2000

Assessing business writing: An examination of scoring methods, writing sample complexity, and rating variability

Larry Weitzel

ASSESSING BUSINESS WRITING: AN EXAMINATION OF SCORING METHODS, WRITING SAMPLE COMPLEXITY, AND RATING VARIABILITY

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Psychology:

Industrial/Organizational

---

by

Larry Weitzel

December 1999

ASSESSING BUSINESS WRITING: AN EXAMINATION OF SCORING METHODS, WRITING SAMPLE COMPLEXITY, AND RATING VARIABILITY

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Larry Weitzel

December 1999

Approved by:

[redacted]

Kenneth S. Shultz, Chair, Psychology

12/10/99

Date

[redacted]

Matt L. Riggs, Loma Linda University

[redacted]

Thung-Rung Lin, Los Angeles Unified School District

ABSTRACT

Defining "good writing" is difficult because there is no set answer as to what constitutes good writing. However, good writing is essential for business. Writing has been evaluated using several techniques. This paper evaluates holistic and analytical assessment of writing samples. It was hypothesized that there would be a difference between the patterns of correlations for simple and complex writing assignments when holistic and analytical methods of analysis were correlated with multiple choice test segments and that there would be less variability among the raters of holistically scored papers than among the raters of analytically scored papers. For hypothesis one, one prediction was fully supported, two were partially supported and two were not supported. In addition, in this paper we did determine that there was less variability among holistic raters than among analytical raters.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Kenneth Shultz, my thesis advisor for all his help and effort in completing this project. I would also like to acknowledge Dr. T. R. Lin, a member of the committee and an excellent friend, for his suggestion of the thesis topic, providing a resource for archived data and for his continued encouragement with this thesis. I would like to acknowledge the large Western School District for the use of their archived data.

This thesis is dedicated to Nancy Weitzel, my wife and my family who stood beside me and encouraged me, and Dr. Kenneth Shultz my professor, advisor and friend from California State University San Bernardino, who went far beyond the call of duty to help when help was needed.

TABLE OF CONTENTS

INTRODUCTION

Defining ''good writing'' is difficult because there is no set answer as to what is ''good writing'' (Quellmalz & Burry, 1983). What would be good writing in one context may not be in another. For example, if an individual were asked to produce a business letter and used the format for a personal note, it would be considered inappropriate. If a letter were to contain specific important information, and the information was omitted it would be considered deficient. If the letter had several errors in grammar and/or punctuation, it would be considered poorly written. Therefore, to write well, the tone, style, and content must be appropriate for the situation.

Method of Discourse

Studies of students' writing have shown that performance on a writing-ability test is greatly influenced by the discourse required (Quellmalz & Burry, 1983). Discourse is defined by Merriam-Webster (1991) as the, ''verbal interchange of ideas'' and the ''formal and usual extended expression of thought on a subject'' (p. 361). Kinneavy (1971) identified four forms of discourse: expressive, where the writer's point of view is most important, and there is little concern for the audience, as

in a diary; literary or narrative, where the focus is on the product and the written work should be appreciated for its own merits, as in a novel; persuasive, which focuses on an attempt to influence the audience, such as a sales presentation; and reference or expository, which is a presentation of facts with no attempt to convey biases. Additionally, expository writing requires the producer to consider facts and ideas and support them with a detailed, logical train of thought, as in a business memo designed to convey information (Alaska State Department of Education, 1989).

Business writing is usually expository. The skillful and logical presentation of ideas is necessary for companies to succeed because ideas which are poorly communicated will not be acted upon. For example, a poorly written memo outlining the need for financial restraint may not convey the needed urgency, and unnecessary expenditures may be made. Hence, it is important for companies to employ individuals who can write appropriately.

It is important for businesses to have a match between the job requirements and the skills of a prospective employee. Therefore, a selection procedure which evaluates an individual's expository writing skills would prove

helpful in selecting the individual best suited for a job which requires expository writing.

How writing is taught

In order to understand how to evaluate writing, one needs to understand how writing was taught and how the teaching of writing influenced the evaluation process. In the past, red penciling papers without providing comments was considered appropriate. As methods of instruction developed, two methods of teaching writing emerged, the product method and the process method.

First came product-oriented writing. It focused on the finished product. Students were given examples of good writing to read, study and analyze. Subsequently, they were given writing assignments and told they could use the readings as models while composing their own essays. The students would write a paper and submit it to the instructor for evaluation. The instructor would read and make notes on the paper, then return it to the student to be rewritten. Writing and rewriting of the paper would continue until the student developed a quality product. This method gave students the ability to compare their product or writing with that of more accomplished authors.

Students learned by example (Hayes & Flower, 1986).

This method of instruction emphasized the final product and asserted that the finished product was most important. As class size grew and instruction became more sophisticated, the direction of writing instruction changed. Instead of emphasizing the product, the importance of the process became more important. Teachers began to emphasize the individual components of the writing process. Activities that targeted the development of specific skills such as brainstorming or idea development, spelling, grammar, and punctuation became popular (Hayes & Flower, 1986). The process method asserts that individual components are more important than the whole (White, 1985).

Hillocks (1984) performed a meta-analysis of writing instruction research to explain, ''the variability among the characteristics of the treatment in relation to the variability of their effect sizes.'' The effect size, or size of the difference in standard deviations, and homogeneity, which tests whether the effect size estimate is greater than would be expected if all the studies had the same effect size, were used as criterion to evaluate the factors common to the 60 studies evaluated. The factors evaluated were: duration of instruction, length of the study; mode of instruction, the teacher/classroom

activities or the interaction between students and teachers and focus of instruction; the activities, such as studying grammar and mechanics or sentence combining. Hillocks concluded that a process orientation is more successful in the development of writing skills than a product oriented approach. However, Applebee (1984) has claimed the effectiveness of process-based instruction is limited because of its focus on the activities or the process of writing without regard to the purpose of the writing. The process method emphasizes building a student's competence in individual factors that are considered important. To measure a student's progress educators need to be able to measure these factors. Multiple-choice tests, a form of indirect assessment, are ideal for measuring the specific factors.

Evaluating Writing

Indirect Writing Assessment. Indirect measurement is measuring one content area and extending the results to another area. For example, measuring a person's ability to spell and claiming that indicates the person can probably write well would be an indirect measurement. Proponents of the indirect method contend that the ability of a writer can be measured by analyzing the individual parts or

competencies of the writer. Therefore, measurement usually involves objective measures, multiple-choice questions centering around areas which are easily evaluated, such as spelling, grammar and punctuation (Quellmalz, 1986). This type of measurement is particularly well suited to the evaluation of specific areas because measurement items can be carefully selected to match the content domain. Honey (1990) claimed multiple-choice tests are often superior to writing samples for testing editing skills of individuals.

