DOCUMENT RESUME

ED 353 327 TM 019 391

AUTHOR McLean, James E., Comp.
TITLE The Utility, Reliability, and Validity of Holistic Scoring for Writing Assessment Samples. Symposium Presented at the Annual Meeting of the Mid-South Educational Research Association (Knoxville, Tennessee, November 11-13, 1992).
PUB DATE Nov 92
NOTE 39p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Collected Works - General (020)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Conference Papers; Cues; Evaluation Methods; Grade 10; High Schools; High School Students; *Holistic Approach; Pretests Posttests; *Reliability; Research Methodology; *Scoring; Skill Development; *Validity; Writing Achievement; *Writing Evaluation
IDENTIFIERS Alabama; Performance Based Evaluation; *Writing Samples

ABSTRACT

Four papers describe a study of the use of holistic writing assessment procedures in a pretest and posttest manner to determine improvement of 10th graders' writing skills. "Problem and Context" (James E. McLean) briefly describes the project and introduces the other three papers. "Holistic Scoring Procedures for Scoring Writing Samples" (Sybil A. Hobson and D. Joyce Steele) describes use of the holistic procedure for scoring the pre- and post-writing samples collected by the project. The need for a homogeneous group of scorers with similar backgrounds who are open to using prescribed scoring methods is discussed. "Reliability of the Holistic Scoring Procedures for the Blue Ribbon Committee Writing Assessment" (Margaret L. Glowacki) assesses the reliability of holistic scoring for two groups of readers (retired teachers trained in holistic scoring specifically for the project and experienced holistic readers). Controlled essay reading, scoring criteria guide, sample papers, checks on reading in progress, multiple independent scoring, and evaluation and record keeping were used. "Validity of the Writing Assessment" (Marcia R. O'Neal) addresses the validity of the writing assessment, using data for 484 10th grade students in Alabama participating in the fall 1991 pretest and spring 1992 posttest. This study provides some limited evidence for the validity of the writing assessment as used in the Blue Ribbon Committee Project. Writing outcomes measure characteristics similar but not identical to those measured by objective methods. Data show stronger relationships with theoretically similar measures. (RLC)

SYMPOSIUM

The Utility, Reliability, and Validity of Holistic Scoring for Writing Assessment Samples

James E. McLean, Organizer

Problem and Content:
James E. McLean*, The University of Alabama

Holistic Scoring Procedures for Scoring Writing Samples:
Sybil A. Hobson* and D. Joyce Steele*, The University of Alabama

Reliability of the Holistic Scoring Procedures:
Margaret L. Glowacki*, The University of Alabama

Validity of the Writing Assessment:
Marcia R. O'Neal*, The University of Alabama

A Symposium Presented at the Annual Meeting
of the
Mid-South Educational Research Association
Knoxville, Tennessee
November 11-13, 1992

* *The authors can be reached at The University of Alabama, P. O. Box 870231, Tuscaloosa, AL 35487-0231, FAX 205/348-6873*

PROBLEM AND CONTEXT

James E. McLean

The University of Alabama

PROBLEM AND CONTEXT

Assessing students' writing has always been a difficult task. It is even more difficult to assess changes or improvement of students' writing. The purpose of this study was to determine the utility, reliability, and validity of applying holistic writing assessment procedures in a pre- and post-manner to tenth grade students to determine improvement of their writing skills.

A group of civic, business, and industrial leaders in a community formed a committee to reward good teaching and improve education. One of their activities was to identify the "best" teacher of tenth grade writing in an area that included three school systems. The committee stipulated that 50% of the determination be based on direct assessments of student performance. After providing appropriate caveats of using pre- and post-measures of student learning to assess teaching performance, a study was undertaken to evaluate its validity. Only volunteer teachers participated in the study. Teachers were given full information regarding the data that would be collected and how it was to be used. Each teacher signed an agreement form before the study began.

Equated writing prompts were presented to the students of the volunteer teachers at the beginning of the year and again at the end of the year. The other presentations in this symposium provide the procedures, reliability, and validity assessments of this effort.

The first paper by Hobson and Steele describes the use of the holistic scoring procedure for scoring the pre- and post-writing samples collected during the project. All identification except code numbers was removed from the papers. Scorers did not know what teacher, school, or system the paper came from nor if they were pre or post. Nonmatched papers were used for training. Both local retired teachers and experienced scorers were used to score the papers. All scorers went through a rigorous training procedure including check points to acquire inter-rater agreement.

The paper by Glowacki assesses the reliability of holistic scoring for two groups of readers.

The first group of readers consisted of retired teachers with varied backgrounds that were trained in holistic scoring specifically for this project. The second group consisted of experienced holistic readers. Pearson correlation coefficients were computed to examine reader reliability. Results indicated moderate rater agreement for the experienced readers, and moderately weak rater agreement for the retired teachers. Additional results reported for both groups included percentages of readers assigning the same scores, contiguous scores, and discrepant scores, and comparison of the means of the first and second readings.

The paper by O'Neal addresses the validity of the writing assessment. The availability of Grade 9 data from the Alabama Basic Competency Test (BCT) administered in fall 1991 to many of the students participating in the writing assessment made possible an examination of the validity of the direct writing assessment. Results of the holistic scoring by two readers were combined to arrive at mean pretest and posttest scores for each of the participants for whom two reader scores were available for both the pretest and posttest. These means were then correlated with BCT results. Correlation of the BCT with the pretest mean was .48. The BCT-posttest correlation was .36.

The full text of each paper follows.

