DOCUMENT RESUME

ED 220 871 CS 207 207

AUTHOR McCready, Michael A.; Melton, Virginia S.
TITLE Feasibility of Assessing Writing Using Multiple Assessment Techniques. Research Report.
INSTITUTION Louisiana Tech Univ., Ruston. Coll. of Education.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Dec 81
GRANT NIE-G-80-0195
NOTE 144p.; Several pages in appendix may be marginally legible.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Conferences; Educational Assessment; Elementary Secondary Education; Essay Tests; Evaluation Criteria; Evaluation Methods; National Surveys; Objective Tests; *School Districts; *State Agencies; State Standards; *Test Format; *Writing Evaluation; *Writing Research
IDENTIFIERS Louisiana

ABSTRACT

Many local and state education agencies in the United States now mandate annual assessments of student writing ability. To comply with these mandates, schools across the country have developed a variety of assessment procedures. To determine the "state of the art" of large-scale writing assessment in the country, a questionnaire was sent to all 50 state education agencies and selected city agencies concerning their assessment practices. From those agencies responding to the questionnaire, 10 representing varying assessment philosophies were invited to send participants to a conference on assessment held in New Orleans. The purpose of this conference was to provide both verification and clarification of the questionnaire findings. In addition, the study drew data about assessment practices from schools throughout the state of Louisiana. These three activities yielded conclusions concerning such problems as (1) the selection of scoring procedures, (2) development of test items, (3) the selection and training of scorers, (4) the value and use of information produced by a writing sample, (5) the scoring of mechanics, and (6) other problems associated with large-scale assessment programs. (Appendixes contain a list of city school systems receiving the questionnaire, a copy of the Louisiana scoring guide for writing samples, and names of participants in the writing conference. Extensive tables of data are included in the paper.) (FL)

FEASIBILITY OF ASSESSING WRITING USING MULTIPLE ASSESSMENT TECHNIQUES

A Research Report

Presented to

NATIONAL INSTITUTE OF EDUCATION

Unsolicited Proposals

by
Michael A. McCready
Virginia S. Melton

NIE-G-80-0195

Louisiana Tech University
College of Education
Ruston, Louisiana

December 1981

Table of Contents

List of Tables

Acknowledgements

The investigators would like to acknowledge the assistance of the following persons in the completion of this research:

Dr. Hugh Peck, Associate Superintendent of Education, Louisiana State Department of Education, Division of Research and Development

Mr. Joe Williams, Director, Bureau of Accountability Louisiana State Department of Education

Mrs. Rebecca Christian, Assistant Director, Bureau of Accountability, Louisiana State Department of Education

Ms. Donna Nola, Supervisor, Bureau of Accountability, Louisiana State Department of Education

Dr. Tom Springer, Associate Professor, Louisiana Tech University (Programmer)

Louisiana Tech University Computer Center

Dr. Alan Purves, University of Illinois at Urbana (Consultant)

Dr. Ina Mullis, National Assessment of Education Progress (Consultant)

Dr. William Lutz, Rutgers University (Consultant)

CHAPTER 1

THE PROBLEM

The purpose of this study was to determine the feasibility of scoring writing samples in addition to objective measures in large-scale assessment of written composition and to determine the status of writing assessment among local and state education agencies. Feasibility was measured in terms of cost, time and management factors using a writing sample in relation to information yielded as compared with information yielded using objective measures. The state of the art was determined by a survey completed by practitioners in local and state education agencies. In light of the central question of the study, alternatives for scoring writing samples were explored. Also possible management techniques were identified which might moderate the cost of scoring large numbers of writing samples.

The Research Questions

This proposed study was designed to determine the feasibility of scoring a writing sample in addition to an objective measure of writing in a statewide writing assessment. The following questions were addressed:

1. What is the relationship between students' primary trait score on a writing sample and their corresponding score on an objective measure of writing mechanics for each group of students in the design?

2. What is the relationship between a syntax score on a writing sample and the syntax score on the objective measure for each group of students?

3. What is the relationship between the number of capitalization errors on the writing sample and the number of capitalization errors on the objective test for each group of students?

4. What is the relationship between the number of punctuation errors on the writing sample and the number of punctuation errors on the objective test for each group of students?

5. What is the relationship between the number of spelling errors on the writing sample and the number of spelling errors on the objective test for each group of students?

6. Are the mean scores on the objective test for each group of students in the design significantly different from the mean scores for every other group?

7. Are the mean scores for primary trait on the writing sample for each group of students in the design significantly different from the mean scores for every other group?

8. Is a profile analysis constructed for each group of students from data yielded from an evaluation of a writing sample similar to a profile analysis constructed for that group from data yielded from an objective test of writing?

9. Is the cost of evaluating a writing sample justified based upon additional information yielded?

10. Is the time involved in a writing sample justified based upon additional information yielded?

11. What is the "state of the art" of writing assessment among SEA's and LEA's with respect to the proposal's central question: "What are the relative costs of assessing writing by means of standardized tests versus judged written essays and what are the additional kinds of information yielded by the latter that may justify the added cost?"

12. What are alternatives for scoring writing samples or what are other strategies for state-wide assessment in which the added cost of detailed analyses can be moderated?

Delimitations of the Study

In order to address the research questions identified in this study, two populations were selected. The delimitations of each population is described below.

Louisiana Student Population

In order to address the questions involving the relationships between student scores from objective measures and corresponding scores on writing samples, results of the Louisiana state-wide writing assessment were used. In the 1979 Louisiana Assessment Program, students in grades four, eight, and eleven were administered proficiency tests in April of that year. All students at the designated grade levels participated in the program. Objective tests for all students were scored. However, a random sample of 2,500 writing exercises per grade level were selected. Scorers consisted of twenty-five classroom teachers who had been recommended by their respective superintendents.

Practitioners in Local and State Education Agencies

In order to determine methods practitioners are using to assess writing, a questionnaire was mailed to each of the fifty state education agencies and to fifty large city school systems (See list in the Appendix.) From the respondents ten persons were randomly selected to be invited to a conference to discuss their respective writing assessment programs.

Importance and Significance of the Study

Accountability legislation in forty states has charged state education agencies and local education agencies to provide information relative to student achievement in the basic skills. The assessment of reading and mathematics appears to have been relatively straight forward. However, the writing assessment has presented a particular challenge in that a variety of methods has been used to score writing. Two schools of thought appear to be emerging, "one decrying objective testing and the other insisting that most important mental processes, including the composing of essays, can be measured well by objective items" (Stanley and Hopkins, 1972).

Apparently, objective tests can be constructed to sample the domain of rules and thereby reliably measure how well students understand them. However, there appears to be some question as to whether a proficiency in the mechanics of grammar assures a proficiency in written composition (Coffman, 1969).

The concern about how to measure writing competencies of American students was intensified in October 1975, with the publication of the Writing Assessment report by the National Assessment of Educational Progress (NAEP). The findings demonstrated an apparent decline in the quality of writing by the nation's students (NAEP, 1975).

Following the NAEP report on writing mechanics, the College Entrance Examination Board, which had recently relied on objective measures, announced that it would once again begin requiring a writing sample as a part of the test of writing ability of college applicants (Godshalk, Swineford, and Coffman, 1966).

Coffman (1971) recognized that writing samples could be scored with a high degree of reliability. At the same time he cautioned that scoring is expensive, requiring large amounts of professional time and may, therefore, be impractical in large numbers. The question, then, is as follows: Does the scoring of a writing sample offer enough additional information to justify the professional time and cost involved?

An underlying objective of assessment is to produce reliable and valid information from which instructional decisions can be made that will increase student achievement in the state. Instructional decisions formulated solely from objective measures of the mechanics of writing are based on the assumption that students can increase their proficiency in writing prose by increasing their proficiency in the mechanics of writing. A review of the literature indicates that several researchers have found a positive relationship between certain quantitative measures of writing mechanics and the quality of writing (Howerton, et al., 1977).

Bloom, et al. (1977) examined the relationship between knowledge of mechanics and the ability to write compositions. Findings indicated that learning mechanics was only weakly related to the quality of compositions and then only certain students demonstrated this relationship. Past a certain level, remedial students demonstrated no significant relationships. These findings suggest the hypothesis that all students can not transfer their knowledge of the mechanics of writing and, at the same time, marshall ideas about a given topic and establish an organization and structure to convey meaning.

Chapter 2

REVIEW OF LITERATURE

This section of the study includes a review of the literature related to assessment of written composition and a summary of a consultation with identified authorities in the field. Two major computer searches were conducted in order to establish a working bibliography. One search was conducted by Educational Research Systems in Arlington, Virginia and the other was conducted through the ERIC System. Printouts were reviewed and related articles were ordered. In addition Dr. Willialm Lutz, consultant to the study, supplied a working bibliography. Authorities in the field of writing as identified by the Project Officer included Dr. Ina Mullis of National Assessment of Educational Progress, Dr. Alan Purves of the University of Illinois at Urbana and Dr. William Lutz of Rutgers University. These consultants were invited to meet with the project directors on the campus of Louisiana Tech University on March 20, 1981. The purposes of the meeting were: 1) to discuss the status of writing assessment, the methods used in assessment of written composition, and the problems associated with writing assessment, and 2) to plan the summer conference with practitioners from local and state education agencies.

Summary of Related Research

This study represents an attempt to determine the status of large-scale writing assessment in the nation. A new emphasis on assessment has been initiated by the Accountability Movement which began its sweep across the nation in the seventies. This new emphasis on assessment has nurtured changes in traditional testing and measurement theory, contributed to the development of criterion referenced tests, and promoted the specification of minimum competencies expected of students.