However, writing is not editing, or the correct usage of grammar and punctuation; writing is assembling ideas or facts in a logical way to convey meaning. Many educators became dissatisfied with the results of indirect measurement as their awareness of the limitations of indirect measurement became apparent. Namely, indirect measurement allows for the evaluation of editing skills, spelling, and grammar (Quellmalz, 1986), but left other areas of the writing process such as organization and analytical ability unevaluated. This dissatisfaction led to the development of direct measures of writing assessment.

Direct Writing Assessment. Many of today's educators are dissatisfied with the quality of writing produced by students (Hayes & Flower, 1986; Reutzel & Hollingsworth,

1988). McCaig (1982) stated the inability of students to compose an intelligible, coherent passage of written English is a national disgrace and a source of outrage in communities throughout the country. This may be one reason that more than two-thirds of the states have adopted writing competency tests as a requirement for graduation from high school (Calkins, 1985). Competency tests movement encompasses direct measurements because they require the individual to actually produce a product that is evaluated based on pre determined criteria. A search of the literature revealed three direct methods of assessment typically used to evaluate papers: primary trait, holistic, and analytical.

Methods of Direct Writing Assessment

Primary Trait. The primary trait method is based on the premise that all writing is directed to a particular audience, and that successful writing will have an effect on that audience. Primary trait papers are test specific; therefore, the primary trait being assessed will differ from one test to another. The instructions given concerning the type of paper to be composed and the aim of the paper will determine the primary trait (Spandel, 1981).

For example, the purpose of the paper may be to

convince an individual to purchase a widget. Therefore, the persuasiveness of the paper would be the primary trait that is being examined. The scoring of primary trait papers is similar to that of holistic papers in that they both look at the paper as a whole (Hansen, 1992), but differs in that the primary trait method focuses on evaluating a specific concept.

Holistic Scoring. Holistic scoring contends that writing is a process that should not be broken down into individual components. The whole is more important than the sum of its parts. Holistic scoring as defined by White (1985) and Spandel (1981) with the assumption that each component is related to all other components; and they cannot be separated. Holistic scoring is used by educational assessment organizations such as the Educational Testing Service (ETS), Advanced Placement (AP) Test, College Level Examination Program, and California Assessment Program (White, 1985).

Holistic scoring provides a way to directly analyze writing that is nearly as cost effective and psychometrically reliable as indirect measurement (Honey, 1990). It is considered quick and reliable, and it has high face validity (Quellmalz, 1986). Frequently, raters can

read single paragraph papers in 30 seconds to 1.5 minutes and a multiple paragraph paper in less than 2 minutes (Quellmalz, 1986; Quellmalz & Burry, 1983). Scoring is simple. Raters read the paper and then quickly score it while they still have a clear impression of the work. The score is based on the overall impression of the paper and considers both the strengths and weaknesses of the paper. Sentence construction, spelling, vocabulary, grammar, and other factors influence the reader only when excellent or poor usage adds or detracts from the overall readability of the paper (Spandel, 1981).

Scoring is usually on a 1-5 scale, where one indicates a very poorly written paper and five indicates a very well written paper. To establish consistent standards at least three benchmark papers are selected from the writing samples. One paper will be of low quality and is given a score of one. Another paper will be of moderate or average quality and is given a score of three and a third paper, of high quality is given the rating of a five. These papers are used as benchmarks and set the criterion for rating the other papers. Unless it is a continuous testing program, the writing samples are usually compared to benchmark papers from the same sample group. The benchmark papers

change with every administration of the writing exercise, therefore, the measurement level is ordinal. Subjects are rank ordered; therefore, the results of one test usually cannot be compared directly with other administrations of the same test (White, 1985).

The inability to compare different administrations of the test raises the issue of the reliability of the raters in holistic scoring. Honey (1990) found the inter-rater reliability of the raters to be .85 on a sample of more than 5000 writing assignments. White (1985) found similarly high reliability in his analysis of two different writing samples. One sample had an inter-rater reliability of .78, and the other had an inter-rater reliability of .85.

Schwartz et al.(1999) used generalizability theory to predict the reliability of holistic scoring and found the reliability of three raters to be .94. Holistic scoring has high reliability, and it also has high face validity because individuals can see the logical connection between the scoring of writing exercise and the quality of written work they performed.

Because holistic scoring is easy, has high inter-rater reliability, and is cost effective, it is often used to evaluate writing when resources are scarce. However,

holistic assessment gives writers little or no direct feedback on their writing ability. Therefore, analytical scoring may be more appropriate when feedback is important, such as when a teacher is using feedback as a way to improve a student's writing.

Analytical Scoring. Analytical scoring is multidimensional (Hansen, 1992) rather than impressionistic. This method assumes it is possible to determine specific characteristics prior to assessment, such as content, organization, and mechanics, which are indicative of good writing. These characteristics then form the basis for the evaluation of writing samples (Alaska Department of Education, 1989). The most commonly accepted format for writing evaluation is based on the works of Quellmalz, Capell, and Chou (1982) and Quellmalz (1986). In analytical scoring, the writing assignment is developed in the same manner as the holistic assignment; however, the format of evaluation varies. Predefined traits, which have been deemed important, are evaluated individually (Hansen, 1992; Spandel, 1981). For example, in a school setting, independent criteria such as ideas and content, organization, voice, word choice, sentence structure, and mechanics may be evaluated (Alaska State Dept. of

Education, 1989).

In a business setting, a writing sample may be judged based on the requirements of the job and the method of discourse most frequently used in the position. For example, a paper may be judged on analytical skills, organization and clarity, spelling, punctuation, and grammar if these are important components of the job (Hoffman & Hoffman, 1990). The inter-rater reliability of analytically scored papers, in an educational setting, is typically .84-.92 and scores of .94-.96 have been achieved when in-depth training is provided (Hoffman & Hoffman, 1990).

One of the common criticisms of analytical scoring is that it is more expensive than indirect or holistic measurement because it requires much more time to evaluate the writing sample. Indirect measures, such as multiple-choice tests, may be machine scored in one second (Hoffman & Holden, 1990). However, when other costs such as test development and lack of test applicability, are factored into the equation, the costs are more similar to those of holistic scoring. Holistic scoring of a writing sample may take 30 seconds to 1.5 minutes depending on the length of the sample (Quellmalz, 1986). Although holistic scoring is

a direct measure of a writer's ability, this method does not provide insight into the individual competencies of the writer. An analytically scored paper, consisting of multiple paragraphs, may take from 2 to 5 minutes score, but the richness of the data supplied by analytical scoring may compensate for the increased costs (Hoffman & Holden, 1990; White, 1985).

