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

Cast, B. M. O. (1940). The efficiency of different methods of marking English compositions -- Part II. British Journal of Educational Psychology, 10, 49-60.

Chase, C. I. (1968). The impact of some obvious variables on essay-test scores. Journal of Educational Measurement, 5, 315-318.

Chase, C. I. (1983). Essay test scores and reading difficulty. Journal of Educational Measurement, 20, 293-297.

Coffman, W. E. (1966). On the validity of essay examinations of achievement. Journal of Educational Measurement, 3, 151-156.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education.

Diederich, P. B. (1974). Measuring growth in English. Urbana, Ill.: National Council of Teachers of English.

Foley, J.J. (1971). Evaluation of learning in writing. In B.S. bloom, J. T. Hastings, & G. F. Maduas (Eds.), Handbook on formative and summative evaluation of student learning. New York: McGraw Hil..

French, J. W. (1966). Schools of thought in judging excellence of English themes. In Anastasi (Ed.) Testing Problems in Perspective. Washington, D.C.: American Council on Education.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 329-338.

Hall, P. B. (1972). Multiple impression marking of essays at a large scale public examination. An unpublished essay submitted in partial fulfillment of requirements for an M. Ed. Department of Education, University of Sydney.

Harkin, J. C. (1983). Measurement of writing. An unpublished paper completed in partial fulfillment of requirements for an M. Ed. University of Sydney.

Hartog, P. (1941). The marking of English essays. London: McMillan.

Harris, W. (1977). Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis for teacher judgments of student writing. Research in Teaching English, 11 175-185.

Huddleston, E. M. (1954). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. Journal of Experimental Education, 22, 165-205.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood. Chicago: International Educational Services.

Kavanagh, M. J., Mackinney, A. C. & Wolins, L. (1981). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 25, 34-49.

Kenny, D. A. (1979). Correlation and causality. New York: Wiley.

Marsh, H. W., Barnes, J., & Hocevar, D. (in press). Self-other agreement on multidimensional self-concept ratings: Factor analysis and multitrait-multimethod analysis. Journal of Personality and Social Psychology.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20, 231-248.

Marsh, H. W. & Hocevar, D. (in press). The factorial invariance of students' evaluations of college teaching. American Educational Research Journal.

Marsh, H. W. & Hocevar, D. (1984). Incorporating item level data into multitrait-multimethod analysis: An application of second order

confirmatory factor analysis. (A paper submitted for publication). Department of Education, University of Sydney, Australia.

Marsh, H. W., Smith, I. D., Barnes, J., & Butler, S. (1983). Self-concept: Reliability, dimensionality, validity, and the measurement of change. Journal of Educational Psychology, 75, 772-790.

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

Michael, W. B., Cooper, T., Shaffer, P., & Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance by professors in English and by professors in other disciplines. Educational and Psychological Measurement, 40, 183-195.

Moss, P. A., Cole, N. S. & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7 and 10. Journal or Educational Measurement, 19, 37-47.

Phillips, G E. (1948). The marking of children's essays. Forum of Education, 7, 19-29.

Quellmalz, E. S., Capell, F. L., and Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 19, 241-258.

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

Smith, L. S. (1979). Measures of high school students' expository writing: Direct and indirect strategies. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Thompson, G. & Bailes, S. (1926). The reliability of essay marks. Forum of Education, 1926, 4, 85-91.

Wiseman, S., & Wrigley, S. (1958). Essay-reliability: The effect of choice of essay title. Educational and Psychological Measurement, 18, 129-138.

TABLE 1

Correlations Among Total and Overall Ratings By Different Teachers

	Mean	SD	T1	T2	T3	T4	T5	T6	O1	O2	O3	O4	O5	O6	E1	E2	E3	C1
Total Scores																		
Student Teachers																		
Total1	36.55	7.60	100															
Total2	38.09	5.64	71	100														
Total3	38.11	8.24	69	68	100													
Master Teachers																		
Total4	35.04	11.23	72	74	77	100												
Total5	33.52	11.64	68	69	77	78	100											
Total6	36.86	8.63	73	73	76	7[illegible]	88	100										
Overall Ratings																		
Student Teachers																		
Over1	6.37	1.37	93	69	67	71	69	69	100									
Over2	6.42	1.24	70	95	66	74	68	71	67	100								
Over3	6.73	1.48	65	64	94	71	71	72	63	62	100							
Master Teachers																		
Over4	5.87	1.90	71	73	72	99	78	71	70	72	70	100						
Over5	5.50	2.26	64	65	75	74	96	65	65	65	69	73	100					
Over6	6.01	1.59	70	69	72	68	67	96	67	70	67	68	68	100				
Early Ratings																		
Master Teachers																		
Early4	53.58	17.83	67	72	71	87	72	69	66	69	66	87	67	65	100			
Early5	60.68	19.54	66	68	71	80	75	68	64	66	64	79	71	66	73			
Early6	64.48	13.81	71	67	73	74	67	84	67	65	71	75	64	81	69	71	100	
Criterion																		
School Perf	67.62	12.95	69	58	72	73	70	70	65	62	67	73	68	67	72	65	68	1