Approximately 40 states are actively developing or using minimum competency tests. Writing assessment is a part of the minimum competency tests in many states (Education Commission of States, 1979). The primary problem facing assessment decision-makers is the question of whether to use direct measurement techniques or indirect measurement techniques in the assessment of writing. Many authorities classify the two methods of evaluation of writing as "holistic" and "atomistic." Atomistic tests include the conventional multiple-choice lists of usage and mechanics, vocabulary tests as a measure of skill in discourse, measures of sentence length and complexity, readability formulas, and other measures of rhetorical conventions which are quantifiable. A user of atomistic tests assumes that the correlation between mastery of the identified feature and the art of discourse is close enough to permit predictions about skill in writing. Holistic tests include those evaluation techniques which depend upon the examination of a sample of writing. According to Cooper, holistic tests include holistic scoring techniques, analytic scoring techniques, and primary trait scoring techniques. However, other authorities criticize the grouping of analytic scoring and primary trait scoring with holistic. (Cooper, page 3).

Atomistic tests traditionally have been short-answer tests which yield information about particular features of language. An example of such a

test is the "GED Writing Skills Test." Skills measured include spelling, punctuation, capitalization, grammar and usage, diction and style, sentence structure, and logic and organization. The examinee simply recognizes errors and makes a choice of effectiveness. Another example of an atomistic test is the "Written English Expression Placement Test." The first part of this test has twenty multiple-choice items on various aspects of punctuation and syntax in which the examinee identifies errors. On the second part, the examinee must identify which of three versions of a sentence is the best one.

Test reviews in the Mental Measurement Yearbook do not indicate any indices for objective tests of writing skills which predict writing ability. The Sequential Test of Educational Progress (STEP) recently changed its name from "Writing" to "Mechanics in Writing" in response to the recent criticism of indirect measures of writing (Burros, 1980). Whether writing ability can validly be measured by multiple choice tests is a very old question. Richard Braddock insists that objective tests measure only proofreading skills (Braddock, 1979). On the other hand a classical ETS study of the 1960's indicated that sixty-minute objective tests of writing skills can correlate above .70 with a reliable criterion of composition scores. However, the problem is compounded by the fact that over 200,000 students took the English Composition Test of the College State Boards in 1976-76 (Godshalk and Swinford, 1969). Regardless of the technique which is used, scoring such a large number of writing samples simply is not feasible.

The Conference on College Composition and Communication (CCCC) has taken a definite position on direct measures of writing. First, CCCC has objected to the inclusion of objective usage tests on the grounds that such tests measure copyreading skills rather than the ability to use language. Further, CCCC felt that such tests discriminate against minority students in that the answers are different from their language patterns. Also, it was felt that secondary English teachers would teach to the test and neglect experiences in writing (CCCC, 1974).

Again in 1978 CCCC passed a resolution concerning the use of direct measures. The resolution stipulated that no student should be given course credit, placed in a remedial writing course, exempted from a required writing course, or certified for competency without submitting a writing sample. The resolution called for further study of the entire issue of testing (CCCC, 1979).

Direct measures for the evaluation of writing effectiveness are not without limitations. Braddock warns that the time limitation will inturn place limitations on students which will cause them to produce a writing sample under artificial circumstances (Braddock, 1979). Sanders and Littlefield concur with this generalization by claiming timed impromptu conditions as well as assigned topics limit both motivation and quality (Sanders and Littlefield, 1979). Diederich insists that at least two writing samples are needed to allow students a chance to do their best (Diederich, 1960). Coffman described direct measurement techniques as expensive in that they require large amounts of professional time to rate a representative sample of papers (Coffman, 1966). The question remains that

while direct measurement may be desirable in large scale assessment it may not be feasible.

At the present time many state departments of educations and large school systems are implementing assessment programs which include direct measurement. The methods of scoring or rating the writing samples vary. A review of the literature indicates three major techniques which are discussed in this following section.

Analytical Scoring

Analytical scoring is based on the assumption that the quality of a writing sample can be judged by comparing the sample to a predetermined rubric. The rubric consists of specified characteristics which are determined to describe effective writing regardless of the mode. The characteristics are usually divided into categories of general merit and mechanics.

A definitive scale was developed by Diederich, French, and Carlton (Diederich, 1974). In order to define the characteristics of effective writing, the researchers selected 300 writing samples from a larger number of college freshmen students enrolled in English. One group of students was assigned the topic, "Who Should Go to College?" A second group of students was assigned the topic, "Why Should Teenagers be Treated as Adults?" From each group a sample of 100 papers was selected. To score the papers, sixty readers were selected from six different fields (Diederich, 1946).

When the scores assigned by the various groups of readers were correlated with the scores assigned by every other group of readers, the resulting correlations were low (.31.) The scores of English teachers produced higher correlations than those of any other group (.41) (Diederich, 1946).

To find out what school of thought tended to exist among the readers, the inter correlations were subjected to factor analysis. The factor analysis produced the following five clusters: ideas, form, flavor, mechanics, and wording. Apparently some readers tended to sort papers based on how well the writer expressed "ideas," while other readers considered "mechanics" as the primary criteria, and others based their judgments on "form" or "flavor" or "wording". While these five clusters had no practical application specified in the original purpose of the study, they did suggest characteristics which might be included in a grading system (Diederich, 1946). A result was the formulation of the analytical method.

Diederich applied the analytical scoring method in grading English compositions written by students in three large high schools. Students wrote one paper per month on an assigned topic. The five factors previously identified were used as the scoring criteria and were assigned weights. After using the scale one year, the researchers and found that the scale divided itself into the two categories of general merit and mechanics. The category of general merit included ideas and organization which were assigned double weight. (See Scale on Next Page.)

DIEDERICH SCALE

TOPIC_____ READER_____ PAPER_____

	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	_____
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	_____
	10	20	30	40	50	
	E	D	C	B Sum	A	_____

*Note that more emphasis is given to "ideas" and "Organization" than to the others.

Each feature on the scale was described in detail with high-medium-low points identified and described along a scoring line for each feature (Diederich.)

I. GENERAL MERIT

1. Ideas

HIGH. The student has given some thought to the topic and writes what he really thinks. He discusses each main point long enough to show clearly what he means. He supports each main point with arguments, examples, or details; he gives the reader some reason for believing it. His points are clearly related to the topic and to the main idea or impression he is trying to convey. No necessary points are overlooked and there is no padding.

MIDDLE. The paper gives the impression that the student does not really believe what he is writing or does not fully understand what it means. He tries to guess what the teacher wants and writes what he thinks will get by. He does not explain his points very clearly or make them come alive to the reader. He writes what he thinks will sound good, not what he believes or knows.

LOW. It is either hard to tell what points the student is trying to make or else they are so silly that, if he had only stopped to think, he would have realized that they made no sense. He is only trying to get something down on paper. He does not explain his points; he only asserts them and then goes on to something else, or he repeats them in slightly different words. He does not bother to check his facts, and much of what he writes is obviously untrue. No one believes this sort of writing -- not even the student who wrote it.

2. Organization

HIGH. The paper starts at a good point, has a sense of movement, gets somewhere, and then stops. The paper has an underlying plan that the reader can follow; he is never in doubt as to where he is or where he is going. Sometimes there is a little twist near the end that makes the paper come out in a way that the reader does not expect, but it seems quite logical. Main points are treated at greatest length or with greatest emphasis, others in proportion to their importance.

MIDDLE. The organization of this paper is standard and conventional. There is usually a one-paragraph introduction, three main points each treated in one paragraph, and a conclusion that often seems tacked on or forced. Some trivial points are treated in greater detail than important points, and there is usually some dead wood that might better be cut out.

LOW. This paper starts anywhere and never gets anywhere. The main points are not clearly separated from one another, and they come in a random order -- as though the student had not given any thought to what he intended to say before he started to write. The paper seems to start in one direction, then another, then another, until the reader is lost.

3. Wording

HIGH. The writer uses a sprinkling of uncommon words or of familiar words in an uncommon setting. He shows an interest in words and in putting them together in slightly unusual ways. Some of his experiments with words may not quite come off, but this is such a promising trait in a young writer that a few mistakes may be forgiven. For the most part, he uses words correctly, but he also uses them with imagination.

MIDDLE. The writer is addicted to tired old phrases and hackneyed expressions. If you left a blank in one of his sentences, almost anyone could guess what word he would use at that point. He does not stop to think how to say something; he just says it in the same way as everyone else. A writer may also get a middle rating on this quality if he overdoes his experiments with uncommon words; if he always uses a big word when a little word would serve his purpose better.

LOW. This writer uses words so carelessly and inexactly that he gets far too many wrong. These are not intentional experiments with words in which failure may be forgiven; they represent groping for words and using them without regard to their fitness. A paper written in a childish vocabulary may also get a low rating on this quality, even if no word is clearly wrong.

4. Flavor

HIGH. The writing sounds like a person, not a committee. The writer seems quite sincere and candid, and he writes about something he knows, often from personal experience. You could not mistake this writing for the writing of anyone else. Although the writer may assume different roles in different papers, he does not put on airs. He is brave enough to reveal himself just as he is.

MIDDLE. The writer usually tries to appear better or wiser than he really is. He tends to write lofty sentiments and broad generalities. He does not put in the little homely details that show that he knows what he is talking about. His writing tries to sound impressive. Sometimes it is impersonal and correct but colorless, without personal feeling or imagination.

LOW. The writer reveals himself well enough but without meaning to. His thoughts and feelings are those of an uneducated person who does not realize how bad they sound. His way of expressing himself differs from standard English, but it is not his personal style; it is the way uneducated people talk in his neighborhood. Sometimes the unconscious revelation is so touching that we are tempted to rate it high on flavor, but it deserves a high rating only if the effect is intended.

II MECHANICS

5. Usage, Sentence Structure

HIGH. There are no vulgar or "illiterate" errors in usage by present standards of informal written English, and there are very few errors in points that have been discussed in class. The sentence structure is usually correct, even in varied and complicated sentence patterns.

MIDDLE. There are a few serious errors in usage and several in points that have been discussed in class but not enough to obscure meaning. The sentence structure is usually correct in familiar sentence patterns but there are occasional errors in complicated patterns; errors in parallelism, subordination, consistency of tenses, reference of pronouns, etc.

LOW. There are so many serious errors in usage and sentence structure that the paper is hard to understand.