Godshalk, Swineford and Coffman (1966) were some of the first researchers to empirically establish a relationship between writing and scores on multiple-choice tests. Their study found correlations from .46 to .75 between the cumulative score on five essays and the College Board English Comprehension Test of high school students. Similar results were found by subsequent researchers such as Veal and Tillman (1971). Quelmalz, Capell and Chou (1982) extended this line of research and compared the effects of analytically scored direct assessment of a student's writing with multiple-choice questions. They formed two conclusions. First, that different modes of discourse draw on different skill constructs; and second, that when referring to a student's writing ability it was necessary to reference the mode of discourse.

Writing for Business. Tests of students' writing ability

are relied upon by college and university admissions officers because these tests provide a way to evaluate or possibly predict the future success of the student (White, 1985). For businesses it is also important to select individuals with the highest probability of success. Employment tests attempt to select employees with the requisite knowledge, skills, and abilities for the job. An employee who is unskilled will require training and may make errors, both of which are costly to the company. Also, an employee who is unskilled may be less successful and may be more likely to withdraw from the organization either voluntarily or involuntarily. To minimize the expense of false positive selection decisions, managers typically use a variety of techniques to evaluate the quality and fit of an individual.

For example, organizations often use interviews, multiple choice knowledge tests and even personality tests to help predict future performance of a perspective employee. However, these ''tests'' are only indirect measures of an individual's ability because the tests measure verbal or written responses to hypothetical situations rather than actual performance (Gatewood & Feild, 1987).

Direct measures or performance tests are also used by businesses and provide a way to measure an individual's skill and ability level. Gatewood and Feild (1987) stated performance tests place an individual in a situation similar to the work environment and require a work sample. Direct measures may be either simple or complex. Simple measures frequently require demonstration of rudimentary skills, whereas, complex skills often require a demonstration of proficiency in the area being evaluated. For example, if a candidate were asked to demonstrate carpentry skills he/she may be called on to demonstrate how to construct an inside wall, which would be a simple project. A complex task may require the candidate to frame an outside wall that includes a door and a window. The latter would be a complex assignment because it requires the same construction techniques as the simple assignment; plus, it requires several additional parts to be plumb, to size for the door and window; and the wall would require the additional strength of a load-bearing wall.

Evaluating writing in business settings

Writing samples are a direct measurement of an individual's writing ability, and writing a job-relevant memo is an evaluation of an individual's performance on the

job (Campbell, McCloy, Oppler, & Sager, 1993). Assessment centers and other pre-employment screening procedures frequently require simple and complex writing samples. A simple writing sample may consist of a short one-page memo or letter; whereas, a complex sample may require the analysis and communication of complex information and be two to three pages in length. The more information a candidate is asked to supply, by participating in varying testing formats, the more accurate the evaluation may be. However, to reach correct conclusions about an individual's abilities not only must a sample of the performance or data be collected but it must also be evaluated properly. There are three factors that must be considered when writing samples are evaluated. First, are the criteria; second, is the method of evaluation; and third, is the expertise of the raters. The criteria for writing samples would include the job requirements and the knowledge, skills, and abilities required to do the job as determined by a job analysis. The method of evaluation influences the amount of detail that will be available. Holistic and analytical scoring, the two most common methods may yield different results due to the difference in emphasis in evaluation. For example, a holistically scored paper will

look at the overall quality of the sample, and an analytically scored paper will look at several specific criteria, such as tone, organization, word usage, data analysis, and grammar. Therefore, when using writing samples as a method of evaluating performance, matching the method of discourse with the correct method of analysis will help the manager select individuals who are more apt to perform well.

The third factor is the experience and ability of the raters. For holistically scored papers it has been shown there is little difference between the ratings provided by different raters, and as training increases, the consistency between raters improves (Hoffman & Hoffman, 1990). For analytically scored papers the consistency of the raters has been evaluated mainly in the educational setting, and scores similar to those of holistic raters are typically found (Quellmalz & Burry, 1983; Quellmalz, 1996). However, there is a difference between the raters used in educational and business settings. Writing samples are frequently evaluated in educational settings by English teachers. In business, the individual who analyzes a writing sample will very likely have been trained in another discipline who also use writing skills on their

jobs. Because of the differences in the individual's training received there may be a difference in the accuracy of the evaluations.

Problem Definition

This research was designed to answer two questions. First, what relationships exist among the five factors of the multiple-choice test and the holistic score and analytical score for each factor? Second, how reliable are the scores produced by raters who are from a business setting?

HYPOTHESES

Hypothesis 1 **There will be a difference between the predicted patterns of correlation coefficients for simple and complex writing assignments (see Table A1) when holistic and analytical methods of analysis are correlated with the multiple choice test segments.** It is believed some of the difference between simple and complex writing samples may be due to the difference in focus of holistic verses analytical scoring. Holistic scoring assigns a generalized or impressionistic score to a paper, while analytical scoring assigns a score to specific factors that are believed to be important. The effects of the general "g" factor should be more apparent across factors which are related and where more expression is required. For

example, multiple-choice questions that measure reason/math and statistics should correlate higher with the ability to organize and analyze than with spelling. In a simple writing sample the quantity of information being communicated may also influence the ratings given. When the quantity of information to be communicated is limited, the opportunity to demonstrate competence is also limited; therefore, it may be more difficult for a candidate to demonstrate proficiency.

Hypothesis 2 **There will be less variability among raters of holistically scored papers than among the raters of analytically scored papers, when the training time is short, approximately 1 to 1.5 hour long**. Quellmalz, (1981) trained raters, in an educational setting, to use both holistic scoring and analytical scoring to evaluate a series of papers. Quellmalz found that when holistic raters were given 3.5 hours of training and analytical raters were given 6 hours of training the analytical training produced slightly better results. Both groups were in the 90% range. This result would be consistent with the concept that as training time increases the inter-rater reliability increases.

In business settings raters seldom have the luxury of

4-8 hours of training. More frequently, they will have thirty minutes to one hour of training prior to rating papers. The holistic rater who only has one factor to analyze will have more time per factor than the rater who has five factors to learn. Therefore, the holistic raters would have more consistent ratings, provided equal limited training time.

METHOD

Participants

Participants were 37 individuals who applied for the position of an entry-level technical assistant at a large Western state urban school district in 1994. Information about the race and ethnicity of the participants was collected when individuals applied for the position but was not available for analysis.