Holistic Scoring Procedures for

Scoring Writing Samples

Sybil A. Hobson and D. Joyce Steele

The University of Alabama, Tuscaloosa, Alabama

Holistic Scoring Procedures for Scoring Writing Samples

Introduction

This paper describes the use of the holistic scoring procedure for scoring the pre- and post-writing samples collected during the project. As was stated previously, the generally equated writing prompts were presented to the students of the volunteer teachers at the beginning of the year and again at the end of the year. Both pre- and post-writing samples were separated from all identification except code numbers. Scorers did not know which teacher, school, or school system the writing sample came from nor if they were pre- or post-writing samples. Writing samples without a match from either the pre- or post-writing administration were used for training purposes. Both local retired teachers and experienced scorers were utilized to score the writing samples. Scorers were trained to use holistic scoring.

The holistic scoring method is a procedure for assessing writing in which a reader judges a writing sample for its overall effectiveness. The scorer must read quickly and make a judgement about the total effectiveness of the writing sample with factors such as organization, spelling, and grammar considered to be of equal importance. The readers receive instruction to read the writing sample quickly and to assign a score based on their total impression of the general quality. The readers are instructed not to reread the writing sample.

During the training, writing samples were used to illustrate the hierarchy of score categories based on the predetermined criterion for scoring. The criterion for each score category was adapted from the General Educational Development (GED) Essay Scoring Guide copyrighted in 1985 by the GED Testing Service. The training was conducted by Mrs. Sybil A. Hobson, from Test Service at The University of Alabama in Tuscaloosa.

Retired teachers were asked by the superintendent representative of the school systems to volunteer their time to score the writing samples. From the teachers' verbal description, it appeared

that the volunteer retired teachers received little or no information about what they were volunteering to do. The school system representatives had difficulty obtaining the number of volunteers needed to score the writing samples in a timely manner. Also, some of the volunteers did not have experience in teaching language arts and writing and some had worked in other positions as counselors for several years since teaching English.

The volunteer retired teachers were a very conscientious group; however, several factors beyond their control or beyond the control of the Assessment Team from The University of Alabama worked adversely toward the successful completion of the scoring task. After the morning session, and after deliberation during lunch by the Assessment Team, it was decided that due to the small number of volunteer retired teachers and their varied backgrounds, it would be more cost effective and more efficient to use trained scorers from a pool of Graduate Teaching Assistants from the English Department at The University of Alabama.

Scoring Procedures

The readers were trained to use a six point scoring guide to assign scores to the papers. The characteristics a paper must have to qualify for each of the score points was discussed at length. The scores ranged from one to six, with six as the most desirable score. The typical characteristics of a six paper are vivid and precise writing which offers sophisticated ideas with particularly effective support for the ideas. The one paper, at the opposite end of the score scale, lacks purpose or development and ideas are difficult or impossible to understand. A score of '0' was reserved for papers that were blank or written on a topic other than the one assigned.

Objectivity issues were presented to assist readers in overcoming potential biases which could influence a reader's perception of a paper. Some factors which could cause reader bias are characteristics of the paper itself such as: neatness, handwriting, skipping lines, unusual margins and length of the paper. Personal reactions to the writing such as preferences or prejudices in style,

reactions to the content and expectations for performance that is inappropriate for the population being tested could also cause reader bias.

Readers were instructed to read a paper and record the score immediately. They should not reread the paper to justify the score in terms of specific errors. Papers should not be compared one to another. Readers were cautioned to avoid making assumptions and score only what is written. Papers without a match from either the pre- or post-writing administration were read by previously trained readers to identify typical papers for each of the six score points. From these papers, a paper which exemplified all the characteristics of a particular score was selected for each of the six score points and rationale was developed to explain why each paper was a specific score. These 'anchor' papers were used as typical papers in training the readers to recognize the characteristics specific to each score point. The readers read and scored each paper in the anchor set. The trainer discussed each paper in the set and gave the rationale for the score for each paper.

Three sets of training papers were prepared and used in this same manner to assist the readers in internalizing a concept of the score points. The readers would read and score a set of eight papers. For each paper, the trainer would ask all readers to show a score card containing the score the reader had given the paper. Various readers were then asked to discuss why they had given the paper that particular score. Explanations were required to be given using the language of the scoring guide. The readers were then given the correct score and the rationale for the score. This was repeated with as many training sets as necessary until the trainer was confident the readers were scoring on target.

The use of flash cards for the readers to indicate their score was useful in that the trainer could glance around the room and know immediately if the readers were scoring on target or if a reader was consistently scoring too high or too low. If this were the case, the trainer could give more detailed information about the scoring procedure the characteristics required for each score

point. Training concluded when the readers were scoring a minimum of 50% on target and the remaining 50% were not more than one point off target.

Papers were placed in packets and a scoring sheet which contained the code number of each paper in the packet was attached. The papers for each teacher were distributed across as many packets as possible to guard against possible bias. Each reader received a packet of papers, read each paper and recorded the score beside the paper's code number on the score sheet. The code number for the reader was recorded on the score sheet also. The reader returned the score sheet and packet after all papers in the packet were scored and received another packet to score. A member of the Assessment Team removed the first reader score sheet from the returned packet, recorded the first reader number on the second reader score sheet and attached the sheet to the packet. This procedure was repeated until all papers were scored by a first reader. The papers were then scored by a second reader. A reader could not be both first and second reader for a packet of papers. Recording the first reader's code number on the second score sheet prevented this from occurring.

Readers were not permitted to judge a paper as "off-topic". Papers that readers considered to be "off-topic" were given to the trainer who decided if they were indeed "off-topic". Papers that had been assigned scores that were more than one point discrepant were also given to the trainer who read the paper as the third reader and assigned the paper a score.

Summary

One of the major findings during this procedure was the need for a homogeneous group of scorers with similar backgrounds who are open to using a prescribed method of scoring. The use of paid Graduate Teaching Assistants from the English Department of The University of Alabama worked very well. The Graduate Teaching Assistants are accustomed to grading papers and are often eager to make extra money. As graduate students, they are receptive to receiving instruction

and using the instructions to score papers. In the long run, the use of Graduate Teaching Assistants was more cost effective in terms of training time, scoring time, and money. Although the retired teachers were volunteers, travel costs and expenses for the Assessment Team exceeded the cost of hiring the graduate teaching assistants.