6. Punctuation, Capitals, Abbreviations, Numbers

HIGH. There are no serious violations of rules that have been taught -- except slips of the pen. Note, however, that modern editors do not require commas after short introductory clauses, around nonrestrictive clauses, or between short coordinate clauses unless their omission leads to ambiguity or makes the sentence hard to read. Contractions are acceptable -- often desirable.

MIDDLE. There are several violations of rules that have been taught -- as many as usually occur in the average paper. Counts of such errors in high, middle, and low papers at various ages and socioeconomic levels would be desirable in order to establish standards.

LOW. Basic punctuation is omitted or haphazard, resulting in fragments, run-on sentences, etc.

7. Spelling

HIGH. Description of spelling levels are most often used in grading test papers written in class. Since there is insufficient time to make full use of the dictionary, spelling standards should be more lenient than for papers written at home. The high paper (at ages 14-16) usually has not more than five misspellings, and these occur in words that are hard to spell. The spelling is consistent; words are not spelled correctly in one sentence and misspelled in another -- unless the misspelling appears to be a slip of the pen. If a poor paper has no misspellings, it gets a high rating on spelling, even if no difficult words are used.

MIDDLE. There are several spelling errors in hard words and a few violations of basic spelling rules, but no more than one finds in the average paper. Spelling standards differ so sharply from grade to grade and from one socioeconomic level to another that each school would do well to make a distribution of spelling errors per hundred words (at least for test papers written in class) and relate its ratings to this distribution.

LOW. There are so many spelling errors that they interfere with comprehension.

8. Handwriting, Neatness

HIGH. The handwriting is clear, attractive, and well spaced, and the rules of manuscript form have been observed.

MIDDLE. The handwriting is average in legibility and attractiveness. There may be a few violations of rules for manuscript form if there is evidence of some care for the appearance of the page.

LOW. The paper is sloppy in appearance and difficult to read. It may be excellent in other respects and still get a low rating on this quality.

Readers learn to use the scale by studying the descriptions of high, mid and low values for each feature. Then, they use the scale to score samples of student writing. Once the scale is used by readers, they discuss the results until they have developed a systematic way of thinking about the papers. In order to attain some reliability, both in terms of the objectivity with which a paper is graded and the variation in the quality of the paper from time to time, Diederick offers the following guidelines:

1. Papers must be judged in accordance with a specified written criteria. Each criteria should be weighted.

2. Each paper must be graded independently by two readers.

Diederick's studies established the analytical scoring method as a reliable measure of writing ability and offered procedures to follow in developing a scale. Procedures, although simple to follow, are time-consuming. The first step is to collect large amounts of writing, both professional writing and student writing. The second step is to analyze the writing carefully so as to identify the prominent features which contribute to an effective writing sample. As the reading continues, the researchers discuss these features. Gradually, the features are shaped and modified until they have become comprehensive enough to describe a piece of writing but short enough to be manageable as a scoring guide.

Once the list of features has been determined, the next step is to describe in writing in nontechnical language what is considered to be high, mid, and low qualities of each feature. Cooper describes this process as helpful in "anchoring" the points along a scoring line (Cooper, 1977).

Finally, the completed guide is implemented to train readers who use the guide to score writing samples. Reliability checks are made to determine inter and intra reader agreement. This step is essential in research or curriculum evaluation studies.

The advantages of the analytical scoring method have been described by a number of researchers including Cooper (1977), Smith (1979), Pitts (1979), and Winter (1979). All agree that the method is quick, efficient, reliable, and yields diagnostic information. However, Cooper and Odell cite as a disadvantage the fact that the criteria for rating is not derived from a particular writing task or stimulus (Cooper and Odell, 1978).

Holistic Scoring

Holistic scoring of a writing sample is based on the assumption that the effectiveness of a piece of writing can be judged by readers' impressions of the writing as a whole. In scoring a writing sample using holistic scoring, an inductive procedure is followed. After the writing samples are collected the readers sort the samples into four stacks. The quality of the essay is judged only in relation to the other essays in the group rather than to a pre-determined rubric (Mellon, 1975). Once the papers are sorted, the readers then identify variables which appear to distinguish better writing from poorer writing.

The holistic procedure was developed by Godshalk, Swineford, and Coffman at Educational Testing Service (ETS). The scoring method was the result of a large-scale investigation which was designed to validate the use of the writing sample as a way to directly measure writing ability. The procedure assumes that the factors that make up writing are so closely related that they cannot be separated (Los Angeles County). The procedure has been extensively researched over the past twenty years, particularly by ETS in connection with the writing sample used in the College Board examinations (Godshalk, Swinefold, and Coffman, 1966).

In order to develop the holistic procedures, Godshalk and his associates collected 646 papers written on three different essay topics as follows: "Pen Pal," "Imagine," and "Teenager." These papers were read by twenty-five experienced readers who scored each paper on a three-point scale. Each paper was given five readings, an analysis of variance was then applied to the scores. Results of the study validated the use of a writing sample as a measure of writing ability. When the writing scores were combined with objective tests results an estimated reader reliability of .92 was established. However, scores tended to use the mid-point of the scale when in doubt about the quality of a paper.

The procedure was field-tested with a four-point scale by Godshalk and his collegues in 1966 in an effort to increase reader reliability by forcing the score away from the middle. In the field test, 533 papers were written on two of the original topics. The papers were read by 145 readers. Each paper was read five times by readers using a four-point scale and five times by readers using a three-point scale. The scores were then subjected to an analysis of variance. Findings demonstrated that scorer reliabilities, test validity, as well as other factors, tended to be more efficient on a four-point scale than on a three-point scale.

The holistic scoring procedure has been adapted to the scoring of writing samples which are a part of the College Board's English Composition Test. The following scoring procedures have been implemented to score writing samples from the College Board Test:

a. Each writing sample is scored independently on a four-point scale.

b. In scoring the paper, each reader makes two judgments:

   (1) First, the reader decides if the paper merits placement in the "upper half" or the "lower half."

   (2) Second, the reader decides if the paper is good enough to rate a "4" or weak enough to rate a "1."

c. The paper is read again by a second reader who follows the same procedures.

d. The scores from the two readings are summed resulting in a total score which may range from one to eight.

e. If there is a discrepancy of two score-points, the score must be reconciled by a third reader (Lutz, 1981).

Readers for the College Board undergo intensive training sessions in which the standards for scoring are set. Under the guidance of a Chief Reader, Readers score samples actually written for the test. The assigned topic is studied by the readers as they discuss and agree on the kind of writing that was required. Then papers are read and scored. Agreement tallies are made to determine that standards are followed. Because the readers are judging writing written on an assigned topic and because each paper is judged in relation to the other papers, scores are expected to fall into a pattern resembling a normal distribution.

Cooper states that general impression marking (holistic) is closer to analytic scoring than it may appear. In analytic marking, a reader compares the writing to a predetermined, printed rubric which describes high, middle, and low writing. While in holistic marking, one does not utilize a printed rubric, scorers do read and discuss sample papers until they do identify features or qualities which guide their judgment. Therefore, even though a rubric is not printed, a common scoring guide is understood among the readers. In analytical marking, it is not assumed that the pattern of scores resemble a normal curve. All of the papers or none of the papers may embrace the specified features (Cooper, 1977).

Primary Trait Scoring

When National Assessment of Educational Progress (NAEP) set out to assess the educational attainment of skill in writing among four age groups (9, 13, and 17 year olds, and adults) on a national basis, a writing sample was required. Holistic scoring as practiced by ETS readers in the scoring of College Board writing samples was used in the first program. The scoring procedures were severely criticized by English educators in the nation (Mellon, 1975).

Criticism of the national writing assessment was varied. The major criticism from the English teachers concerned the ranking procedures employed. According to Mellon, ranking essays yields very little information about a piece of writing other than that some writing samples are better than others (Mellon, 1975). Mullis charged that the score points were almost impossible to describe. For example, if one paper were scored a 2, and another a 4, one did not know what factors contributed to one paper being better than the other (Mullis, 1980). The most convincing argument against holistic scoring is presented by Lloyd Jones (Cooper and Odell, 1977). According to him, the assumption is made in holistic scoring that if a writer is effective in one mode of writing, that writer is effective in all modes. In response to this criticism, NAEP asked Lloyd-Jones and Carl H. Klaus to develop a system of scoring writing samples which defined precisely the writing mode being evaluated and to develop scoring guides to evaluate that mode. In developing the Primary Trait Scoring System, Lloyd-Jones and Klaus first chose a three-part discourse model which included explanatory persuasive and expressive writing modes. Next, they described each mode of writing in terms of the purpose of the discourse, the role of the writer, the effect on the

audience, and the content required by the writing task. After describing the discourse the researchers then developed writing exercises which would stimulate respondents to address the writing mode as defined. Since the writing mode had been carefully defined, the writing exercise was thereby restricted. This restriction created new problems in item development for test makers. Chances were increased that the writing exercise would fall outside the experiences of the respondents. With no knowledge of the specified situation, writer's response would not be a valid indication of his ability to write effectively (Lloyd-Jones, 1977).

Each exercise was field tested and the responses analyzed to determine if writers interpreted the exercise as testmakers intended. If writers did not respond as expected, the writing task was revised. The task was recategorized, rephrased, or completely changed.

Once the task was refined, a scoring guide was created for the exercise. A complete scoring guide consists of the following elements:

1. The exercise itself.
2. A description of the rhetorical trait of the writing.
3. An interpretation of the exercise indicating how each element of the stimulus is expected to affect the respondent.
4. An interpretation of how the situation of the exercise is related to the primary trait.
5. The actual scoring guide which is a shorthand system to be used in reporting descriptions of writing.
6. Sample papers which have been scored as representative of each score point.
7. Discussions of why each sample paper was scored as it was.

When a complete guide had been developed, it was given a feasibility check by Westinghouse Learning Corporation under the direction of Louise Diana. The observers used the scoring guide to rate papers which had previously been rated by the testmakers. The two scores were analyzed to determine if they were related. The reliability data were judged to be at least as good as data obtained in holistic scoring.