Position Requirements

This entry-level technical assistant position required a college degree. After a thorough job analysis was conducted, a content-oriented strategy was utilized to develop the exam plan. This staff job required frequent analysis of information and writing of executive summary reports. The written reports often required the incumbent to judge the criticality of the information collected and

determine what information needed to be included.

Therefore, part of the assessment package for the candidates included a writing project intended to assess candidate's ability to determine the criticality of the information and the ability to effectively organize and communicate this information

Procedures

The exam plan adapted a multiple hurdle design, with a written test, an interview and a writing project. The first hurdle, the written test, consisted of a 100-item multiple-choice test which evaluated five critical factors: reasoning skills, basic statistics and research methods, reading comprehension, interpretation of data, and writing skills. Candidates were given 3 hours and 15 minutes to complete the written exam. The scores on the written test were used as the criteria to determine which candidates would advance to the second hurdle of the exam, the interview. Of the 251 individuals participating in the exam, the score of 39 candidates exceeded the cut off score, which was set based on the job analysis and business necessity. These 39 individuals were then invited to participate in an interview given at a later date. The third hurdle, the writing project, was given immediately

following the written test and required the candidate to prepare a simple letter requesting a meeting with an administrator. The writing samples were not graded until after the interview portion of the exam. The second hurdle, the interview, was conducted by two raters who were trained by the Personnel department. The training covered the rating standards to be used in evaluating a candidates training experience, professional development, and management skills.

After the interview, 37 candidates remained and were asked to produce a second writing sample, the complex writing assignment. For this assignment the candidate was asked to prepare a letter concerning the staffing of a building after an earthquake.

Writing Assignment One, Simple. Two hundred fifty- one candidates participated in Writing Assignment One. Each candidate was given a maximum of one hour to prepare a letter to the Administrator of Region H. The purpose of the letter was to establish a meeting with the administrator to discuss his continual attempts to use funds inappropriately. Provisions were to be made within the letter to allow follow-up and verify a meeting time. The candidates were given six assumptions which were to be

included in the letter:

The writer's name, is Pat Smith.

The supervisor's address is 350 N. Grand Avenue, Los Angeles, CA 90015.

The Administrator of Region H is Mr. Ryan Alexander.

The Administrator's address is 3421 West Second Street Los Angeles, Ca 90004.

The Administrator is at a higher level than the supervisor.

You have never dealt with the Administrator before.

The subjects were also given six additional items, which ''should'' also be included in the letter:

1. Today's date.

2. The supervisor's name, job title, and mailing address.

The Administrator's name, job title, and mailing address.

Greeting.

Reason for the letter and the topic of the meeting.

Request for a meeting and follow-up to meeting request.

(Your supervisor is available during the week of June 27, 1994.)

The final letter was to be signed by the writer's supervisor, the Financial Manager.

Writing Assignment Two, Complex. Only candidates who passed the first two hurdles, the written exam and the

interview, participated in this portion of the selection procedure. Candidates were given a maximum time of 2 hours and 15 minutes for the writing exercise.

The candidates were asked to assume they were a new employee with a local company, and their supervisor had requested they attend a meeting that concerned the safety and use an administrative building that was damaged during a recent earthquake. Because of the damage to the building, weekend shifts in the building were to be limited while the building was being repaired. The specific purpose of the meeting was to discuss how to maintain minimum computer coverage while repairs were being made to the building. The company's mainframe computers and peripheral equipment, which were used for payroll, are housed within the administrative building. Following the meeting, the staff member was to prepare a memo outlining the important facts to the regional vice-president responsible for the project.

The candidates were provided a 29-item information sheet that represented their ''notes'' from the meeting. These notes were to be the basis of their report; however, the candidate was allowed to create supporting information necessary to construct a summary of the meeting or make reasonable assumptions in order to fill in the gaps. Facts

that could not be clearly supported from the notes were not to be included. Notes included such items as:

1. Repair work would require some equipment to be down during the weekends.
2. Minimum staff needed; four per shift, except on weekend prior to warrants, where a full staff is required; however, minimum staffing could be maintained with 8-8-8 on Saturday and 4-4-4 on Sunday.
3. Warrants go out the third Wednesday of each month, and additional staff is usually required for processing.

Evaluation of the Writing Assignments: Scoring Guide

Holistic. Both the simple and complex writing samples were holistically scored on a six-point scale:

The scale was designed to provide a continuum against which the candidate's papers would be evaluated. See Appendix B1 for an example of the Holistic Scoring standards used in the research.

Analytic. An analytical scoring guide was developed to measure specific elements deemed important for job success. Elements from the scoring guide developed by Hoffman and Holden (1990) and the Alaska State Department of Education (1988) were combined for this evaluation. Five dimensions were operationally defined:

1. Organization and clarity
2. Analytical skills
3. Spelling
4. Tone
5. Grammar and Mechanics

The five factors were scored on a six-point scale similar to the scale used for holistic scoring. Each point was operationally defined.

Selecting Benchmark Papers

Prior to the evaluation, an English teacher (SME), who had been trained by the state of California to conduct the evaluation of student's writing, met with the researcher to select the papers to be used as benchmark papers. The researcher chose 40 papers at random from the candidates' first writing papers and the SME evaluated them based on the holistic criteria. After reading the papers, the SME selected four papers from the sample to be used as benchmark papers. Because no paper in the 5 or 6 range was found, the expert produced a writing sample for both the 5th and 6th benchmarks.

Evaluator Training for Holistic Scoring

A panel of six SME's was convened to evaluate the first writing samples produced by the candidates. At the

start of the training, the test administrator discussed the specific job criteria, proper letter format, and the level of competence required for the position. This was to insure that the raters used the same criteria during the evaluation. The papers selected to be used as benchmark papers were read and scored by the raters; then the scores were compared and discussed until consensus was achieved. These papers and their scores were used as the benchmark papers against which all other writing samples were evaluated. A second set of four papers was read and scored by the raters to assure the consistency of grading. After each break, the benchmark papers were reviewed to assure consistency among the raters. This procedure was used throughout the rating session. During the actual assessment, each paper was read by two raters. If the scores differed by more than one point, both raters reread the paper and discussed it until consensus was achieved with no more than a one point spread.

Evaluator Training for Analytical Scoring

Part of the analytical scoring procedure was similar to that used for holistic scoring. Four papers were selected at random to be used as benchmark papers. The papers were read by three individuals who had participated in the

training and rating of simple writing samples. Consensus was reached quickly, possibly due to a transfer of training effect. The agreed upon score was used as the benchmark. Because of the small sample size, the papers used as the benchmark papers were also used in the analysis.