References

American Council on Education (1985). Adding an essay to the GED Writing Skills Test: Reliability and Validity Issues. Washington, D.C.: Authors.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., and Rock, D. A. (1987). Assessing Writing Skill. New York: College Entrance Examination Board.

Reliability of the Holistic Scoring Procedures for the Blue Ribbon Committee Writing Assessment

Margaret L. Glowacki

The University of Alabama

Reliability of the Holistic Scoring Procedures

for the Blue Ribbon Committee Writing Assessment

This aspect of the study investigated the reliability of the holistic scoring for the Blue Ribbon Committee Writing Assessment project. Holistic scoring involves having two or more readers evaluate a writing passage for its overall effectiveness, as a whole, rather than by considering its individual features such as word use, grammar, punctuation, organization, and style in isolation. Its purpose is not to correct, edit, or diagnose the weaknesses of a writing sample, but to form a general impression of the writing (Charney, 1984). This is in contrast to analytic scoring which involves the identification of characteristics viewed as important to good writing and judging papers by the number of these characteristics the papers contain. Each characteristic is scored separately and the scores are totalled to provide a rating (Huot, 1990).

Another method of assessing writing, which is similar to holistic scoring, is primary trait scoring. According to White (1985) there is little difference between primary trait scoring and holistic scoring except that the criteria used in primary trait scoring are defined with greater precision and exclusiveness than in holistic scoring. According to Kilpatrick (1986), the premise of primary trait scoring is that "any piece of writing is directed toward a particular audience for a particular purpose" (p. 29). Primary trait scoring was developed by committees at the National Assessment of Educational Progress (White, 1985).

Reliability refers to the consistency with which an examinee obtains the same score on the same measure at different administration times or the ability of independent observers to agree. Reliability is generally reported as a decimal fraction, stated in terms of a positive correlation coefficient ranging from .00 to 1.00. A correlation coefficient of +1.00 means there is a perfect linear agreement between two sets of observations. A correlation coefficient of .00 means there is no linear relationship. Although there are many types of reliability, inter-rater or inter-reader

reliability is the type used in holistic scoring. Inter-reader reliability is the degree of agreement between two or more readers on a writing sample and is of great importance to the reliability of holistic assessments.

According to White (1985), no measure can be completely reliable due to changes in performance by individual examinees that have nothing to do with the measure. "Holistic scoring is able to achieve acceptably high reliability by adding a series of constraints to the economically efficient practice of general impression scoring" (p.23). White maintained that when examining the reliability of holistic scoring, many of the objections to holistic scoring's reliability are actually objections to the unreliability of general impression scoring. General impression scoring was the early, less reliable scoring method used by Diederich (1974) in which there were no guides or controls. During the 1950s and 1960s, it was discovered that if readers were trained in the holistic scoring methods, reliable results could be produced and researchers and individuals involved in testing began to use holistic scoring. Educational Testing Service (ETS) has been largely responsible for the development of techniques that have led to the present reliability of holistic scoring (Humes, 1980; White, 1985) and routinely uses holistic scoring to assess writing samples (Charney, 1984). Holistic scoring also is the most commonly used method for writing assessment in the elementary schools (Charney, 1984; Freedman & Calfee, 1983). According to White (1985), by following a set of procedures designed to reduce unnecessary variability in scoring, it is possible to achieve acceptably high reliability. A carefully developed writing prompt is also a necessity. The six steps as discussed by White are:

1. Controlled Essay Reading - All papers are read at the same time and in the same place.
2. Scoring Criteria Guide - The scoring guide provides descriptions for papers at different points on the scoring scale.

3. Sample Papers - They also are called anchor papers and are examples of papers at each point on the scoring scale. These papers are unmarked and given to readers to score during training. Their purpose is to help the readers internalize the scoring scale.

4. Checks on the Reading in Progress - Experienced readers act as leaders and check scores to determine that all readers continue to score at the same level.

5. Multiple Independent Scoring - Two readers read each essay independently. If there is a difference larger than one point between the two readers, a third reader reads the paper and the scores are averaged.

6. Evaluation and Record Keeping - Records of scoring by the readers should be kept and the most reliable scorers identified for future scoring.

As discussed in the section of this paper regarding procedures, these steps were carefully followed during the reading of the test papers. Since, there was not sufficient time for the readers in Gadsden to read both the pretests and posttests, it was decided that the pretests would be read again by the University of Alabama (UA) readers so that both the pretests and posttests could be scored by the same readers. Since the UA readers are required to have certain qualifications (American Council on Education, 1991), this also provided an opportunity to examine how well the training worked for readers who might not possess these qualifications or similar backgrounds by comparing the scores of the Gadsden readers with those of the UA readers.

Cooper (1977) discussed the unreliability of essay ratings, but stated that reliability "can be improved to an acceptable level *when raters from similar backgrounds are carefully trained*" (p. 18). He stated that

> when raters are from similar backgrounds and when they are trained with a holistic scoring guide--either one they borrow or devise for themselves on the spot--they can achieve nearly perfect agreement in choosing the better of a pair of essays; and they can achieve scoring reliabilities in the high eighties and low nineties on their summed scores from multiple pieces of a student's writing. (p. 19)

Qualifications for the UA readers include that they are required to be graduate teaching assistants in English; have a bachelor's degree, preferably in English; be able to write effectively; have taught secondary or post-secondary English language arts for a minimum of two years; be willing to accept the established scoring standards; and demonstrate an ability to work well in groups. Stalnaker (1934) illustrated that inter-reader reliability could be improved with training from a range of .30 to .75 to a range of .73 to .98.

Cooper (1977) reported a study conducted by Lee Odell in which Odell obtained 80%, 100%, and 100% agreements between two raters who chose the better essay from each of thirty pairs of pre- and posttest essays for three types of writing. The raters included experienced teachers and graduate students in English education who received an average of one hour of training time for each type of writing.