The Primary Trait Scoring System was implemented by NAEP in the Second Writing Assessment (1974) and used again in the Third Writing Assessment (1979). Holistic scoring was also used to determine if a group of papers written in 1979 was better than a group written in 1974. For this purpose, the same items were used in 1974 and 1979. Primary Trait Scoring was used to provide specific information about particular rhetorical aspects of papers. Papers were also scored for cohesion, that is, the general ways words and ideas were linked together in writing so as to create a sense of wholeness. In addition to being rated for quality, papers for two items were analyzed in terms of their syntactic and "mechanical" features. Thus, NAEP utilized a variety of scoring procedures in order to determine the status of writing in the nation.

Summary of Consultation with Identified Authorities

On March 20, 1981, Dr. Alan Purves, Dr. William Lutz and Dr. Ina Mullis met on the Louisiana Tech campus with project directors McCready and Melton to discuss the current status of writing assessment in the United States and to plan a national conference on writing assessment to be held in New Orleans in the summer. On page is the agenda of topics addressed in the March 20 conference.

A basic question that was addressed initially was "what is the role of large-scale assessment and can its purpose be achieved with an objective measure?" The consultants all agreed that the political impact of writing assessment is significant and that the impact of a writing sample is greater than that of the objective measure. The consensus was that the objective measure can yield valuable information but that the inclusion of the twenty minute essay adds a dimension to the assessment in terms of information yielded and political impact. Purves pointed out that in numerous other countries, writing samples are considered important because of political spinoffs. All agreed that in order to determine one's writing ability, we must have him write.

For purposes of meeting a legislative mandate, a statistically sound sample is sufficient. However, to provide feedback at the classroom level, every paper would need to be scored. Whether or not scoring every paper is feasible depends on the purpose of the assessment. If written composition is a priority item in the education budget, than it is feasible to score every paper down to the individual school level. In fact, scoring every paper might tend to have more of an impact on instruction while at the same time providing information (data) to meet the need for a statewide generalization about writing. If every paper is not scored, impact is diminished because it takes a long time for information to filter down from the top. Everyone seemed to agree that the writing sample is desirable and that a sample would suffice if the resulting data is adequately wrung for information and if that information is reported in a usable form to local education agencies.

Rebecca Christian, of the Louisiana Bureau of Accountability acknowledged the need to test every youngster because of the instructional implications. However, the cost of scoring is the source of the problem. She supported the scoring of a sample, and then making examples of responses available to teachers from their own classrooms. The state could make provisions for familiarizing teachers with the scoring procedure and give these teachers the opportunity to evaluate their own students' papers. Mrs. Christian also commented on the inherent difficulties of reporting the results of the primary trait system of scoring. As a result, Louisiana is simply not getting the full benefit of scoring a writing sample. In her opinion, if the primary trait system of scoring is used in its present form, even scoring every paper may not yield the state enough information to justify the cost.

The problem with a general writing assessment report to a state legislature is that there is a tendency of people looking at this data to say that the state has been engaged in the process for five years and yet

nothing is happening. Therefore student writing is not improving. If all papers are scored and the data is broken down into various categories, it becomes easier to show that students improved in one category but did not improve in another and that students in this system or school exhibit these strengths and weaknesses. Therefore if every student is assessed, a data base becomes available which can be reassembled in a variety of ways. If the need is for a generalization for the whole state, that is available. However if the need is for reporting at the classroom level, data is available for that purpose.

All the consultants recommended a writing sample and agreed that it has political fallout. All agreed that assessing every student is desirable but that a random sample would be effective also, if an effort is made to wring the data.

The consultants for the March 20 conference were selected based on their individual expertise about and experience with the various techniques of scoring written composition. In order to meet the objectives of the meeting, the directors asked that Ina Mullis represent primary trait scoring, that Bill Lutz represent holistic scoring, and that Alan Purves represent analytic scoring.

Purves shared the research findings of a study done by some of his graduate students using papers of 17 year-olds. The papers were scored using the Diederick Scale, a holistic scale, and primary trait. In terms of cost, according to the study, there was not a significant difference between primary trait and holistic. Both were quick. The Diederick Scale required three times as much time as either of the other two methods. However, there was a higher inter-rater reliability (.92) and intra-rater (.80) reliability with the Diederick Scale.

One of the topics discussed at some length by the consultants was what to do about scoring mechanics. NAEP makes a practice of pulling an additional sample to score for mechanics with one reading. In New Jersey, papers representing the various score points on the holistic scale are pulled and are scored descriptively for mechanics. At NAEP separate scorers are used to score mechanics. These scorers do not score for primary trait. One consultant strongly felt that scoring for mechanics reinforces the public misimpression that quality of writing is directly related to punctuation. His finding was that by scoring mechanics, we simply tend to feed that idea. He also asserted that it seems unfair to score a first draft for mechanics.

All consultants agreed that American students lack a clear perception of what is involved in the process of revision. When given the opportunity to revise, students generally want to recopy or, if any revision is done, it is cosmetic in nature.

One point of view which was projected where scoring of mechanics was concerned was that by singling out and scoring for various aspects of mechanics, a message is sent to teachers which might be interpreted as, "I need to teach more capitalization in my classes," rather than "I need to emphasize writing in my classes." The message which should be sent to

teachers in answer to the question, "What should I do to get students ready for the writing assessment?" is teach them to write by having them write.

There was some discussion among the consultants concerning differences among the three methods of scoring: holistic, primary trait, and analytical. It was pointed out that the New Jersey program was designed to be as non-threatening as possible. As a result a decision was made not to develop a scoring guide. Through discussion the scorers move toward agreement and verbally describe each level of the paper.

It was pointed out that the major difference between the holistic scoring and primary trait scoring is that in primary trait scoring, student writing is being measured using external criteria rather than being compared with one another. Also, the holistic approach tends to measure general fluency whereas primary trait is more concerned with the specified purpose of the writing. If general fluency were being measured for the national assessment, NAEP would consider it necessary to use some external criteria. It was pointed out that with holistic scoring the assumption is that fifty percent of those tested can write better than fifty percent, which is known before the assessment begins. Rank ordering can be used for measuring change. Assessment years can be compared to see which ones come out on top. The feeling at NAEP is that holistic scoring is efficient for measuring change in general fluency, but in terms of comparing writing data with the rest of the data of the national Assessment (mathematics, reading, etc.) the Primary Trait method is more efficient and more suitable.

One of the problems cited with holistic scoring is getting scorers to score a paper a "four." English teachers are particularly reluctant to give a paper a "four." Trainers have to be moved away from the idea that a four represents an "A." There is also a reluctance to assign the lowest score. The six-point holistic scale is being used more now which seems to give the scorers a feeling of a wider spread of scores; this brings about the assigning of more high and low score points. English teachers tend to be difficult to convince that if the sample is sufficiently large, statistically there will be a normal distribution of papers.

Criteria for selection of scorers was discussed at some length. All three consultants indicated that in their experiences, persons with backgrounds in English, specifically the teaching of composition, were used. All felt that one of the most important variables in the training of scorers was background of experience. One of the main reasons that Louisiana selected the Primary Trait System of scoring was that since classroom teachers were to be used in scoring the assessment, the consensus was that teachers with relatively limited backgrounds could be used due to the tight structure of the scoring guides. One consultant indicated that in selecting prospective scorers, one has to be careful of the person who talks a good game but who has very deep-seated unmovable prejudices, some of which even the person himself is not aware of. In training scorers, ground rules must be set which everyone agrees to follow, regardless of whether he/she agrees or disagrees.

Problems associated with development of writing tasks were discussed at some length. All agreed that identifying topics appropriate for boys as

well as girls, for city as well as rural youngsters, does present problems. Using teachers to generate topics does not solve the problem, since many topics which teachers think will work do not work in an assessment situation. Several variables were pointed out as being important in developing tasks. For example, is the topic one which will contribute to reader fatigue, or will the topic outrage, upset, or depress the reader. Some topic will stimulate only one type of response, a situation which can cause "brain numbness" on the part of the reader. One consultant suggested that topics which require the writer to assume and defend a particular position should be avoided, because the writer might not have a genuine position. If this position is not sincere, the writing becomes artificial and lifeless. NAEP is moving toward use of prompts which allow the writer to draw on personal experiences.

Of all the modes of writing, narration is the most difficult to score.

Audience identification varies in importance, depending upon the mode of writing. One consultant felt that identification of audience places artificial constraints on the writer. Students know the audience is make-bleieve and tend to write to the teacher anyway. Audience identification was defended on the grounds that once a student leaves school he no longer writes for a teacher, but rather for varying audiences. Therefore instructionally there should be emphasis on variation of audience in the classroom.

The question of the training of scorers was also addressed. When holistic scoring is used, much depends on the chief reader who selects the samples to be used in training. Most often training begins by looking at a three. If Primary Trait is used, begin with a solid "two," and then move to a "three." One difficulty in using primary trait scoring is that there is often more difference between "two's" than between the best "two" and the worst "three." In training scorers to an analytical scale, begin by using holistic scoring for the first papers, then share with scorers the rationale underlying the scale before they look at any more papers. In analytical scoring some ground rules have to be established as to where to score for sentence fragments or punctuation.

In training scorers to use the holistic method, the chief reader assumes responsibility for leading discussions with scorers to lead to agreement. Usually these discussions occur about three times during a scoring day. Also the chief reader moves among the scorers randomly picking up papers, scoring them and checking for agreement. The tables are constantly being monitored. Besides the chief readers, there are table leaders at each table.

All the consultants felt that scoring a writing sample is feasible depending on its purpose. All agreed that it constitutes the only way to find out if students can write or not. Multiple choice items tend to measure editing skills more than writing skills. A problem with the writing sample is that student achievement is measured based on a first draft, one which has not been revised. All agreed that in order for

scoring to be feasible, it must be streamlined. Primary trait and holistic scoring are more efficient, time wise, than analytical scoring, especially when the practice is used of pulling a smaller sample to score for mechanics.