Analytical scoring standards were developed based on the research of Quellmalz (1986), elements from the scoring guide developed by Hoffman and Holden (1990), and the Alaska State Department of Education (1988). The standards provided a method of evaluating the specific elements deemed important for job success. The five dimensions were operationally defined by the researcher, and the definitions were discussed with the volunteer raters.

Table 1

Analytical Factors

Organization and clarity	Is there a good opening? Does the paper proceed in a logical fashion? Are there enough details?
Analytical Skills	Were the correct points selected to convey the intent of the paper? Were any important points left out? Were many unimportant points included?
Spelling	Were there any spelling errors?
Tone	Was the paper written in proper tone? Was the structure formal, informal, condescending or

	congenial?
Punctuation and Grammar	Did the writer follow standard conventions in writing, such as capitalization, punctuation, and paragraphing?

Analysis

A correction for direct restriction of range was preformed by using the candidates' score on the multiple choice test as a cut off for selecting individuals whose papers would be read by the evaluation committee. The following formula was used to compensate for the direct restriction of range.

$$R_u = \frac{R_{xy}\frac{S_u}{S_r}}{\sqrt{1 + R_{xy}^2 + R_{xy}^2\frac{S_u}{S_r}}}$$

R_u = correlation coefficient after correction for restriction of range (the unrestricted correlation coefficient)

R_{xy} = the correlation coefficient for the holistic scores and score of the analytical factors on the multiple choice test factors.

S_u = the standard deviation of the holistic scores on the multiple choice factors.

S_r = the standard deviation of the analytically scored papers on the multiple choice factors.

RESULTS

Pattern of Relationships

Hypothesis one, which was partially supported, stated there would be a difference between the patterns of correlations for simple and complex writing assignments when holistic and analytical methods of analysis were correlated with multiple choice test segments. Specifically there were eight predictions regarding the direction and level of increase (see Appendix A1). For example the correlation between organization and clarity, and the factors of reason and statistics was expected to be high, and the correlation between grammar and the same factors was only expected to be moderate. Appendix A2 shows the raw data correlations for both simple and complex writing samples before correction for direct restriction of range. To compute the difference in correlation, the r-value for complex writing sample was subtracted from the r-value for the simple writing sample.

Appendix A3 shows that only one of the eight predicted

correlations, the correlation between reading and analysis of data, approached the predicted change before correction for restriction of range. Appendix A4 is the data for the factors after correction for restriction of range. Appendix A5 lists the difference in correlation after correction for restriction of range and the 95% confidence interval for each correlation is listed under each correlation. This table shows that after correction for restriction of range, the correlation between reading and analysis had increased to the predicted level, and two others, data and tone, and reason and grammar approached the predicted level; three of the remaining predictions were near 0.05, and two were moderately strong in the opposite direction.

H2 stated there would be less variability among raters of holistically scored papers than among the raters of analytically scored papers, when the training time is short, approximately one to one and one half-hour long. Pre consensus scores were used to calculate the values. The concept of Euclidian distance was used to determine the variance, distance is equal to the square root of the sum of the squared differences. For analytically scored papers the variance of the five individual factors was summed and divided by five to produce the combined variance. The mean

and SD were calculated for holistic and complex/simple and analytical factors (see Table A6). This hypothesis was supported.

Discussion

The ability to write clearly has been taught in schools for hundreds of years. One of the first methods used to teach writing was the product method where students learned by imitating the works of other authors. This was followed by the process method where students spent time learning specific skills such as spelling and grammar; skills which educators deemed important to ''good writing''. Measuring an individual's mastery of a skill, such as spelling, was quick, easy and inexpensive, using indirect methods of measurement. However, educators were dissatisfied because the indirect methods did not evaluate other factors, such as the organization or clarity, tone or grammar of a paper. Educators wanted a method that was highly reliable and directly measured an individual's writing ability. This dissatisfaction lead to the development of holistic and analytical writing evaluation. These methods are slower and more costly than multiple-choice questions, or indirect measures; but they had an inter-rater reliability of about 90%, and they provide a direct measurement of a writer's

skills. The evaluation of writing samples, once the domain of education, has become vitally important to business.

For a business to be successful, it must hire individuals who are not only technically qualified, but who can also convey ideas. Business people write memos, letters, and contracts. They use writing skills to develop, for example, training manuals, write advertising copy and document the disciplinary actions taken. Therefore, the evaluation of writing samples is important in selecting employees who will be able to write appropriately. As in education, businesses also used indirect measurements like multiple-choice tests to evaluate candidates writing ability; however, they too are also somewhat dissatisfied with the results. Businesses want the scoring speed and economy of a multiple-choice test, but they also want the additional information that holistic and analytical scoring can provide.

A nomological net was used in the present study to investigate the relationship among the five factors of a multiple choice and holistic and analytical scoring of simple and complex writing samples.

Hypothesis 1 stated there would be a difference

between the patterns of correlations for simple and complex writing assignments when holistic and analytical methods of analysis were correlated with multiple choice test segments. The results of the study were mixed. One prediction was fully supported; two were partially supported, and two were moderately correlated in a negative direction. Although the magnitude and direction of change did not equal the expected results, the results may indicate that factors of an extremely well constructed and tested multiple-choice test may be able to measure factors frequently measured in analytical evaluation of writing, factors such as grammar, tone, and organization and clarity. However, even when additional factors are measured, multiple-choice tests may still miss factors that would be important to good writing.

Hypothesis 2, it was anticipated there would be less variability among the raters of holistically scored papers than among the raters of analytically scored papers when the training time is short. This hypothesis was supported. Quellmalz (1981) found that when holistic raters were given 3.5 hours of training and analytical raters were given 6 hours of training the analytical training produced slightly better results, and both groups were in the 90% agreement. However, in businesses raters seldom receive 3.5 to 6 hours

of training; more frequently they receive 1 to 1.5 hours of training. By using the appropriate method of evaluation based on training time available, businesses will be able to better evaluate candidates.

One limitation of the study may be the small sample sized used. Although the scales were corrected for restriction of range, the small sample size may have caused instability of the correlation coefficient.

A second limitation is the method of analysis. In comparing the method of analysis, both holistic and analytical scoring are inherently subjective in nature. The difference depends on the degree of specificity of the evaluation. Holistic grading looks at a paper as a single unit, however, any factor, such as tone, which is poorly or expertly executed may substantially influence the evaluation of the raters. This was revealed by the comments of the raters who were holistically rating the papers when a rater commented on the inappropriate format of the letter. The rater would comment that the project was to write a letter but the individual wrote a memo, or that the tone of the entire letter was very negative or inappropriate for the situation. In each instance, the scores given were very low. The low score may have been a reflection of a single

factor's influence on the total score.