Charney (1984) maintained that individuals who use holistic scoring assume that if the following conditions are met, the writing assessments will be reliable and valid:

> [I]f the design of the training and rating sessions take the factors necessary for reliability into account;
> if the readers are qualified, and come from similar backgrounds;
> if the readers are "calibrated," that is, trained to conform to agreed upon criteria of judgement;
> if the criteria, which either are supplied to the readers in the form of a rating guide, or are decided upon by the readers as a group, are appropriate; and
> if readers work quickly, usually under supervision. (p. 69)

According to Swartz, Patience, and Whitney (1985), there are two aspects of writing samples that affect the reliability of a large-scale writing assessment: reading reliability and score reliability. Reading reliability refers to the consistency with which different readers give comparable scores to the same paper. Score reliability refers to the similarity of scores received by an examinee on two

similar, equally difficult topics. In this paper, only reading reliability will be examined. The objective of the writing assessment program was to improve the students' writing skills, so differences were expected between the pretest and posttest scores and there was no opportunity to administer two writings for either the pretest or the posttest.

White (1985) expressed uncertainty as to the reliability of controlled scoring of essays because of a lack of agreement about ways of measuring reliability of writing samples, and the many different ways available to compute reading reliability. Breland, Camp, Jones, Morris, and Rock (1987) discussed several sources of error in their study including examinees, having six different topics, the use of several modes of discourse, and the readers who scored the essays.

The most common method for estimating reading reliability is a correlation between the scores assigned by different readers to the same writing sample. Breland, et al. (1987) maintained that this type of estimate generally is inflated because the only source of error recognized is the readers. A study by French (1962) as described by Breland, et al. (1987) was conducted at ETS in 1961 to examine the scoring of essays by readers representing several professional fields. Three hundred essays written by college freshmen were scored by 53 readers using a 9-point scale. No essay received fewer than five of nine possible ratings, seven different ratings were received by 23 percent of the essays, eight different ratings were received by 37 percent of the essays, and all possible ratings were received by 34 percent of the essays. Inter-reader reliability was .31. The conclusion of this study was that the essay score depended to a large degree on who did the scoring. The problem of reliability was emphasized by recognizing that readers represent only one source of error.

Several studies have been conducted to examine the reliability of holistic scoring. Holistic scoring using a 7-point scale was used in the writing assessment program conducted in Grosse Pointe, Michigan public school systems each spring. McCaig (1984), as cited by Kilpatrick (1986),

discussed the program, which had been in place since 1975 in grades one through ten. The average reliability coefficient between readers and a final rating was .75.

Swartz and Whitney (1985) conducted a study to examine the relationship between scores on the Writing Skills test of the Test of General Educational Development (GED tests), which is an indirect measure, and scores on holistically scored essays. During the fall of 1983, while equating new forms of the GED tests, three essay topics also were administered. The sample was national and included 170 high schools. The holistic scoring method was selected for scoring because holistic scoring is quick and relatively simple, and it offered the most accessible method of achieving high degrees of inter-reader reliability. The results indicated that a high degree of inter-reader agreement can be attained with holistically scored writing samples. Correlations between first and second readers ranged from .69 to .76.

Godshalk, Swineford, and Coffman (1966) used a three-point holistic scale to have 25 readers score 646 essays. The result was an estimated inter-reader reliability of .921, but evidence existed that some readers assigned a score of 2 to essays whose quality they were unsure of. Godshalk, et al. (1966) then conducted a field study in which 146 readers re-read 533 of the essays from the original study. Both a three-point and four-point holistic scale were used. The four-point scale was found to be more highly favored and reliable.

Several studies have been conducted to compare analytic and holistic scoring. The California State Department of Education developed sample writing exercises and scoring guides for assessment of elementary students' writing using three discourse modes: expressive, explanatory, and persuasive. A 4-point holistic scale and an analytic scoring guide were developed. Correlational coefficients between the holistic and analytic scores for expressive, explanatory, and persuasive writing samples were .75, .76, and .76, respectively (Prater & Padia, 1980).

A study by Bauer (1981) compared analytic, holistic, and primary trait scoring methods. The results reported included intra- and inter-reader reliabilities. An analysis of variance based on Snedecor's intra-class correlation formula was used to calculate the inter-rater reliabilities of the three scoring methods. The analytic method had the highest reliability (.954), holistic was next highest (.928), and primary trait was lowest (.838). Using a z-test with a .01 level of significance to compare the reliabilities of the three methods, no significant difference was found between the analytic and holistic scoring methods.

White and Polin (1986) compared three types of direct writing assessment scoring scales: holistic, development and focus, and correctness and efficiency. The inter-reader correlation was highest for the holistic scale (.75). The inter-reader correlation for the correctness and efficiency scale was .67, and .66 for the development and focus scale.

Other studies have compared direct and indirect methods of writing assessment and have found them to correlate highly. Swartz and Whitney (1985) found that holistic scoring can be used to score writing samples with a high degree of inter-reader agreement. In their study, readers agreed or gave a contiguous score in 93% of cases. Results also indicated a substantial relationship between writing samples and multiple choice items, although the two measures appear to be measuring different skills.

Veal and Hudson (1983) conducted a study comparing potential validity, reliability, and costs of several direct and indirect writing measures. Direct measures included holistic, analytic, primary trait, and mechanics count scoring. Indirect measures included language arts items from several objective tests. Approximately 100 tenth grade students from 24 high schools were randomly selected. Holistic scores were determined to be most widely representative scores. The inter-reader reliability for the holistic methods ranged from .69 to .76. Also, the scores for the analytic method had the highest correlation with the holistic methods for the direct measures (.64) while the scores

of the Iowa Test of Basic Skills had the highest correlation with holistic scoring for the indirect measures (.70).