It was pointed out that the "State of the Art" of writing assessment is not clearly defined at this time but that much progress has been made over the past ten years. Everyone hoped that impending budget cuts would not cause writing assessment programs to disappear.

CHAPTER 3

RESEARCH PROCEDURES AND FINDINGS

This study included two evaluation designs. One design was utilized to determine the relationship between an objective measure of writing ability and the use of a writing sample. The other design was used to determine the "State of the Art" of writing assessment in the nation. This section presents the research methodology employed to answer the research questions related to each of the two designs. As this study embraced more than one design, the findings for each study are reported immediately following the description of the design.

The Relationship Between Scores on Objective Tests and Scores on Writing Samples for Louisiana Students

The Louisiana State Assessment Program included written expression for the first time in the 1979-80 school year. As a part of the writing assessment, (October 1979) a sample of writing was secured from students in grades four, eight and eleven. From the total population at each grade level, 2500 writing samples were selected. The writing samples to be scored were identified by means of a stratified random sampling technique based on racial-ethnic group, socio-economic level of the region and population density. All special education students were deleted from the sample.

The Bureau of Accountability in the Louisiana State Department of Education had decided to use an adaptation of the primary trait scoring system developed by NAEP to score the writing samples. The Louisiana Trait system was developed by National Testing Service based on minimum competencies determined by Louisiana teachers. Items and scoring guides were developed under the direction of Dr. Stella Lieu (NTS).

In the fall of 1978, the Louisiana writing assessment instruments were field-tested under the direction of NTS. At that time NTS provided training for Louisiana state department personnel. Based on the observations and recommendations of NTS, items and scoring guides were revised for the spring testing.

In June of 1979 the directors of the project were contracted by the State Department of Education to train twenty-five classroom teachers as scorers. This training took place in a three-week institute described in the narrative below.

The Training Process for Scorers

The training of the Institute participants began with an overview and history of the Louisiana assessment program presented by staff members from the Bureau of Accountability in the State Department of Education. The objective of this phase of the training was to enable the participants to see their charge in the total context of the state assessment program.

An overview of the history of assessment was presented to the participants, with particular emphasis on the role that criterion-referenced testing has played in recent years. This phase of training led to a review of the problems associated with the assessment of composition. Various methods were reviewed including the holistic approach, writing scales, computerized scoring and others. The trainers felt that participants should be made aware of the pitfalls and strengths of various approaches to assessment of writing so that the primary trait system could be viewed with a clearer perspective.

With this background information, the trainers felt that the participants were adequately prepared for an introduction to the Primary Trait System of scoring by Dr. Wayne Martin of the National Assessment of Educational Progress.

After an introduction to the Primary Trait system, participants were prepared for their introduction to the fourth grade items and scoring guides. Demonstration scoring was done by the instructors using test papers that demonstrated the various levels of the guides. Several sample papers for each score point were discussed in detail. Then fourth-grade papers were distributed to participants so that practice scoring could be carried out. Packets had been prepared so that instructors knew the score designation of the papers being distributed. Discussion always followed each scoring. Then several sets of test papers demonstrating various score-point designations were distributed so that participants could learn to distinguish between papers of varying quality. After completing practice scoring for each item at the fourth-grade level, the process was completed for items at grades eight and eleven. The rationale underlying this approach was that all participants should be familiar with the items and scoring guides at all three levels so that a better perception of sequential development in writing could be attained. The perception of the scorers must be similar from grade level to grade level if the total evaluation is to be valid. The process required each participant to verify his rating with the minimum proficiencies and, therefore, facilitated a more thorough familiarization on the part of the participants with the state minimum standards in writing.

The next phase of training consisted of practice scoring by grade level. The participants were divided into three groups, one for each of the levels of testing. Packets of five papers of random mixture of score points were prepared by the trainers. Participants scored each set of practice papers, which were discussed until each reader was in agreement with the trainer about the score point. This phase of training was continued until scorers were scoring consistently. Reliability coefficients were determined daily.

The third phase of training consisted of practice scoring by individuals. This training simulated actual scoring procedures. Packets of ten booklets were prepared by trainers. After the participants scored the first packet of ten papers, the trainers realized that ten papers are too many to be scored in the training process. The discussion following scoring appeared to be crucial to reaching agreement. Therefore, the packets of ten were divided into packets of five. This procedure made it possible for all members of the group to score a packet and then to discuss the ratings.

Training continued with the participants scoring the same five tests, comparing their results, and discussing their differing opinions. The crucial phase of this stage of training appeared to be reaching a concensus of opinion on a score point. Always, the importance of achieving a high degree of reliability was emphasized by the trainers.

Reliability of Scorers

The major objectives of this Institute was that scorers would use the Primary Trait System with a high degree of scorer reliability. Reliability checks were taken after each packet was scored by each member of the group. Trainers recorded each scorer's rating of a paper on each primary and secondary trait and the percent of agreement on each was determined. Table 1, on the next page, shows the percent of agreement for two checks for each group on each trait.

From the resulting data it appears that scorers for fourth grade achieved a higher scorer reliability than scorers at other grade levels. Further, reliability tended to vary from item to item. The higher reliability might be explained by the fact that the writing task required students to write only a simple sentence at the fourth-grade level. The scoring guide was carefully defined for scoring the primary trait, a factor which eliminated subjectivity.

Some of the writing stems stimulated students to write more. For example, item four caused children to write more than item one and the reliability was lower. Apparently, the more students write, the more difficult it is to achieve a high degree of reliability of scorers. This conclusion is supported by the reliability per cents for eighth and eleventh grade. It appears that it is more difficult to define a primary trait at higher grade levels. Responses of students at these levels tend to be more varied and not as predictable as are responses of students at the fourth grade level. When subjective decisions must be made, reliability decreases.

Reliability was consistently higher on the secondary traits of capitalization and punctuation. These traits can be explicitly defined and are clearly specified in the minimum standards.

Lower reliability per cents were noted on syntax and spelling. This was caused largely by disagreement over the difference between "correct" spelling and "correct" usage. The fact that usage was counted as a syntactical error caused some confusion and led to disagreement.

Description of Instrument

Accountability personnel in the Louisiana State Department elected to use a scoring guide similar to guides developed by NAEP. Writing tasks were selected from the Louisiana Minimum Standards Document. Once a task was identified, an item stimulus was developed and a guide was developed to score the item. Development of items and guides were under the direction of Dr. Stella Liu (Wayne State University) who was contracted by NTS.

TABLE 1

		Primary Trait		Secondary Traits					
				Syntax		Spelling		Capitalization		Punctuation	
4th	1.	90	90			85	84	91	96	94	89
	2.	82	95			77	80	91	89	84	87
	3.	69	81			77	83	86	82	76	84
	4.	76	73			85	93	93	71	76	81
8th	1.	76	77	75	82	92	61	90	92	88	80
	2.	69	73	88	85	84	71	91	81	73	67
	3.	78	77	89	74	96	81	94	80	85	84
11th	1.	77	76	83	75	93	88	90	87	94	80
	2.	77	76	75	70	93	89	94	89	89	75

Scoring guides for a primary trait consisted of eight score points defined as follows.

0 - No response

1 - An attempt is made to respond to the request but fails to complete the specified task. This score indicates writing that is below minimum.

2 - The writer completes the task with minimal elaboration. This score indicates a minimum proficiency as specified in the Louisiana Minimum Standards for Writing.

3 - The writer completes the task with elaboration through added detail. Organization may be lacking. This score indicates writing that is above minimum but not excellent.

4 - The writer completes the task with elaboration through expanded details and descriptive terms. Organization is readily discernible and mature. This score indicates excellence in writing.

7 - Illegible. No further scoring.

8 - Illiterate. No further scoring.

9 - Misunderstands the task or writes on a totally different subject. No further scoring.

A scoring guide was developed for each item. Criteria for determining each score point were defined for each item. Scoring guides were used in scoring the writing responses secured in the field test. Problems were identified and revisions were made in guides. The revised guides were then implemented in scoring the state assessment program.

Mechanics were scored as secondary traits. The four secondary traits which were scored included spelling, syntax, capitalization, and punctuation. The scoring guide which were developed were based on the minimum standard document. The score point 2, which represented minimum proficiency, was assigned if the writer did not violate any convention specified for mastery at the designated grade level. If the writer violated a convention specified for mastery, a score of 1 was assigned. Above minimum was represented by a score of 3. The score was assigned if the writer attempted and used correctly any convention which was designated to be introduced but not mastered at a given grade level. A score of 4 was assigned if no errors were made. Examples of scoring guides are provided in the Appendix.

1979 Statewide Writing Assessment

The statewide writing test was administered in October 1979. In December 1979, scorers were reassembled for a one-day refresher session at the State Department of Education. State Department staff distributed

writing samples to be scored. Each scorer received a packet of one hundred papers, which they were to take home and score. When the entire packet was scored by the first scorer, the packet was mailed to a second scorer with the score sheets being sent to the State Department. When the second scorer had completed his/her work, any discrepancies were resolved by a third scorer.

All score sheets were sent to Multi-Media in Brazille, Louisiana to be key-punched. All data were filed on a tape which was forwarded to the researchers at Louisiana Tech. The tape was not compatiable to the computer at Tech. Therefore, the data were entered into the Tech computer for analysis.

Scorer Reliability

To determine rater agreement, a chi square test for independence was applied to determine the association of the first rater's scores to the second rater's scores on a designated trait for a writing sample scored by both raters. A SAS computer program was used to analyze the data. Contingency tables were generated for each trait scored on each item. Chi square provided a measure of discrepancy between observed cell frequencies and those expected on the basis of independence. If the chi square was found to be significant at the 01 per cent level, the null hypothesis that no difference existed between the observed and expected values was rejected. The alternate hypothesis that the two values were associated was accepted. All values were found to be associated.