Note: All coefficients, presented without decimals, are statistically significant. The numbers at the end of each variable name refer to raters where 1, 2, and 3 were student teachers, and 4, 5, and 6 were master teachers.

TABLE 2

MTMM: Student Teacher and Master Teacher Ratings of Six Essay Traits

	M1	S1	O1	W1	C1	Q1	M2	S2	O2	W2	C2	Q2	M3	S3	O3	W3	C3	Q3	M4	S4	O4	W4	C4	Q4	M5	S5	O5	W5	C5	Q5	M6	S6	O6	W6	C6	Q6
M1	100																																			
S1	77	100																																		
O1	59	63	100																																	
W1	69	68	65	100																																
C1	55	56	73	62	100																															
Q1	61	57	64	69	72	100																														
M2	62	55	47	52	45	45	100																													
S2	52	57	51	48	45	41	70	100																												
O2	44	47	42	46	46	39	63	65	100																											
W2	49	48	47	62	47	49	58	55	59	100																										
C2	38	41	34	47	41	50	43	39	47	60	100																									
Q2	55	55	44	60	50	52	66	62	63	77	71	100																								
M3	58	57	48	59	40	47	58	49	37	55	41	50	100																							
S3	48	45	49	55	43	45	58	59	41	57	38	51	81	100																						
O3	50	52	50	52	51	42	57	56	48	51	37	48	67	67	100																					
W3	53	57	53	67	56	55	56	50	41	65	47	60	74	72	69	100																				
C3	41	40	52	51	51	51	47	43	36	44	32	46	62	66	67	70	100																			
Q3	47	41	46	53	50	54	54	38	34	52	45	56	65	70	68	72	76	100																		
M4	61	60	55	66	48	53	63	51	45	64	46	62	75	64	63	67	55	59	100																	
S4	62	66	58	67	49	54	63	59	51	65	44	65	74	70	64	67	52	59	90	100																
O4	56	58	52	63	48	55	58	56	56	65	49	65	66	62	62	65	54	58	86	86	100															
W4	59	58	54	70	49	55	60	54	58	66	48	63	72	67	60	72	57	57	86	86	85	100														
C4	55	55	46	65	45	57	57	49	48	62	54	68	67	62	57	66	63	57	80	81	81	89	100													
Q4	63	62	54	73	50	56	60	54	49	65	52	68	71	67	60	71	53	59	85	87	84	91	90	100												
M5	63	58	46	58	45	54	62	53	42	54	46	56	66	61	61	68	50	59	69	71	65	65	63	66	100											
S5	58	59	46	60	49	52	64	59	44	53	42	57	67	66	63	69	57	65	66	68	64	64	62	66	89	100										
O5	52	50	45	58	41	48	51	42	44	56	47	54	62	60	63	67	53	62	68	68	67	65	63	66	81	83	100									
W5	57	58	45	67	46	56	51	42	42	58	46	58	62	59	60	68	56	64	71	73	68	72	70	74	75	76	77	100								
C5	53	50	44	61	45	57	52	45	42	60	53	62	58	57	57	64	51	59	66	67	63	69	69	66	76	74	81	83	100							
Q5	56	54	45	61	42	53	55	48	43	58	53	60	63	63	60	69	51	60	69	72	68	70	67	71	84	82	86	84	88	100						
M6	70	65	51	60	46	49	62	59	49	54	36	58	62	63	57	56	46	50	60	66	52	56	56	61	63	66	54	58	55	59	100					
S6	67	64	50	58	45	49	61	56	50	58	42	58	62	63	62	63	55	60	58	63	54	54	55	60	64	68	60	62	55	60	89	100				
O6	53	52	52	53	46	44	48	51	58	55	38	50	51	55	62	55	50	49	58	63	60	55	54	61	45	53	57	51	48	52	63	69	100			
W6	56	54	49	62	52	59	56	46	51	60	42	57	63	63	58	70	63	67	63	65	62	61	57	63	55	60	55	63	54	56	68	79	72	100		
C6	42	40	49	53	48	55	49	35	58	57	44	53	46	47	48	49	55	49	53	54	56	50	53	54	36	40	45	48	43	42	49	57	74	76	100	
Q6	54	48	48	61	49	59	54	46	51	60	46	56	62	61	58	61	59	65	60	63	61	57	57	60	52	58	55	61	55	57	71	77	73	88	77	100

