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 perhaps better able to differentiate among different components of writing effectiveness.
Multidimensional Evaluations of Writing Effectiveness

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

28 March, 1984

Running Head: Writing Effectiveness

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. W. Marsh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.
Multidimensional Evaluations of 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 Markings:

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
Writing Effectiveness

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 appear to represent "the limit of the extent of agreement one can generally expect between single judges marking an 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 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 0.65 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
Writing Effectiveness

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 0.89 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, Capeli, 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
Writing Effectiveness

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-
Writing Effectiveness

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
Writing Effectiveness

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 straight-forward 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; Haruyama & 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., $O_1$ & $T_1$) 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:
Writing Effectiveness

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.
Writing Effectiveness

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


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


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


### TABLE 1
Correlations Among Total and Overall Ratings By Different Teachers

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>O1</th>
<th>O2</th>
<th>O3</th>
<th>O4</th>
<th>O5</th>
<th>O6</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total1</td>
<td>36.55</td>
<td>7.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total2</td>
<td>38.09</td>
<td>5.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total3</td>
<td>38.11</td>
<td>8.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total4</td>
<td>35.04</td>
<td>11.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total5</td>
<td>33.52</td>
<td>11.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total6</td>
<td>36.86</td>
<td>8.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall Ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over1</td>
<td>6.37</td>
<td>1.37</td>
<td>93</td>
<td>62</td>
<td>67</td>
<td>71</td>
<td>69</td>
<td>69</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over2</td>
<td>6.42</td>
<td>1.24</td>
<td>70</td>
<td>95</td>
<td>66</td>
<td>74</td>
<td>68</td>
<td>71</td>
<td>67</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over3</td>
<td>6.73</td>
<td>1.48</td>
<td>65</td>
<td>64</td>
<td>94</td>
<td>71</td>
<td>71</td>
<td>72</td>
<td>63</td>
<td>62</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over4</td>
<td>5.87</td>
<td>1.90</td>
<td>71</td>
<td>73</td>
<td>77</td>
<td>71</td>
<td>70</td>
<td>72</td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over5</td>
<td>5.50</td>
<td>2.26</td>
<td>71</td>
<td>73</td>
<td>77</td>
<td>71</td>
<td>70</td>
<td>72</td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over6</td>
<td>6.01</td>
<td>1.59</td>
<td>73</td>
<td>69</td>
<td>72</td>
<td>68</td>
<td>67</td>
<td>67</td>
<td>70</td>
<td>67</td>
<td>68</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Early Ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Master Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early1</td>
<td>53.58</td>
<td>17.83</td>
<td>67</td>
<td>72</td>
<td>71</td>
<td>87</td>
<td>72</td>
<td>69</td>
<td>69</td>
<td>66</td>
<td>66</td>
<td>87</td>
<td>67</td>
<td>65</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early2</td>
<td>60.68</td>
<td>19.54</td>
<td>66</td>
<td>68</td>
<td>71</td>
<td>80</td>
<td>75</td>
<td>68</td>
<td>64</td>
<td>66</td>
<td>64</td>
<td>79</td>
<td>71</td>
<td>66</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early3</td>
<td>64.48</td>
<td>13.81</td>
<td>71</td>
<td>67</td>
<td>73</td>
<td>74</td>
<td>67</td>
<td>84</td>
<td>67</td>
<td>65</td>
<td>71</td>
<td>75</td>
<td>64</td>
<td>81</td>
<td>69</td>
<td>71</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td><strong>Criterion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School Perf</td>
<td>67.62</td>
<td>12.95</td>
<td>69</td>
<td>58</td>
<td>72</td>
<td>73</td>
<td>70</td>
<td>70</td>
<td>65</td>
<td>62</td>
<td>67</td>
<td>73</td>
<td>68</td>
<td>67</td>
<td>72</td>
<td>65</td>
<td>68</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>59</td>
<td>57</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>77</td>
<td>68</td>
<td>65</td>
<td>68</td>
<td>65</td>
<td>68</td>
<td>65</td>
<td>68</td>
<td>65</td>
<td>68</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>56</td>
<td>73</td>
<td>61</td>
<td>67</td>
<td>44</td>
<td>45</td>
<td>57</td>
<td>45</td>
<td>61</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>62</td>
<td>55</td>
<td>47</td>
<td>32</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>47</td>
<td>42</td>
<td>42</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>45</td>
<td>49</td>
<td>53</td>
<td>48</td>
<td>45</td>
<td>57</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Trait</th>
<th>Student Teachers</th>
<th>Master Teachers</th>
<th>All Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Mechanics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>7</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>Organization</td>
<td>17</td>
<td>57</td>
<td>9</td>
</tr>
<tr>
<td>Word Usage</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Content/Ideas</td>
<td>20</td>
<td>67</td>
<td>18</td>
</tr>
<tr>
<td>Quality of Style</td>
<td>3</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>
### Writing Effectiveness 26