From the chi square statistic, a contingency coefficient was generated. The contingency coefficient appeared to be the best correlation coefficient suited to the data. The contingency coefficient is a descriptive measure of the association between two nominal values and is independent of the ordering of the rows and columns on a contingency table. The minimum value is zero. The maximum value of the contingency coefficient .943 for the primary traits and .894 for the secondary traits. A summary of the results is shown in the Tables 2.1, 2.2, and 2.3 on pages

Conclusions

From the data analysis it is evident that there is little relationship between the first scorer's rating of traits and the second scorer's rating of the same traits. No variables contributing to this lack of relationship can be identified by the statistical analysis of the data. However, a study of the data analysis suggests certain variables that may contribute to the variance. In the following section, the variables are identified and recommendations are presented.

Scoring Guides

1. Scoring guides tended to be so general that too much interpretation was left to the discretion of the scorers.

2. There was an apparent confusion between scorers in recognizing a writing response rated as a three and one rated as a four. Seemingly, scorers could not agree on a writing response that was simply above minimum (3) and one that was exemplary writing (4).

3. Scorers could not agree on writing that was illegible or illiterate. For some scorers, if a sample was illegible, it was considered illiterate.

4. Scorers had difficulty agreeing upon writing that was below minimum (1).

Scoring Process

Scorers all evaluated the writing response at home over the Christmas holidays. Each scorer established his/her own hours and scoring schedule. At no point did scorers come together for sessions to clarify interpretations. Therefore, the entire scoring process was uncontrolled.

Items

An examination of the items resulted in the following conclusions.

1. Format of some of the items caused students to lose sight of the primary trait in their writing. The provision in the test format of too many lines caused students to write to fill those lines rather than to address the trait.

2. On fourth grade, Item 1, the direction "describe as much as you can about the picture" violates testing and measurement guidelines as well as the scoring guide. The scoring guide was structured to evaluate an item describing the location of persons. When students wrote as much as they could about the picture, they lost sight of location.

3. The eighth grade Item 2 posed two conflicting issues for the student to address, one on smoking and one on integrity. Students had difficulties in organization because of the nature of the item.

4. The eleventh grade Item 2 presented alternatives for the student to select and write a persuasive response. The nature of the alternatives varied so much that one scoring guide could not be adapted to all of them in the same way.

Training Process

There was too great a lapse of time between the training process and the actual scoring.

Recommendations

1. The following recommendations are suggested for revision in the scoring guides:

a. Each rating on both primary and secondary trait scoring guides must be clearly defined. In the appendix is a recommended scoring guide for each secondary trait at each grade level. The rating of two has been defined as writing which adheres to all conventions designated for mastery by the respective grade level in the minimum standards document (i.e., those items designated with three stars). The rating of three has been defined as writing which adheres to all conventions which have been introduced and are on-going (i.e., those items that are designated with two stars). The rating of one is simply defined as writing which does not meet the qualifications of a two.

b. A scoring guide for syntax has been designed for grade four. Separating the scoring of syntax from the scoring of spelling will remove the need for interpretations by the scorer. Also, scoring for syntax at fourth grade makes the scoring the same for all three grade levels.

c. On the primary trait scoring guides the following recommendations are made:

   1) Eliminate the rating of four.

   2) Collapse the ratings of seven and eight (illegible and illiterate).

2. The following recommendations are suggested for revisions in the scoring process.

   a. Scoring must be tightly controlled. Scorers must score as a group under the direction of trainers. Scoring guides must be strictly followed. A scoring schedule should be outlined by which scorers would systematically score a set of papers, exchange with a second scorer, with a third scorer reconciling discrepancies. Opportunities should be provided for scorers to regularly come together for conferences. Conference periods appear to be essential to consistent interpretations of the scoring guides.

3. The training process should immediately preceed the scoring of the samples or be a part of the scoring of the papers.

Revisions in the Louisiana Scoring Process

The Teacher Education Department in the College of Education at Louisiana Tech University was contracted to sponsor the scoring and analysis of writing samples for the Bureau of Accountability, Department of Research and Development, Louisiana State Department of Education. The writing samples were secured from a sampling of students in grades 4, 8, and 11 who responded to writing tasks developed in June 1980.

The writing tasks were developed by a committee of classroom teachers who had participated in scoring the 1979-1980 Assessment. At each grade level tested, four writing tasks were developed for each objective tested. The writing tasks had been formatted into four test forms at each grade level. Each test form was field-tested on a sample of 100 students. The number in the sample population was limited in order to facilitate scoring of the test.

It is difficult for the classroom teachers who score the writing sample to be released from the classroom. With this sample, teachers would be required to score 400 tests. As each test must be scored twice, the number to be scored was actually 800.

The sampling was conducted by the Bureau of Accountability; therefore the principal investigators had no control of the sampling techniques or distribution and administration of tests.

The Problem

The responsibilities of the Teacher Education Department at Louisiana Tech included the following:

1) Direct the scoring of the writing samples
2) Analyze the data resulting from the writing sample
3) Furnish reports relating to item analysis and scorer reliability

The procedures followed by the principal investigators are detailed in this report. Findings resulting from the data analysis are presented with conclusions and recommendations.

Scoring the Writing Sample

The findings from the 1979-1980 Writing Assessment suggested several changes that needed to be made in the scoring process. In an effort to improve scorer reliability, the principal investigators designed a form of "specialized-team" scoring. It was the opinion of the investigators that if scorers specialized in one area that they would increase scorer reliability because they would not have to consider but two scoring guides. Speciality teams were organized as follows:

1. One person to score Primary Trait
2. One person to score Spelling and Syntax
3. One person to score Capitalization and Punctuation

One person was assigned to score spelling and syntax because of the close relationship of the two traits. In order to get a valid measure of a student's proficiency with a given trait, it is essential that an error be counted consistently the same way each time. If syntax and spelling are scored separately, there is a tendency to count a wrong word choice both as a syntactical error and as a spelling error. With the same person scoring both traits, the scorer tends to score errors consistently in the same way.

Punctuation and Capitalization were grouped together for similar reasons. Correct terminal punctuation signals correct use of capital letters. Scoring guides relating to decisions often require that consideration be given to both capitalization and punctuation. By grouping the two together, it was hoped that scorer reliability would be improved.

In order to facilitate the scoring of secondary traits, the scoring guides for each trait were revised. All conventions which were identified in the State Minimum Standards Document as being "minimum" were listed for a score of "2". All conventions identified in the State Document as being introduced and maintained were listed for a score of "3". (Copy in the Appendix) The fact that conventions which should be adhered to by writers at each grade level were specified for scorers, removed the necessity of reference to the Document by the scorer. In this way, the principal investigators hoped to improve the scoring of secondary traits.

Two scoring teams were assigned to each grade level. One resolver and one clerk was assigned to each grade level. Test booklets were divided into packets of eight. Each scorer on each team was given a packet to score. The scoring guide shown on the next page was used to record the scores. As soon as all three scorers on a team scored a packet, the packets were given to the clerk who removed the first score sheet and placed in the packet the score sheet for the second scoring. A copy of the second scoring sheet is shown on page . Scoring teams then scored each packet a second time. When the second scoring was completed, the clerk placed the first score sheet beside the second score sheet and circled in red all scores that did not agree. The reconciler then considered each score where there was a disagreement and resolved it by deciding on one of the specified scores.

Findings

In order to determine scorer reliability, a chi-square test for independence was applied to determine the association of the first scorer's ratings on a designated trait for a writing sample scored by both scorers. An ASA computer program was used to analyze the data. Contingency tables were generated for each trait scored on each item. Chi square provided a measure of discrepancy between observed cell frequencies and those expected on the basis of independence. If the chi square was found to be significant at the .01 percent level, the null hypothesis that no difference existed between the observed and expected values was rejected. The alternate hypothesis that the two values were associated was accepted. All values were found to be associated.

From the chi square statistic, a contingency coefficient was generated. The contingency coefficient appeared to be the best correlation coefficient suited to the data. The contingency coefficiency is a descriptive measure of the association between two nominal values and is independent of the ordering of the rows and columns on a contingency table. The minimum value is zero. The maximum value of the contingency coefficient for the primary traits is .81 and .81 for the secondary traits. The contingency coefficient determined for each pair of scorers for each trait is shown on the Tables 2.1, 2.2 and 2.3. The contingency coefficient

TABLE 2.1

GRADE FOUR
Scorer Reliability:
Contingency Coefficient: First Scorer to Second Scorer

	1980			1979
Test Form	A	B	C	Last Year
Primary Trait				
Item 1	.75	.52	.63	.77
Item 2	.76	.66	.80	.80
Item 3	.61	.72	.72	.78
Syntax				
Item 1	.45	.60	.61	----
Item 2	.46	.55	.59	----
Item 3	.50	.65	.67	----
Spelling				
Item 1	.05	.36	.62	.64
Item 2	.57	.67	.69	.60
Item 3	.53	.72	.53	.59
Capitalization				
Item 1	.56	.56	.62	.60
Item 2	.74	.61	.78	.58
Item 3	.63	.50	.56	.60
Punctuation				
Item 1	.66	.69	.78	.62
Item 2	.70	.65	.78	.64
Item 3	.66	.60	.78	.61
Maximum Value				
Primary Trait	.81	.81	.81	.943
Secondary Trait	.81	.81	.81	.984

TABLE 2.2

GRADE EIGHT
Scorer Reliability:
Contingency Coefficient: First Scorer to Second Scorer

	1980			1979
Test Form	A	B	C	Last Year
Primary Trait				
Item 1	.45	.77	.79	.84
Item 2	.57	.82	.79	.81
Item 3	.61	.80	.68	.81
Syntax				
Item 1	.64	.67	.72	.60
Item 2	.62	.64	.82	.59
Item 3	.72	.80	.81	.47
Spelling				
Item 1	.44	.70	.65	.52
Item 2	.44	.61	.82	.52
Item 3	.71	.79	.79	.62
Capitalization				
Item 1	.40	.56	.45	.51
Item 2	.34	.46	.74	.51
Item 3	.66	.70	.76	.53
Punctuation				
Item 1	.62	.70	.79	.44
Item 2	.42	.65	.82	.48
Item 3	.67	.70	.77	.48
Maximum Value				
Primary Trait	.81	.81	.81	.943
Secondary Trait	.81	.81	.81	.894

TABLE 2.3

GRADE ELEVEN
Scorer Reliability:
Contingency Coefficient: First Scorer to Second Scorer

	1980			1979
Test Form	A	B	C	Last Year
Primary Trait				
Item 1	.56	.81	.75	.78
Item 2	.72	.80	.73	.78
Syntax				
Item 1	.57	.64	.67	.61
Item 2	.62	.78	.66	.57
Spelling				
Item 1	.53	.52	.45	.59
Item 2	.59	.75	.55	.52
Capitalization				
Item 1	.51	.40	.52	.55
Item 2	.41	.73	.49	.52
Punctuation				
Item 1	.29	.39	.51	.52
Item 2	.55	.73	.43	.51
Maximum Value				
Primary Trait	.81	.81	.81	.943
Secondary Trait	.81	.81	.81	.894

is reported for each test form field-tested. The last column of each table reflects the contingency coefficient for the 1979-1980 scorers.