A reliability coefficient of .76 between a multiple choice test, Standard Test of Written English, and an essay test was reported by Breland and Gaynor (1979). Vacc (1989) reported a significant correlation between holistic and analytic scoring by four teachers. Cooper (1977) maintained that "since holistic evaluation can be as reliable as multiple-choice testing and since it is always more valid, it should have the first claim on our attention when we need scores to rank-order a group of students" (p. 4).

The Pearson Product Moment Correlation procedure was used in this study to examine inter-reader reliability. Many studies examine both inter-reader and intra-reader reliabilities (one reader reading two papers by the same individual). One of the limitations of this study was that only one sample was collected for the pretest and one sample for the posttest topics with treatment occurring between the tests, so intra-reader reliability could not be examined. To examine intra-reader reliability, two writing samples by the same individual without treatment would have to be collected and this was not possible. Results of the current study provided an inter-reader reliability of .35 for the pretests scored by the readers in Gadsden, and reliabilities of .52 and .57, for the pretests and posttests, respectively, scored by the UA readers.

Of the writing samples included in the study (not scored as off topic by either reader), the number of papers that required a third reading for the Gadsden readers, the UA pretest readers, and the UA posttest readers were 128, 31, and 41, respectively. Other factors examined included the percentages of readers assigning the same score, contiguous scores, and discrepant scores (i.e. scores differing by more than one point). These are presented in Table 1. Although the sample originally consisted of 484 subjects, only those subjects who took both the pretest and posttest, and who were not off topic for either the pretest or posttest were used in all calculations throughout this

paper. Descriptive statistics are presented in Table 2. The frequency distributions of scores and the percentage of times each score was assigned are presented in Table 3. The scores given by readers 1 and 2 for each writing sample were averaged to produce a single score for each sample in the two pretest scoring sessions and the posttest scoring session. The averaged scores for the pretest writing samples read by the Gadsden readers and the UA readers then were compared to examine the percentages of same scores, contiguous scores, and discrepant scores for the two groups of readers. These are presented in Table 4.

Table 1

Percentage of Same Scores, Contiguous Scores, and Discrepant Scores between Reader 1 and Reader 2 for Each Scoring Session

Readers	N	% Same Scores	% Contiguous Scores	% Discrepant Scores	Inter-reader Reliability
Gadsden Pretest	363	26%	39%	36%	.35
UA Pretest	366	39%	52%	8%	.52
UA Posttest	366	33%	56%	11%	.57

Table 2

Means and Standard Deviations

		Mean		Standard Deviation	
Group	N	Reader 1	Reader 2	Reader 1	Reader 2
Gadsden Pretest	363	3.76	3.59	1.34	1.37
UA Pretest	366	3.70	3.09	.76	.71
UA Posttest	366	3.74	3.03	.80	.74

Table 3

Frequency Distribution of Scores

	Pretest - Gadsden Readers		Pretest - UA Readers		Posttest - UA Readers	
Score	Absolute Frequency	Percent of Total Papers	Absolute Frequency	Percent of Total Papers	Absolute Frequency	Percent of Total Papers
1	5	1.37%	0	0%	1	0.27%
1.5	12	3.30%	2	0.55%	4	1.09%
2	21	5.79%	6	1.64%	9	2.46%
2.5	37	10.19%	40	10.93%	44	12.02%
3	49	13.50%	110	30.05%	92	25.14%
3.5	64	17.63%	118	32.24%	121	33.06%
4	71	19.56%	49	13.39%	56	15.30%
4.5	42	11.57%	33	9.02%	33	9.02%
5	35	9.64%	7	1.91%	3	0.82%
5.5	16	4.41%	1	0.27%	3	0.82%
6	11	3.03%	0	0%	0	0%
		100.00%		100.00%		100.00%

Table 4

Percentage of Same Scores, Contiguous Scores, and Discrepant Scores Between the Gadsden Readers and the UA Readers on the Pretest

N	% Same Scores	% Contiguous Scores	% Discrepant Scores
332	36%	41%	23%

The difference in reliability coefficients between the Gadsden readers and the UA readers may have been due to several factors. The Gadsden readers consisted of retired teachers with varied backgrounds and various teaching specialties. The subjects taught included psychology, world

geography, composition, English literature, grammar, language arts, government, economics, world history, science, social studies, reading, and writing. Positions held included high school counselor, principal, and assistant principal. Years of experience teaching writing classes ranged from 0 to 27. All UA readers had similar backgrounds because they are required to be graduate teaching assistants in English; have a bachelor's degree, preferably in English; be able to write effectively; have taught secondary or post-secondary English language arts for a minimum of two years; be willing to accept the established scoring standards; and demonstrate an ability to work well in groups. This also was the first time the Gadsden readers had any experience with writing assessment and the first time they had received training. In contrast, the UA readers have from nine months to one and one-half years of experience reading writing samples on many different topics, with training for each new set of topics.

As Kilpatrick (1986) pointed out, one limitation of holistic scoring is the possibility of reader inconsistency. Readers may come to the training having different ideas of evaluation and must accept and learn to apply the scoring guides provided for the writing assessment accurately and consistently. Sometimes, even readers who have achieved high levels of agreement will show less agreement when they score large numbers of papers. It is believed that following the training steps delineated by White (1983) will aid in reducing this inconsistency. Follman and Anderson (1967), as quoted by Cooper (1977), concluded

> [i]t may now be suggested that the unreliability usually obtained in the evaluation of essays occurs primarily because raters are to a considerable degree heterogeneous in academic background and have had different experiential backgrounds which are likely to produce different attitudes and values which operate significantly in their evaluations of essays. The function of a theme evaluation procedure, then, becomes that of a sensitizer or organizer of the rater's perception and gives direction to the attitudes and values; in other words, it points out what he should look for and guides his judgment. (p. 19)

Coffman (1971) concluded that "in general, when made aware of discrepancies, teachers tend to move their own ratings in the direction of the average ratings of the group. Over a period

of time, the ratings of the staff as a group tend to become more reliable" (p. 36). It is possible that with an opportunity for more training, the Gadsden readers may have increased their inter-reader reliabilities.