A third limitation may be the difference between the formal education, experience and training of the individuals doing the rating in schools and in businesses. Raters in a school setting are frequently English teachers or heads of English departments. They may have a degree in English and have the experience of evaluating between 30 and 150 papers every other week. Additionally, educators are required to review writing samples from graduating high school students to determine whether or not the student meets the state's mandated minimum English competency level. In preparation, teachers or teacher trainers undergo extensive training in how to evaluate papers. In some states the training may last up to three days and is conducted by a national organization which specializes in training teachers how to evaluate papers.

In business settings the individuals who read or evaluate writing samples are seldom trained to evaluate writing. Generally, they are managers, superiors or technical analysts whose main responsibility lies in other areas. Managers or supervisors usually check or edit writing only when needed. Technical analysts, on the other hand, may have infrequent opportunities to edit the writings of

others. Because of their lack of training, their standards may be different than those of a teacher. A manager's standard may be, ''is it acceptable for the purpose,'' whereas, a teacher's standard would probably be ''is it correct.''

Implications for Future Research.

A nomological net was used to investigate the differences in the patterns of correlations between objective, multiple-choice items and subjective, writing samples for simple and complex writing samples. Although the results were only partially supported, more research needs to be done to determine the relationships between the different factors. By understanding the links between the factors, researchers may be able to establish more appropriate scoring guides and standards for the analytical evaluation of writing. This research, and past research, indicates that when training times are short the inter-rater variability of raters who holistically score writing projects may be higher than the inter-rater reliability of analytically scored papers. Therefore, a rater should only use analytical scoring when there is sufficient time available for the development of the standards and for adequately training the raters. This substantially increases

the cost of evaluating writing samples. Because of the increased time and expense required to develop and administer analytical scoring perhaps a study that investigates the utility of the two forms of scoring should be conducted.

Appendix A

Table A1

Predicted Pattern of Hypothesized Relationships

Analysis	Reason	Stats.	Reading	Data	Writing
Simple Writing Sample					
Holistic	low	low	high	low	low
Org/Clarity	low	mod.	low	high	high
Analysis	high	high	low	high	low
Spelling	low	low	high	low	high
Tone	low	low	high	low	high
Grammar	low	low	high	mod.	high
Complex Writing Sample					
Holistic	low	low	high	low	low
Org/Clarity	***high**	***high**	low	high	high
Analysis	high	high	***mod.**	high	***mod.**
Spelling	low	low	high	***mod.**	high
Tone	low	low	high	***mod.**	high
Grammar	***mod.**	***mod.**	high	mod.	high

mod. is an abbreviation for moderate

*and Bold indicates areas of predicted change

Table A2

Correlations Before Correction for Restriction of Range

Analysis	Reason	Statistics	Reading	Data	Writing
		Simple Writing Sample			
Holistic	-.043	.386	-.174	-.300	.096
Org/Clarity	-.041	-.123	.011	-.196	.268
Analysis	.010	-.017	.139	-.285	.304
Spelling	-.044	-.141	-.018	-.095	.409
Tone	.001	-.079	.074	.041	.367
Grammar	-.064	-.108	.015	-.208	.051
		Complex Writing Sample			
Holistic	-.179	-.249	.112	-.187	.007
Org/Clarity	**-.082**	**.160**	-.268	-.113	.187
Analysis	-.188	.127	**-.374**	-.345	**.254**
Spelling	-.149	-.169	-.214	**-.112**	.153
Tone	-.153	-.009	-.186	**-.144**	.200
Grammar	**-.175**	**-.064**	-.259	-.077	.213

*and Bold indicates areas of predicted change

Table A3

Difference between Complex and Simple Samples

Before Correction for Restriction of Range

Analysis	Reason	Statistic	Reading	Data	Writing
Holistic	.136	.635	-.286	-.113	.089
Org/Clarity	**.041**	**-.283**	.279	-.083	.081
Analysis	.198	-.114	**.513**	.060	**.05**
Spelling	.105	.028	.196	**.017**	.256
Tone	.154	-.070	.260	**.185**	.167
Grammar	**.111**	**-.044**	.274	-.131	-.162

*and Bold indicates areas of predicted change

Table A4

Correlations After Correction for Restriction of Range

Analysis	Reason	Stats.	Reading	Date	Writing
		Simple Writing Sample			
Holistic	-.103	.464	-.280	-.414	.146
Org/Clarity	-.100	-.174	.018	-.290	.373
Analysis	.024	-.025	.227	-.398	.412
Spelling	-.106	-.198	-.030	-.147	.507
Tone	.003	-.113	.124	.065	.471
Grammar	-.153	-.155	.026	-.306	.078
		Complex Writing Sample			
Holistic	-.396	-.331	.186	-.279	.499
Org/clarity	**.196**	**.223**	-.403	-.173	.275
Analysis	-.412	.180	-.512	-.458	**.357**
Spelling	-.339	-.235	-.336	**-.174**	.228
Tone	-.347	-.013	-.300	**-.219**	.291
Grammar	**-.389**	**-.092**	-.391	-.121	.308

*and Bold indicates areas of predicted change

Table A5

Difference between Simple and Complex

After Correction for Restriction of Range

Analysis	Reason	Stats.	Reading	Data	Writing
Holistic	.293	.795	-.466	-.135	-.353
	-.034/.546	.634/.890	-.681 /-.167	-.440/.198	-.608/-.033
Org/Clarity	**-.296**	**-.397**	.421	-.117	.098
	-.566/.031	-.639/-.084	.112/.656	-.425/.215	-.234/.372
Analysis	.436	-.205	**.739**	.060	**.055**
	.130/.666	-.496/.128	.546/.858	-.269/.377	-.274/.372
Spelling	.233	.037	.306	**.027**	.279
	-.099/.518	-.291/357	-.020/.537	-.300/.347	-.050/.553
Tone	.350	-.100	.424	**.284**	.180
	.025/.605	.411/.232	.116/.658	-.044/.557	-.153/.476
Grammar	**.236**	**-.063**	.417	-.185	-.230
	-.095/.520	-.379/.267	.107/.653	-.480/.148	-.516/.102

*Bold indicates areas of predicted change
*Small numbers are the confidence intervals for each value.