Conclusions

Scorers at grades eight and eleven appear to have improved scorer agreement. The improvement is particularly noted in the secondary traits, especially at grade eight. Scorers at grade eight appeared to have excellent rapport with each other and to have a commitment to the achievement of scoring agreement.

The eighth-grade scorers came together throughout the day to review how they scored papers, to discuss points of disagreement, and to establish scoring guidelines. This discussion appears to contribute to scorer agreement. This is reflected in the fact that the contingency coefficients improved for each test form scored.

An increased agreement was not as noticeable at grades four and eleven. Several unexpected problems emerged during the scoring process that may have contributed to this. These problems are discussed below.

First of all, there was a "people" problem. The scoring process had been carefully planned and scorers were notified well in advance of the session. However, on the day scoring was to begin, a scorer at the fourth-grade level and at the eleventh-grade could not come. Two substitutes had to be secured at the last minute. The fact that the substitutes were not prepared to score and missed the training session contributed to the lack of scorer agreement for the teams on which they scored. Also, one scorer at the fourth-grade level tended to be more concerned with "keeping up" with the rest of the scorers rather than achieving agreement. The scorer became so distraught that she left before the task was completed. Another situation arose when an eleventh-grade scorer objected to scoring decisions which had been made concerning the relationship between handwriting and capitalization. It is essential that basic scoring guidelines be followed by all scorers. Another attitude that was noted for the eleventh-grade scorers wasthe feeling "do not worry about it - let the resolver take care of it." The attitude of the scoring team appears to be an important element in the achievement of scoring agreement.

The second problem was the amount of training. The primary purpose of this scoring session was to score the items which were field-tested for the spring assessment. All scorers had received training the previous summer or had served as scorers in the 1979-1980 State assessment. Therefore, only two hours were scheduled for retraining. This was not enough time. More time may have been needed because the scoring guides for both primary traits and secondary traits had been revised. However, it was apparent that a sufficient amount of time must be allocated to the retraining of scorers so that scorers will think in the same way. Discussions throughout the scoring process contributed to scorers thinking alike, also.

An analysis of the agreement of scorers at the fourth-grade level seemed to indicate a need for a more extensive training session at this

level. Scorers appear to need help in recognizing errors in syntax. This secondary trait was added to fourth-grade in an effort to improve scorer agreement by removing the confusion between syntax and spelling. Fourth-grade scorers may need guidelines that are more detailed. Certainly, trainers must devote time to a concentrated training session with this group.

The third factor which affected scorer agreement was the items themselves. All items and scoring guides had been developed the previous summer. One serious problem at eleventh-grade level was with a "persuasion item", item 1. The item objective had been confused by the test developers with a description task and the scoring guide tended to score for description rather than persuasion. Therefore, agreement was difficult to reach. A similar situation was noted at eighth-grade level with the third item of Form C. Students had difficulty fulfilling the task and responses were harder to grade. The item must stimulate the student to address the specified task and the scoring guide must clearly describe each possible score.

Opportunity for Students to Edit

At grades eight and eleven two forms of the test were formatted so as to provide an opportunity for students to edit the writing sample. Space was provided for the student to rewrite the sample. The investigators counted each test which included a rewritten writing sample. At grade eight, of the 100 tests providing the opportunity for rewriting a sample, thirty-three (33) included a rewritten sample. At grade eleven, twenty-seven of the 100 tests included a written sample.

Observations made by scorers indicate that students merely copied the original paragraph including all errors. In fact, errors on the rewritten sample were noted that were not observed on the original sample. Apparently, students did not utilize the opportunity to edit and rewrite to an advantage.

Time Required for Scoring

Scorers were scheduled to arrive at 1:00 p.m. on Thursday afternoon. Plans were to spend Thursday afternoon in retraining and Friday and Saturday from 8:00 to 4:30 in Scoring.

Retraining was conducted as scheduled with scoring beginning on Friday morning. Scoring moved slowly Friday morning. Obviously, the retraining session was not sufficient time for scorers to score in agreement. Much time was devoted to discussion. it was hoped that Form A could be scored by both teams in four hours. However, Form A was only scored one time at the eighth grade level. At the eleventh grade level and at the fourth grade level, scorers had scored about 50 booklets for the second round. However, by 2:00 p.m. all groups had completed Form A and had begun Form B with eighth grade being the last to complete Form A. By the end of Friday afternoon with all teams scoring to 5:30 and eighth grade scoring even later, 50 booklets of Form B had been scored twice.

Several scorers arrived early Saturday morning. Many felt pressured at being behind schedule. However, by 10:00 a.m. Form B was completed. At this time the principal investigators realized that 400 booklets could not be scored in two days. They recommended to the State Department Supervisor to delete one form from the required scoring. The Supervisor did agree, reluctantly, to delete Form D from scoring. This move tended to remove pressure from the scorers. Form C was completed by all groups by 5:30 p.m. with eighth-grade finishing last.

Conclusions

Although two teams of three scorers with one clerk and one resolver, scored 150 test booklets in one day, it was a supreme effort with extreme pressure. This number of scorers could, with ease, score 100 booklets.

Scoring is a tedious task requiring complete concentration. An individual can only attend such a task for a few hours and then he/she must break the concentration. Scorers cannot score with reliability for long hours.

After a time, agreement breaks down. If an agency is sincerely interested in accurate scoring, then consideration must be given to an adequate number of scorers and adequate time to accomplish the task.

As the eighth-grade group was under much pressure, a task analysis was conducted. It was noted that the eighth-grade test was composed of three items as compared to two items at eleventh-grade. In order to adjust scoring time, the investigators recommended that only two items be included on the eighth-grade test. This should balance the time required to score the eighth-grade test with the time required to score eleventh and fourth.

Summary

If the adjustments as recommended are made in scoring, it appears that two teams of three scorers, one resolver, and a clerk - a total of 8 people - can score efficiently 100 test booklets in eight hours. This would indicate that about 12.5 test booklets can be completely scored and resolved in an hour.

If for any reason the number of scorers per team should be reduced, "specialized scoring would in all probability be rendered ineffective. The only purpose of "Specialized Scoring" was to reduce the number of scoring guides to be considered by one individual. Scorers tended to like the procedure and indicated that it did make scoring easier.

If the number of scorers were reduced then the investigators recommend that the original method of scoring be utilized with one scorer scoring all traits.

Relationship Between the Various Trait Scores on the Objective Test and the Writing Sample

Items on the objective test measuring each domain of spelling, capitalization, punctuation, and language structure were determined. The number of items correct for each domain was determined for each student.

Next, each trait score for each item on the writing sample measuring that trait was summed for each student. This provided a trait score for each student on the primary traits and on capitalization, punctuation, spelling, and syntax, (except at grade 4). Each student's trait score was paired with his corresponding domain score on the objective test. A Pearson's r was run to determine the relationship between those two scores.

Findings

The following table shows the correlation coefficient for each trait-domain score.

THE RELATIONSHIP OF EACH DOMAIN ON THE WRITING SAMPLE TO A CORRESPONDING DOMAIN OBJECTIVE TEST

	4	8	11
Syntax		.37	.46
Spelling	.20	.27	.26
Capitalization	.17	.33	.29
Punctuation	.12	.19	.17
Total Test to Primary Trait	.50	.54	.42

Clearly, there was very little relationship between the trait scores and the corresponding domain scores as measured on the objective test. Apparently, the two tests are measuring two different functions. However, there appeared to be some relationship between the Primary Trait score received on a writing sample and the total score received on an objective test.

The next question which was addressed in this study concerned the information evaluated on the two tests. Did the Scoring Guides for the writing samples evaluate the same skills as measured on the objective test? Each objective measured on the objective test was determined and compared to the respective scoring guide. Findings are reported below.

Fourth Grade Test

Spelling. In spelling the objective test was structured to evaluate the student's ability to spell beginning consonant sounds, color words, numerals, and regular plurals. The writing sample was evaluated in terms of the high-frequency words used by the writer. One descriptive item stimulated the student to use color words. Other than these specifics, it was strictly by chance if the student used the other specified skills (plurals of nouns and number names)

Capitalization. The objective test contained items which measured the student's ability to capitalize proper nouns; the pronoun, I; and the beginning of the sentence. The writing sample tended to only measure the student's ability to capitalize the beginning of the sentence. One item stimulated the student to write in the first person and use the pronoun I. Unless a student tended to name the characters in a stimulus picture, the student was not stimulated to use proper nouns.

Punctuation. The objective test measured the student's ability to use end punctuation of period, exclamation point, and question mark. The writing sample only stimulated the student to use the period.