References

Bauer, B. A. (1981). A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability. (ERIC Document Reproduction Service No. ED 216 357)

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill. New York: College Board Publications.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical review. Research in the Teaching of English, 18(1), 65-81.

Coffman, W. E. (1971). On the reliability of ratings of essay examinations in English. Research in the Teaching of English. 5, 24-36.

Cooper, C. R., & Odell, L. (1977). Evaluating writing: Describing, measuring, judging. National council of Teacher of English.

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

Follman, J. C., & Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English 1, 190-200.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. W. Walmsley (Eds.). Research on writing: Principles and methods (pp. 75-98). New York: Longman.

French, J. (1962). Schools of thought in judging excellence of English themes. Proceedings of the 1961 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

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

Humes, A. (1980). A method for evaluating writing samples. Southwest Regional Laboratory Technical Note. (ERIC Document Reproduction Service No. ED 193 631)

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research, 60(2), 237-263.

Kilpatrick, N. G. (1986). A study comparing holistic ratings of writing samples and spelling and language standardized test scores of sixth grade students. (Doctoral dissertation, The University of Alabama, 1986). Dissertation Abstracts International, 48, 03A.

McCaig, R. A. (1984). A model for the evaluation of student writing: A handbook (2nd ed.). Grosse Pointe, MI: Grosse Pointe Public School System.

Prater, D., & Padia, W. (1980, February). Developing parallel holistic and analytic scoring guides for assessing elementary writing samples. Paper presented at the meeting of the Southwest Educational Research Association in San Antonio, Texas, for the California State Department of Education. (ERIC Document Reproduction Service No. ED 194 530)

Stalnaker, J. M. (1934). The construction and results of a twelve-hour test in English composition. School and Society, 39, 218-24.

Swartz, R., Patience, W., & Whitney, D. R. (1985, July). Adding an essay to the GED writing skills test: Reliability and validity issues. GED Testing Service, Research Studies, Number 7, American Council on Education.

Swartz, R., & Whitney, D. R. (1985). The relationship between scores on the GED Writing Skills test and on direct measures of writing. Washington, D.C.: American Council on Education.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. The Elementary School Journal, 90, 1, 87-95.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. Research in the Teaching of English, 17(3), 290-296.

White, E. M. (1985). Teaching and assessing writing. San Francisco, CA: Jossey-Bass, Inc.

White, E. M., & Polin, L. G. (1986). Research in effective teaching of writing: Volumes I and II. Los Angeles: California State University Foundation. (ERIC Document Reproduction Service No. ED 275 007)

Validity of the Writing Assessment

Marcia R. O'Neal

The University of Alabama

Validity of the Writing Assessment

Introduction

As with any measure, a number of validity issues must be considered for holistically scored writing assessments. As indicated in the Standards for Educational and Psychological Testing (1985), "Validity is the most important consideration in test evaluation" (p. 9). McKendy (1992), among others, pointed out that maintaining the validity of an instrument involves careful attention at the test development, administration, and scoring stages. Assessing the validity of holistically scored direct writing assessments presents its own unique set of challenges. These will be addressed in the context of the major types of validity, which will be discussed following a description of the data.

Data

A total of 484 tenth grade students from the Etowah County, Attalla City, and Gadsden City school systems in Alabama participated in both the fall 1991 pretest and the spring 1992 posttest. The holistic scoring of the essays for all students participating in both writing assessments was described in a paper presented earlier in this symposium. A score of zero (off topic or unreadable) by either of two readers or a difference of more than one point by the two readers resulted in the paper being read by a third reader. In order for the essay to receive a valid average, at least two readers had to rate the essay higher than zero. The average was based on the scores of either two or three readers, using only those scores higher than zero.

The test results for the spring 1991 administration of the Grade 9 Alabama Basic Competency Tests (BCT) were available as a criterion measure against which to compare direct writing assessment results. The BCT is a minimum competency test given statewide to students in Grades 3, 6, and 9 in the areas of reading, mathematics, and language (Teague, 1989). Although

total subject area scores typically are not reported, they can be obtained from the sum of items correct on each competency of the subtest. The language portion of the test contains competencies in areas typically found on tests traditionally considered indirect, objective assessments of writing. These competencies include such skills as spelling, punctuation, subject-verb agreement, pronoun-antecedent agreement, finding errors in paragraphs, and recognizing complete sentences.

The spring 1991 BCT results were obtained for students in the three school systems participating in the Blue Ribbon Committee Project. A total of 1,262 students in these systems took one or more sections of the Grade 9 BCT. These records were matched to the records containing writing assessment holistic scoring results using student name. A total of 423 matches were found.

It was possible that for any given student among the 423 matches, the student may not have taken one or more of the BCT subject areas, or the student may not have received a valid mean holistic score for the direct writing assessment pretest or posttest (two or more readers giving the essay a score of zero). Therefore, it was possible for a student record to contain from one to five scores, including holistic pretest mean, holistic posttest mean, BCT reading score, BCT mathematics score, and BCT language score. Analyses for the validity study were conducted in two ways. First, each analysis was completed using all pairs available for that analysis (pairwise analyses). The Ns for these pairwise analyses ranged from 337 to 414. Then, the same analyses were completed using only the 331 records that had all five scores present (common N analyses). The results of both sets of analyses are reported in this study. Means and standard deviations are provided in Table 1 for all available scores and for the set of scores on which common N analyses were performed.