Note: All coefficients are presented without decimal points. Each variable is labelled with a letter-number combination where the letters stand for traits (M=mechanics, S=sentence structure, O=organization, W=word usage, C=content/ideas, Q=quality of style) and numbers stand for raters (1, 2 and 3 are student teachers; 4, 5, and 6 are master teachers).

TABLE 3

Comparisons Involving the Second Campbell and Fiske Guideline: Number and Percentage Rejections

	Comparisons Involving					
	Student Teachers		Master Teachers		All Teachers	
Trait	N	%	N	%	N	%
Mechanics	2	7	9	30	23	15
Sentence Structure	7	23	8	27	35	23
Organization	17	57	9	30	64	43
Word Usage	0	0	6	20	15	10
Content/Ideas	20	67	18	60	91	61
Quality of Style	3	10	3	10	48	32

TABLE 4

ANOVA Analyses of MTMM: Combined, Student Teachers, Master Teachers

	Combined				Master Teacher				Student Teacher			
Source	df	MS	F-Ratio	Var	df	MS	F-Ratio	Var	df	MS	F-Ratio	Var
Cases (convergence)	138	24.7	164.4	.68	138	12.2	68.5	.68	138	10.2	35.0	.55
C x Trait(T) (divergence)	690	0.5	3.4	.06	690	0.2	1.4	.01	690	0.5	1.6	.06
C x Method(M) (halo)	690	1.1	7.3	.16	276	1.4	7.6	.20	276	1.4	4.7	.18
C x T x M (error)	3450	0.2		.15	1380	0.2		.18	1380	0.3		.29

Note: All effects are statistically significant ($p < .01$) except the divergence effect for Master Teachers ($p > .05$). Variance Components (Var) are defined as described by Marsh and Hocevar (1983).

TABLE 5

Summary of Models Designed To Explain MTMM

Model	Chi-square	df	ratio	RMS	Coef d
0 -- Null	6407	630	10.17	.572	.000
1 -- 6 Traits & 6 Methods	845	528	1.60	.051	.868
2 -- 1 General Factor	2429	593	4.10	.076	.621
3 -- 6 Traits	2294	579	3.96	.073	.642
4 -- 6 Methods	1296	579	2.24	.051	.800
5 -- 6 Traits & 1 General	1776	543	3.27	.062	.723
6 -- 6 Methods & 1 General	997	543	1.84	.040	.844

TABLE 6
LISREL Estimates For Model With 6 Methods, 6 Traits, and a 13th Factor Representing an External Validity Criterion.