**TABLE 4**

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

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi-square</th>
<th>df</th>
<th>ratio</th>
<th>RMS</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 -- Null</td>
<td>6407</td>
<td>630</td>
<td>10.17</td>
<td>.572</td>
<td>.000</td>
</tr>
<tr>
<td>1 -- 6 Traits &amp; 6 Methods</td>
<td>845</td>
<td>528</td>
<td>1.60</td>
<td>.051</td>
<td>.868</td>
</tr>
<tr>
<td>2 -- 1 General Factor</td>
<td>2429</td>
<td>593</td>
<td>4.10</td>
<td>.076</td>
<td>.621</td>
</tr>
<tr>
<td>3 -- 6 Traits</td>
<td>2294</td>
<td>579</td>
<td>3.96</td>
<td>.073</td>
<td>.642</td>
</tr>
<tr>
<td>4 -- 6 Methods</td>
<td>1296</td>
<td>579</td>
<td>2.24</td>
<td>.051</td>
<td>.800</td>
</tr>
<tr>
<td>5 -- 6 Traits &amp; 1 General</td>
<td>1776</td>
<td>543</td>
<td>3.27</td>
<td>.062</td>
<td>.723</td>
</tr>
<tr>
<td>6 -- 6 Methods &amp; 1 General</td>
<td>997</td>
<td>543</td>
<td>1.84</td>
<td>.040</td>
<td>.844</td>
</tr>
</tbody>
</table>
### Table 6
LISREL Estimates For Model With 6 Methods, 6 Traits, and a 13th Factor Representing an External Validity Criterion.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor Loadings</th>
<th>E/U</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>66 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 25 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>m2</td>
<td>65 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 29 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>m3</td>
<td>75 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 34 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>m4</td>
<td>80 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 36 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>m5</td>
<td>82 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 39 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>m6</td>
<td>66 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 31 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t1</td>
<td>65 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 35 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t2</td>
<td>66 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 37 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t3</td>
<td>75 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 46 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t4</td>
<td>80 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 50 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t5</td>
<td>82 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 52 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>t6</td>
<td>66 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 54 00 00 00 00 00 00 00 00 00 00</td>
</tr>
<tr>
<td>V</td>
<td>70 00 00 00 00 00 00 00 00 00 00 00</td>
<td>00 61 00 00 00 00 00 00 00 00 00 00</td>
</tr>
</tbody>
</table>

| Factor Correlations | m1  m2  m3  m4  m5  m6  t1  t2  t3  t4  t5  t6  V |
|---------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| m1                  | 100   |       |       |       |       |       |       |       |       |       |       |       |
| m2                  |       | 71 100|       |       |       |       |       |       |       |       |       |       |
| m3                  |       |       | 72 100|       |       |       |       |       |       |       |       |       |
| m4                  |       |       |       | 72 73|       |       |       |       |       |       |       |       |
| m5                  |       |       |       |       | 66 65|       |       |       |       |       |       |       |
| m6                  |       |       |       |       |       | 65 72|       |       |       |       |       |       |
| t1                  |       |       |       |       |       |       | 00 00|       |       |       |       |       |
| t2                  |       |       |       |       |       |       |       | 00 00|       |       |       |       |
| t3                  |       |       |       |       |       |       |       |       | 00 00|       |       |       |
| t4                  |       |       |       |       |       |       |       |       |       | 00 00|       |       |
| t5                  |       |       |       |       |       |       |       |       |       |       | 00 00|       |
| t6                  |       |       |       |       |       |       |       |       |       |       |       | 00 00|
| V                   |       |       |       |       |       |       |       |       |       |       |       | 66 56|

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.
**Writing Effectiveness 29**

**TABLE 7**

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

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

**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.