Table 1

Means and Standard Deviations for Holistic Pretest and Posttest Means and BCT Results

	All Available Scores			Common N Analyses (N=331)	
	N	$\bar{X}$	SD	$\bar{X}$	SD
Pretest Mean	340	3.37	.67	3.37	.67
Posttest Mean	416	3.33	.69	3.38	.70
BCT Reading	420	69.58	8.32	70.88	7.22
BCT Mathematics	420	75.06	15.21	76.82	14.26
BCT Language	418	77.84	12.71	79.52	11.94

Content Validity

The content validity of an assessment instrument should be built in during the development process (Standards, 1985). In the case of direct writing assessment, Miller and Crocker (1990) suggested criteria for good prompts based on their summary of suggestions from other researchers:

1. Prompts must be thought provoking and allow some latitude for expression.

2. Prompts must be specific enough that all examinees are responding to a common theme or a common core of content.

3. Prompts must provide structure so that examinees know what is expected of their responses. For example, prompts should include specific instructions such as "address both sides of the issue" or "give examples."

4. In developing prompts, many options exist in terms of the intended audience and mode of discourse. . . . Thus, experts should be in agreement about the mode of discourse and type of audience that are suggested to the examinee by the prompt.

5. The content of prompts should be within the realm of general experience of all examinees and should not provide an advantage to any particular subgroup, whether by race, gender, or culture. . . .

6. Prompts should avoid issues that are controversial either politically or socially (e.g., abortion or capital punishment) to reduce the possibility that raters might be biased against compositions that express views conflicting with their own beliefs.

7. Expectations for length, time limits, and scoring criteria should be clearly stated in the prompts. (pp. 287-288)

After careful consideration of such criteria, two expository prompts were selected (one for the pretest and one for the posttest) and adapted from published prompts (Breland, Camp, Jones, Morris, & Rock, 1987). The prompts published by Breland et al. had already undergone careful scrutiny, revision, pilot testing, and use, and their findings have been published.

Another approach to the validity issue as it relates to holistic scoring was described by Charney (1984) as practical validity. The issue concerns the manner in which essays are scored. Charney had this to say:

> A valid assessment of writing ability includes a natural human response to a writing sample. If readers can be trained to respond in a consistent, acceptable way, then the ratings will be reliable and valid. The requirements for achieving this condition are fairly complex. Those who use holistic scoring assume that the assessments will be valid and reliable:
>
> if the design of the training and rating sessions takes the factors necessary for reliability into account;
>
> if the readers are qualified, and come from similar backgrounds;
>
> if the readers are "calibrated," that is, trained to conform to agreed upon criteria of judgment;
>
> if the criteria, which either are supplied to the readers in the form of a rating guide, or are decided upon by the readers as a group, are appropriate; and
>
> if readers work quickly, usually under supervision. (p. 69)

Most elements that Charney presented have been discussed earlier in this symposium. The single issue not yet discussed is also the one that is most difficult to resolve. It centers on the issue of the appropriateness of the scoring criteria. Charney commented further on this issue, saying:

> Since, under normal circumstances, it is difficult to secure agreement on quality, any method that simply selects one standard is likely to be rejected by those who uphold alternative standards. As a result, the face validity of a given test of writing ability depends on whether one agrees with the criteria for judgment established for the ratings. (p. 73)

The holistic scoring criteria used to rate the writing samples in the present study were selected after careful consideration of several sets of criteria. Among the criteria examined and considered were scoring criteria from the Alabama State Department of Education (A. Moody, Personal Communication, May 14, 1992); Swartz, Patience, and Whitney (1985); and White and Polin (1986). The criteria finally selected were adapted from the criteria given by Swartz et al. (1985). Ultimately, judgments of the validity of the rating process must be a local decision based on rating scale users' agreement with the criteria.

Criterion-Related Validity

As Charney (1984) indicated, problems are inherent in any correlations that purport to assess the criterion-related validity of a direct writing assessment. Correlations with indirect measures have been criticized because the indirect measures have been "rejected as valid measures of writing ability" (p. 76). Correlations with grades have been rejected as reflecting other abilities in addition to writing. Correlations with other writing samples have been rejected because they are the very measures whose validity is in question. Although the issues presented by Charney are not without merit, such measures appear to be the only ones available for assessing criterion-related validity. Thus, despite their limitations, two of these measures have been considered in this study.

Criterion-related validity was first assessed by obtaining correlations of each of the holistic writing means with the BCT language score. Results are reported in Table 2. The correlations in parentheses are those obtained after correcting for attenuation using the formula given by Ferguson (1981). The reliability coefficients used to compute the corrected correlations were .52 for the holistic pretest, .57 for the holistic posttest, and .952 for the BCT language test.

Table 2

Correlations of Holistic Means with BCT Language Scores

	Pairwise Analyses	Common Analyses
Pretest Mean With Language	.50*(.70) (N=339)	.49*(.70) (N=331)
Posttest Mean With Language	.41*(.56) (N=414)	.36*(.49) (N=331)

* $p < .001$

These results are consistent with findings from other studies that compare holistic writing scores to scores on indirect measures of writing. For example, Veal and Hudson (1983) found correlations of .42, .40, .70, and .57 respectively between a holistically scored direct writing assessment and the language portions of each of four objective tests: Tests of Achievement and Proficiency (TAP), California Achievement Test (CAT), Iowa Test of Basic Skills (ITBS), and Writing Proficiency Test (WPT). Miller and Crocker (1990) conducted a review of studies comparing indirect and holistically scored direct writing assessment and reported among their findings seven studies in addition to the Veal and Hudson results in which the correlations ranged

from .46 to .76. Correlations of .55 to .59 were found between writing samples and an objectively scored writing skills test in a study reported by Swartz et al. (1985).

McKendy (1992) indicated that another way of assessing the criterion-related validity of a direct writing assessment was to examine its relationship to other writing samples. He cited earlier studies by Godshalk, Swineford, and Coffman (1966) and by Breland and Gaynor (1979) in which such a procedure was suggested.

Following this suggestion, the pretest mean writing sample scores were correlated with the posttest mean scores. The correlation between these two measures was .43 ($p < .001$) for the 336 students who had valid pretest and posttest mean scores. This relationship accounts for 18% of the variance. The correlation between the two measures for the 331 students with all five scores available was .43 ($p < .001$). These corrections, when corrected for attenuation, become .78 and .79 respectively.