Variables	Factor Loadings													
	m1	m2	m3	m4	m5	m6	t1	t2	t3	t4	t5	t6	V	E/U
M1	.66	00	00	00	00	00	50	00	00	00	00	00	00	25
S1	65	00	00	00	00	00	00	54	00	00	00	00	00	29
O1	79	00	00	00	00	00	00	00	25	00	00	00	00	34
W1	75	00	00	00	00	00	00	00	00	47	00	00	00	22
C1	80	00	00	00	00	00	00	00	00	00	15	00	00	36
Q1	82	00	00	00	00	00	00	00	00	00	00	23	00	29
M2	00	72	00	00	00	00	45	00	00	00	00	00	00	31
S2	00	64	00	00	00	00	00	50	00	00	00	00	00	37
O2	00	71	00	00	00	00	00	00	42	00	00	00	00	35
W2	00	74	00	00	00	00	00	00	00	46	00	00	00	29
C2	00	61	00	00	00	00	00	00	00	00	48	00	00	44
Q2	00	78	00	00	00	00	00	00	00	00	00	52	00	18
M3	00	00	78	00	00	00	35	00	00	00	00	00	00	26
S3	00	00	79	00	00	00	00	33	00	00	00	00	00	27
O3	00	00	78	00	00	00	00	00	38	00	00	00	00	28
W3	00	00	81	00	00	00	00	00	00	37	00	00	00	20
C3	00	00	84	00	00	00	00	00	00	00	08	00	00	32
Q3	00	00	85	00	00	00	00	00	00	00	00	15	00	27
M4	00	00	00	90	00	00	24	00	00	00	00	00	00	12
S4	00	00	00	88	00	00	00	32	00	00	00	00	00	11
O4	00	00	00	91	00	00	00	00	21	00	00	00	00	13
W4	00	00	00	86	00	00	00	00	00	36	00	00	00	09
C4	00	00	00	82	00	00	00	00	00	00	44	00	00	12
Q4	00	00	00	84	00	00	00	00	00	00	00	40	00	09
M5	00	00	00	00	81	00	39	00	00	00	00	00	00	14
S5	00	00	00	00	82	00	00	38	00	00	00	00	00	15
O5	00	00	00	00	87	00	00	00	34	00	00	00	00	14
W5	00	00	00	00	78	00	00	00	00	39	00	00	00	21
C5	00	00	00	00	82	00	00	00	00	00	40	00	00	17
Q5	00	00	00	00	85	00	00	00	00	00	00	38	00	11
M6	00	00	00	00	00	62	73	00	00	00	00	00	00	11
S6	00	00	00	00	00	72	00	63	00	00	00	00	00	11
O6	00	00	00	00	00	76	00	00	48	00	00	00	00	24
W6	00	00	00	00	00	93	00	00	00	28	00	00	00	11
C6	00	00	00	00	00	84	00	00	00	00	19	00	00	30
Q6	00	00	00	00	00	92	00	00	00	00	00	27	00	13
Val	00	00	00	00	00	00	00	00	00	00	00	00	100	00

Factor Correlations

	m1	m2	m3	m4	m5	m6	t1	t2	t3	t4	t5	t6	V
m1	100												
m2	71	100											
m3	72	71	100										
m4	72	73	79	100									
m5	66	65	79	77	100								
m6	72	72	78	67	65	100							
t1	00	00	00	00	00	00	100						
t2	00	00	00	00	00	00	99	100					
t3	00	00	00	00	00	00	64	76	100				
t4	00	00	00	00	00	00	70	74	61	100			
t5	00	00	00	00	00	00	57	58	62	84	100		
t6	00	00	00	00	00	00	78	79	68	93	98	100	
V	66	56	69	68	66	67	28	25	14	30	17	22	100

Note: The model illustrated here contains 6 method factors (m1 to m6), 6 trait factors (t1 to t6), and a 13th factor corresponding to the external validity criterion (V). It differs from the design of model 1 (see Table 5) only in that the external validity criterion was added as a 13th factor. "E/U" stands for the error/uniqueness component.

TABLE 7

MTMM matrix for Summed Master Teacher and Student Teacher Ratings, and Correlations with Overall Ratings, Total Ratings, and the Validity Criterion

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Student Teachers																		
1 - STMech	100																	
2 - STSent	86	100																
3 - STOrg	77	82	100															
4 - STWord	79	80	77	100														
5 - STCont	69	71	80	79	100													
6 - STQual	76	74	75	83	87	100												
Master Teachers																		
7 - MTMech	87	82	74	80	68	76	100											
8 - MTSent	86	86	77	81	70	79	95	100										
9 - MTOrg	76	76	79	80	70	75	85	88	100									
10 - MTWord	80	78	74	86	75	81	87	89	87	100								
11 - MTCont	75	72	73	82	75	82	82	83	87	91	100							
12 - MTQual	81	78	74	85	74	81	90	91	91	94	92	100						
Overall/Total Ratings																		
13 - STOverall	88	86	86	89	89	93	84	86	81	86	86	86	100					
14 - STTotal	90	91	90	92	89	91	86	88	84	88	85	87	98	100				
15 - MTOverall	83	80	79	84	75	82	92	94	93	95	93	97	88	89	100			
16 - MTTotal	85	83	79	86	76	83	95	96	94	96	94	97	89	91	99	100		
17 - MTEarly	81	80	75	84	71	79	86	88	86	91	85	90	85	87	91	92	100	
Validity Criterion																		
18 - Essay Test	74	70	63	73	62	69	75	75	71	80	72	76	74	76	78	79	76	100

NOTE: All coefficients, presented without decimal points, are statistically significant. MT and ST refer to ratings by master teachers and student teachers, which are summed across ratings by the three teachers in each group.