*Confidence intervals were computed using the following formula:

First r is transformed into Fisher=s Zr
The Stand error of Zr is computed: SE (Zr) = SQRT {1/N-3]
95% Confidence interval for Zr is found as
CI(Zr) = Zr +\- (z(crit.) * SE(Zr))
Using the reverse Zr to r function the values were transformed back into units of r to yield the confidence interval

Table A6

Variability Among Raters

Simple and Complex Writing

	Mean	SD
Holistic simple	.68	.67
Holistic complex	.59	.35
Analytic simple	.97	.83
Analytic complex	.74	.46

Variability Among Raters

Comparison of Means and Standard Deviations

	S_H		S_A
Simple M	.68	<	.97
Complex M	.59	<	.74
Simple SD	.67	<	.83
Complex SD	.35	<	.46

Appendix B1
Writing Project
Holistic Rating Standards

Holistic evaluation of writing assumes all components of writing are related, and the whole is more than the sum of its parts. Therefore, the work should be judged as an integrated work, separate elements should not be considered. The score for the writing sample should reflect an overall score for the assignment.

Minor errors, such as a few misspelled words or a few punctuation errors should not influence the evaluation of the writing sample.
Major errors which influence the readability, coherence, or understandability may influence the overall score if they are serious enough to detract from the overall quality of the sample.
Please use the following guideline as the criteria for evaluating the proficiency of the writing samples.

6 Superior The paper executes all of the elements completely. The main theme is clearly and logically supported. There are very few minor flaws.

5 Proficient The paper is very competent in all basic areas. There may be a few minor flaws, but they may not seriously detract from the coherence of the paper.

4 Basically competent All elements of the assignment are covered. There are a few flaws in convention which range from minor to serious. The flaws may not be so serious as to detract from the clarity of the paper.

3 Inadequate memo The paper lacks competence in one or two elements. There are several minor flaws or a few minor flaws and one major flaw.

2 Unacceptable memo The paper lacks competence in two or more major areas. There are serious flaws which dramatically influence the competence of the writing sample.

1 Incompetent memo The writing sample has only one or two of the elements required for good writing. The paper

lacks coherence and unity.

Appendix B2

Writing Project
HOLISTIC SCORING GUIDE
Simple Writing sample

Candidate=s S.S. #_____ Date: May 15, 1997

Score: Rater 1_____ Score: Rater 2_____ Total Score = $R_1+R_2/2$_____

Holistic Scoring Guide					
1	2	3	4	5	6
Incompetent letter: The writing sample has only one or two of the elements required for good writing. The paper lacks organization. Incorrect response or informal tone. Addresses a few elements of a letter.	Unacceptable letter: The paper lacks competence in two or more major areas. Serious flaws dramatically detract from the competence of the writing sample. Informal tone. Addresses several elements of a letter.	Inadequate letter: The paper lacks competence in one or two elements. There are several minor flaws and one major flaw. Slight business tone. Addresses the main elements of a letter.	Basically competent letter: All elements of the assignment are covered. There are a few flaws in convention. Written in a business tone. The paper follows correct form. All major/required elements are included.	Proficient: The paper is competent in all basic areas. There may be a few minor flaws, but they may not seriously detract from the coherence of the paper. Professional business tone. The paper follows correct form .	Superior: The paper executes all of the elements completely. The main theme is clearly and logically supported. There are very few minor flaws. The paper demonstrates competence in convention.

Appendix B3

Writing Project

ANALYTICAL SCORING GUIDE

Analytic evaluation of writing samples assumes the whole is equal to the sum of the parts, and that important parts can be identified apriori. Factors such as organization, tone, content, and mechanics are considered important to good writing; however, other factors can be measured depending on the needs of the examiner. When evaluating writing samples each factor contributes to the final total score of the product. For example, poor organization may make the content of a letter difficult to understand, and poor punctuation may completely change the intended meaning of a sentence. Therefore, correct usage of individual writing factors is very important. Analytical assessment gives an examiner the ability to evaluate the competencies of an individual in relation to each specific factor by evaluating each factor separately. In analytical assessment each factor is measured separately; therefore, the work, as a whole, should be judged based on the combined scores of the individual components measured.

Minor errors, such as a few misspelled words or a few punctuation errors should not seriously influence the evaluation of the writing sample. Major errors which influence the readability, coherence, or understandability may influence the overall score if they are serious enough to detract from the overall quality of the sample.

Please use the following guideline as the criteria for evaluating the proficiency of the writing samples.

Components of Writing

Organization.

The paper is very focused and has a readily identifiable theme. The theme of the paper may take the form of a purpose or point of view. Successful papers create a strong impression and convey the correct meaning.

Strong papers tend to:

- ∃ be clearly written so even an uninformed reader would know the writing assignment and have no trouble knowing what the writer was trying to say.
- ∃ have a strong opening. The main theme is clearly and logically supported by the facts provided.
- ∃ have sentences which are well developed and convey the meaning intended.
- ∃ use transitional words and phrases which clearly lead the reader from one point to another, they help the paper flow, and makes the paper easy to read.
- ∃ have paragraphs which reflect a sense of order, details seem to show a logical progression.
- ∃ not dwell on trivia.

Weak papers tend to:

- ∃ be unfocused or unclear. They have no identifiable central theme.
- ∃ be disjointed and ramble. They present facts without order.

- ∃ lack content.
- ∃ have week or non existent paragraphing. Transitions do not transport the reader from one thought to another.
- ∃ dwell on unimportant details.

Analysis:
The writer identifies important details and includes them in the paper. Unimportant details are omitted.

Strong papers tend to:

- ∃ focus on the importance of the information being presented. The most Important details are presented first and are supported with facts.
- ∃ omit unimportant elements.

Weak papers tend to:

- ∃ be poorly focused. They do not isolate the important details, nor do they present them once identified.
- ∃ include many unimportant details which may distract the reader from the purpose of the writing sample.

Spelling:
Correct spelling is important to convey the meaning intended. Correct spelling of easy and moderately difficulty words is considered fundamental to good English. Occasionally difficult words may be misspelled. Improper hyphenation is not considered spelling. Strong papers tend to have proper spelling. Simple and difficult words are correctly spelled. Weak papers tend to have many misspelled words.

Tone and word choice:
Tone and word choice are important because they convey the meaning of the writing sample. When tone and word choice are correct, the paper will be interesting and easily read. When tone and word choice are incorrect, the reader may be unable to detect what is intended. Raters should listen to how the words fit and flow together.

Strong papers tend to:

- ∃ use the correct tone for the assignment.
- ∃ have words chosen which convey the meaning in an interesting manner.
- ∃ use words that sound right and not forced.
- ∃ speak directly to the reader.
- ∃ be capable of evoking a mood or feeling.

Weak papers tend to:

- ∃ use an incorrect tone for the assignment.
- ∃ use words incorrectly and in ways which obscure the meaning of writing sample.
- ∃ ignore the reader.
- ∃ bore the reader.