Construct Validity

One method of examining construct validity involves the consideration of convergent and discriminant validation. Such analyses have as their aim gathering evidence that a measure correlates highly with variables with which it is theorized to correlate, and the same measure does not correlate highly with measures theoretically different from it (Anastasi, 1982). Results considering the relationship of the holistic writing pretest and posttest mean scores with the three BCT subject area scores are reported in Tables 3 and 4. Again, corrections for attenuation are shown in parentheses. Additional reliability coefficients used were .951 for the BCT reading test and .957 for the BCT mathematics test.

Table 3

Correlations of Holistic Writing Pretest and Posttest Means with Grade 9 BCT Language, Reading, and Mathematics Results - Pairwise Analyses

	Language	Reading	Mathematics
Pretest Mean	.50*(.71) (N=339)	.51*(.72) (N=337)	.44*(.62) (N=338)
Posttest Mean	.41*(.56) (N=414)	.40*(.54) (N=413)	.32*(.43) (N=413)

* p < .001

Table 4

Correlations of Holistic Writing Pretest and Posttest Means with Grade 9 BCT Language, Reading, and Mathematics Results - Common N Analyses (N=331).

	Language	Reading	Mathematics
Pretest Mean	.49*(.70)	.50*(.72)	.43*(.62)
Posttest Mean	.36*(.49)	.35*(.48)	.29*(.40)

* p < .001

Although the evidence is not striking, a trend is apparent. The correlations involving the mathematics scores tend to be somewhat lower than those for language or reading (a subtest measuring verbal skills considered to be more closely related to language subtest skills than are mathematics skills). The trend becomes more apparent when the individual competencies from each subtest are correlated with holistic results. These correlations can be found in Table 5.

A total of 21 out of 23 (91%) of the correlations between the pretest mean scores and the language competencies were statistically significant. The numbers decline with reading

competencies (18 out of 21, or 86%) and mathematics competencies (19 out of 25, or 76%). A similar trend can be seen in the correlations between the competencies and the posttest mean scores. For language competencies, 13 out of 23 (57%) were significant, whereas 8 out of 21 (38%) of reading competencies and 7 out of 25 (28%) of mathematics competencies were significant.

Table 5

Correlation of Holistic Mean Scores with BCT Competencies (N=331)

Language			Reading			Math		
Comp	Pre	Post	Comp	Pre	Post	Comp	Pre	Post
1	.25*	.17	1	.35*	.24*	1	.23*	.20*
2	.13	.11	2	.22*	.14	2	.16	.14
3	.26*	.15	3	.23*	.22*	3	.23*	.18
4	.34*	.16	4	.21*	.08	4	.19	.14
5	.20*	.06	5	.34*	.20*	5	.26*	.16
6	.32*	.24*	6	.20*	.15	6	.27*	.16
7	.38*	.26*	7	.35*	.10	7	.25*	.17
8	.25*	.20*	8	.27*	.17	8	.16	.08
9	.24*	.23*	9	.25*	.10	9	.26*	.11
10	.26*	.13	10	.29*	.21*	10	.22*	.15
11	.34*	.30*	11	.24*	.25*	11	.17	.14
12	.22*	.19	12	.19	.16	12	.22*	.17
13	.29*	.17	13	.19	.16	13	.33*	.31*
14	.26*	.24*	14	.34*	.23*	14	.20*	.15
15	.28*	.24*	15	.33*	.23*	15	.24*	.11
16	.21*	.13	16	.23*	.18	16	.20*	.20*
17	.32*	.22*	17	.20*	.19	17	.30*	.17
18	.39*	.27*	18	.33*	.16	18	.34*	.19*
19	.30*	.29*	19	.23*	.22*	19	.36*	.20*
20	.14	.07	20	.25*	.15	20	.25*	.23*
21	.29*	.20*	21	.14	.14	21	.19	.11
22	.31*	.25*				22	.29*	.24*
23	.24*	.21*				23	.20*	.11
						24	.33*	.15
						25	.14	.10

* $p < .001$

Conclusion

The evidence accumulated in this study provides some limited evidence for the validity of the writing assessment as used in the Blue Ribbon Committee Project. Procedures for selecting writing prompts and for selecting and using scoring guides were in keeping with those suggested in the literature. The moderate corrected correlations of holistic essays with the BCT language scores suggest that the writing outcomes are measuring characteristics similar but not identical to those measured by objective methods. The trend in the correlations of the holistic scores with other objective measures suggests the stronger relationships with theoretically similar measures than with theoretically dissimilar measures.

References

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan Publishing Co., Inc.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill, Research Monograph No. 11. College Entrance Examination Board, New York.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18, 65-81.

Ferguson, G. A. (1981). Statistical analysis in psychology and education (5th ed.). New York: McGraw-Hill.

Swartz, R., Patience, W., & Whitney, D. R. (1985, July). Adding an essay to the GED Writing Skills Test: Reliability and validity issues. GED Testing Service Research Studies Number 7, Washington, DC: American Council on Education.

McKendy, T. (1992). Locally developed writing tests and the validity of holistic scoring. Research in the Teaching of English, 26, 149-166.

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. Applied Measurement in Education, 3, 285-296.

Standards for Educational and Psychological Testing. (1985). Washington, DC: American Psychological Association.

Teague, W. (1989). Minimum standards and competencies (reading, language, and mathematics) for Alabama Schools (1989 edition). Bulletin 1989, No. 37. Montgomery, AL: Alabama State Department of Education.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. Research in the Teaching of English, 17 290-296.

White, E. M., & Polin, L. G. (1986). Research in effective teaching of writing: Volumes I and II. Los Angeles: California State University Foundation. (ERIC Document Reproduction Service No. ED 275 007)