Grammar and Mechanics:

This area covers capitalization, punctuation, usage, spelling, and sentence structure. These elements when properly used will convey the meaning intended. When improperly used meaning will be unclear. Sentences are characterized by direct, energetic sentences free of problems.

Strong papers tend to:

- ∃ have few if any punctuation errors. Words are properly capitalized
- ∃ have proper subject/verb, noun/pronoun, and pronoun/modifier agreement.
- ∃ use the correct tense for the writing sample.
- ∃ use complete sentences.
- ∃ be easily read aloud.

Weak papers tend to:

- ∃ have many punctuation errors which make it difficult to understand the meaning.
- ∃ use incorrect verb tense or make frequent agreement errors.
- ∃ have fragmented sentences.
- ∃ be very difficult to read silently or aloud.

Appendix B4

Analytical Scoring Guide

	Organization/ Clarity ____	Analysis _____	Spelling____	Tone_____	Grammar/Mechanics _____
6 Superior	Paper is well organized. Main theme is clearly and logically supported. Paragraph transitions flow, are logical, and aid understanding of the paper.	Paper shows a thorough analysis of all details. All important elements are reported and no unimportant elements are reported.	No spelling errors.	Proper tone and word usage. Words are specific and accurate.	One or two minor punctuation errors. No major errors. Words properly capitalized. Subject/verbs, noun/pronoun, and pronouns and modifier agreement is correct.
5 Proficient	Paper is well organized, few flaws. Main theme is logically supported. Paragraph transitions are clear, with only minor distractions.	Paper shows some analysis of the details. Most important elements are reported. Some unimportant items are reported.	Very few spelling errors even with difficult words.	Good tone and word choice.	A few minor punctuation errors. Minor errors in subject/verb, noun/pronoun, and pronouns/ modifier agreement.
4 Competent	Paper is organized and meaning is clear. Logical progression of ideas. Paragraph transitions clear with few major distractions.	Paper shows adequate analysis of details. Most important elements are reported.	Some common words misspelled.	Correct tone and suitable word choice.	Several minor punctuation errors and one major error. Several errors in agreement.
3 Inadequate	Paper lacks competence in one element, several minor flaws or one major flaw. Some information not understood. Poor/unclear paragraph transitions.	Paper shows poor analysis of details. Many important elements are omitted and a few unimportant items are included.	Many common words misspelled. Spelling detracts from readability.	Improper or poor tone or word choice.	Major errors in punctuation which detract from readability of sample. Sentence construction is awkward and hard to read. Several errors in agreement.
2 Unacceptable	Paper lacks competence in both organization and clarity. Paragraph transitions missing or very poor. Paper difficult to understand.	Paper shows very poor analysis. Many important elements are omitted and some unimportant items are included.	Spelling seriously detracts from readability. Many words misspelled.	Improper or poor tone and wording.	Major errors in punctuation detract from the readability of the sample. Sentence construction is awkward and hard to read. Paper is difficult to understand
1 Incompetent	Paper is unclear, rambles and does not provide required information. No paragraph transitions.	Paper shows no sign of analysis. Most important elements are omitted and many unimportant elements are included.	Spelling detracts from readability, and paper is hard to understand.	Improper or wrong tone/ wording	Many major errors in punctuation. Paper is difficult to understand. Very poor sentence construction. Many agreement problems.

REFERENCES

Alaska Writing Assessment Pilot Survey 1988-1989(1989). Alaska State Department of Education, Juneau. (ERIC Document Reproduction Service No. ED 332 223).

Applebee, A. N. (1984). Writing and reasoning, Review of Educational Research, 54, 577-596.

Calkins, L. M. (1985). I am one who writes': New approaches to children's writing. American Educator, 9, 26-29.

Campbell, J. P., McCloy, R. A., Opppler, S. H. & Sager, C. S. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.) Personnel Selection in Organizations (pp. 35-43). San Francisco: Josey Bass.

Gatewood, R. D., & Feild, H. S. (1987) Human Resource Selection (3rd ed.). Fort Worth: TX. Harcort Brace.

Godshalk, F. I., Swineford, F., & Coffman, W.E., (1966). The measurement of writing ability. New York College Entrance Examination Board

Hansen, J. B. (1992). A purpose driven assessment program. Paper presented at the symposium Multidimensional Assessment Models: Matching tools with Tasks. By the American Educational Research Association San Francisco, April (ERIC document reproduction services No. ED 344 928).

Hayes, J. R. & Flower, L. S. (1986). Writing research and the writer. American Psychologist, 41, 1106-1113.

Hillocks, G. Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. American Journal of Education, 93, 122-170.

Hoffman, J. K. & Hoffmam C. C.(1990, June). Training assessors to produce High Interrater reliability in evaluating writing samples. Paper presented at the International Personnel Management Association Assessment Council 14th Annual Conference. San Diego, CA.

Hoffman, C. C., & Holden, L. M.(1990, June). Development

and Operational use of an analytic writing evaluation instruction. Presented as part of the Symposium Alternate Strategies for Assessing Writing Skills at the 1990 International Personnel Management Association Assessment Council Conference.

Honey, R. (1990, June). Assessment of writing ability: The effectiveness of different models and formats. Paper presented at the International Personnel Management Association Assessment Council, San Diego, CA.

Kinneavy, J. (1971). A theory of discourse. New York: W.W. Norton & Company.

McCaig, R. (1982). How to (and how not to) conduct an assessment of student writing. (ERIC Document Reproduction Service No. 229 396)

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

Quellmalz, E. (1986). Writing skills assessment. In R. Berk (Ed.) Performance Assessment Methods and Application. Baltimore: Johns Hopkins Press.

Quellmalz, E. S.(1986). Towards successful large-scale writing assessment: Where are we now? Where do we go from here? . Educational Measurement Issues and Practices, 3(1), 29-32,35.

Quellmalz, E. S. & Burry, J. (1983). Analytic scaler for assessing students' expository and narrative writing skills. CSE Resource paper No. 5 (ERIC document reproduction services No. ED277006O).

Quellmalz, E. S., Capell, F. J. & Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 19 (4), 241-258.

Quellmalz, E. S. (1981). Report on Conejo Valley's fourth-grade writing assessment. Los Angeles: Center for

the Study of Evaluation, University of California.

Reutzel, D. R. & Hollingsworth, P. M. (1988). Whole language and the practioner. Academic Therapy, 23, 405-415.

Spandel, V. (1981). Class Room Applications of Writing Assessment: A Teachers Handbook. Northwest Regional Education Lab., Portland Oregon. (ERIC Document reproduction services No. ED 214 996).

Veal, L. R., Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 5, 37-45.

White, E. (1985). Teaching and assessing writing. San Francisco: Josey-Bass.