Eighth Grade and Eleventh Grade Test

Spelling. The objective test included items which evaluated the student's ability to spell the months, contractions, hyphenated compound numbers, plurals of nouns ending in s, past tense forms of verbs, and holidays. On the writing sample the students were responsible for specific spelling patterns. Only the words attempted by the student were evaluated on the writing sample. Some students wrote a lengthy sample which increased the probability that he would misspell a word. Other studetns wrote shorter samples, attempted fewer words, and made higher scores. There was a great variance in the words attempted by students on the writing samples.

Capitalization. When the items on the objective test were analyzed, the items designed to measure capitalization tended to measure conventions related to letter writing, writing title, and other specialized uses of capital letters. None of the writing tasks stimulated students to attempt the conventions evaluated on the objective test.

Punctuation. The objective tests evaluated the use of hyphens, the use of comma with items in a series, periods at the end of sentence and between dollars and cents. None of the writing tasks stimulated the use of these conventions.

Syntax. Findings relating to syntax paralleled the findings relating to capitalization and punctuation. Conventions which were evaluated on the objective tests were not stimulated on the writing task.

Conclusions. The basic conclusion which can be made is that there was very little relationship between scores students received on a writing sample and scores students received on an objective test. Obviously, from the findings reported, a student's ability to write a minimumally acceptable writing sample can not be predicted from a knowledge of the student's score on the Louisiana State Assessment of Writing (objective test). No indication of the lack of relationship can be made from the correlation study. A content analysis was made of each objective test and each corresponding scoring guide. The content on the two instruments were compared. Clearly, from the findings, the two instruments were not designed to measure the same things. Therefore, from this study the generalization can not be made as to whether or not a student's ability to write can be predicted from a score on an objective test of writing skills.

The Relationship of Mean Scores for Different Socio/Economic Levels of Students

An evaluation design was outlined in the proposal to answer the following three questions:

1. Are the mean scores on the objective test for each group of students in the design significantly different from the mean scores for every other group?

2. Are the mean scores for primary trait on the writing sample for each group of students in the design significantly different from the mean scores for every other group?

3. Is a profile analysis constructed for each group of students from data yielded from an evaluation of a writing sample similar to a profile analysis constructed for that group from data yielded from an objective test of writing?

Circumstances beyond the control of the investigators prevented the completion of the research to answer these questions. In order to address these three questions, a unique computer program was required which was not available in the university computer center.

Personnel changes in the computer center resulted in changes in responsibility so that the person with whom the investigators had been working was promoted. The new person was not familiar with the project nor with the multi-trait design. Although the investigators were successful in getting the necessary data loaded into the computer, a program could not be written by the personnel. Therefore, a programmer was needed. By this time it was late in the project year. Although budget categories could have been adjusted to accommodate this cost, the program could not have been developed and run by the closing date of the project. University and state policy mandates that all work be completed and billings filed in the Comptroller's office by the termination date of the project. Therefore, it was not possible to address these questions.

SURVEY OF NATIONAL TRENDS IN WRITING ASSESSMENT

One of the purposes of this study was to determine the extent to which large- scale writing assessment is being implemented in the nation.

Specifically, the following questions were to be answered:

(1) How many states and large city school systems are attempting to assess writing?

(2) If writing is being assessed, which measurement technique is being used: direct measurement, indirect measurements or both?

(3) If direct measurement is being used, what scoring method is used?

(4) How are assessment results reported and utilized?

(5) Do decision makers consider direct measures essential to the assessment of writing?

In order to answer these questions, the questionnaire shown on pages 43 and 44 were mailed to each state department of education and large school systems (See Appendix page 83 for selected school systems)

Findings and Results

Responses were received from 42 state departments of education with 24 of the SDE's indicating a writing assessment program. Table 3 on page 45 provides descriptive data about each of the writing programs. Of the 24 states claiming to have a writing assessment program, 22 claimed to assess writing using a writing sample. Most of the states using a writing sample indicated that they used holistic scoring procedures with three states using primary trait techniques, one state using analytical, and three states using both holistic and analytic scoring. States using only objective measures included only two.

Of the school systems responding, 20 school systems have a writing assessment program. Table 4 shown on page 47 summarizes the descriptive data for each program. Of the 20 school systems which indicated a writing assessment program, 17 use a writing sample. Methods of rating the sample tended to be more varied than those used by the SDE's. However, holistic scoring procedures were indicated to be used by seven systems with a combination of holistic and analytical used by three, and holistic and primary trait by one system. At least two systems claimed to use a writing scale and three systems claimed to use analytical scoring.

The data reveal that in most instances where a writing assessment has been implemented, the decision resulted from policy rather than by mandate. At the state level, 12 states are responding to a mandate with only three school systems responding to a mandate.

When asked if an objective test was used, eleven states indicated that an objective measure was used as well as a writing sample. In the eleven states not using an objective test, a writing sample was used. Only two states used an objective test without a writing sample. In cases of large school systems, twelve systems used an objective test. In eight systems where an objective was not used, a writing sample was used. Only two systems used an objective test without a writing sample.

One of the features set forth by W. James Popham as being characteristic of a high quality minimum competency program is that the program should assess defensible pre-determined competencies (Popham, 1981). The respondents were asked if minimum standards had been determined and if the writing sample measured the predetermined standards. As indicated on Table 3, fourteen states had minimum competencies and thirteen states indicated that the writing sample measured those competencies. Table 4 indicates that fifteen districts have minimum standards and eleven districts indicate that the writing sample measured predetermined standards. Minimum Standards were developed in a variety of ways, with the representative committee being the most frequently used process. (See Tables 5 and 6)

One of the major difficulties in utilizing the writing sample is managing the scoring of the large numbers of samples which result when the entire population is tested. On the other hand, if only a sample of the population is tested, there can be no direct feedback to the individual students. Respondents were asked if they tested an entire population or only a sample. Tables 7 and 8 report the results. Only ten states and fourteen LEA's test the entire population. Sampling size, in most cases, was less than 5,000 (13 SDE's and 5 LEA's).

Another major problem which has appeared in the implementation of writing assessment programs is the development of writing tasks which are used in the assessment to prompt students to write. Assessments utilizing a writing sample are relatively new and as a result very few writing tasks are available in item banks. This is not the case where multiple choice items for objective tests are concerned. NAEP has developed a bank of writing tasks because most agencies are faced with the problem of generating effective ones. Respondents were asked who developed the writing tasks in their respective samples. As state and local agencies indicated on Table 9 and 10, most of them utilized a committee composed of state department personnel, university personnel, teachers, and administrators (12 SDE's and 14 LEA's). Contractors were used by six SDE's and four LEA's) NAEP items were used by 3 State Departments.

When a student's ability to write is tested by means of an objective test, the majority of items are related to mechanics. (See Tables 11 and 12) Therefore, respondents were asked if mechanics were scored on the writing sample. As shown on Tables 9 and 10, fourteen SDE's and fourteen LEA's indicated that mechanics were evaluated. However, when the fact is considered that most of the agencies employ holistic scoring, it becomes apparent that mechanics is evaluated in the total impression of the writing. Louisiana has made an effort to evaluate mechanics by applying the minimum standards as a criteria. That is, to be judged as minimally competent, a student can not make an error in the use of a convention

identified for mastery at a given grade level. Other states scoring by the primary trait system, identify percentage of errors by the process of counting words written and errors made. Both processes are tedious and time consuming.

Since so much time, effort, and money is being expended in the assessment of writing, the agencies were asked how the information was reported and utilized. As shown on Table 13, most agencies report in terms of percentage correct or percentage of students demonstrating mastery. According to state agencies responding, the information is utilized by the LEA's and local schools (Table 14). On the other hand, LEA's report that the information is utilized mainly by the school and classroom. Only three LEA's report their local testing results to the state department and two LEA's report to parents. Apparently, the local education agency is expected to utilize the results.

1894
LOUISIANA TECH UNIVERSITY
Ruston, Louisiana 71272

Teacher Education *College of Education*

Dear Colleague:

The department of Teacher Education at Louisiana Tech University has been contracted by the National Institute of Education to research the "State of the Art" of large-scale assessment of written composition. As a part of that study, we must determine the status of large-scale writing assessment programs currently being used by both local and state education agencies. Therefore, we are asking you to complete this questionnaire and return it to us in the enclosed self-addressed envelope.

From the respondents to this questionnaire, ten individuals will be selected to participate in an expense-paid conference on writing assessment to be held in Louisiana in the summer of 1981. Invitations will be issued to others to attend at their own expense. Consultants at this conference will include Dr. William Lutz of Rutgers University, Dr. Ina Mullis of National Assessment of Educational Progress, and Dr. Allan Purvis of the University of Illinois-Urbana.

Your prompt response to this questionnaire would enable us to complete this important NIE assignment and would place your name among those to be considered for participation in the NIE conference on writing in the summer of 1981.

Sincerely,

Virginia L. Melton

Michael A. McCready

Please mark an X in the box beside each response which best describes the assessment program in your school system.

A. ASSESSMENT PROGRAM IN WRITTEN COMPOSITION

1 Does your local education agency have a systemwide assessment program in written composition?

☐ YES ☐ NO

***If you answered NO, do not complete the remainder of the questionnaire, but please return it to us.*

***If you answered YES, please continue.*

2. How was your assessment program in written composition initiated?

☐ mandated by state legislature
☐ result of policy set by state board of education
☐ result of action by local board of education.
☐ other _____

3 If your school system does have an assessment program in written composition, place an **X** in the box indicating the grade levels tested:

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
1	2	3	4	5	6	7	8	9	10	11	12

4. Does your system have a published set of minimum standards in writing?

☐ YES ☐ NO

5. How was the minimum standards document developed?

☐ by a committee of classroom teachers
☐ by a committee of school administrators
☐ by a committee of university personnel
☐ other _____
☐ by a committee representative of all of the above
☐ by state department personnel
☐ by a contractor

6. How is information from writing assessment utilized?

☐ reported to local districts
☐ reported to individual schools
☐ reported to individual classrooms
☐ reported to the SDE
☐ other _____

7. How are the results reported? (Respond to each section)

☐ by subject area tested
☐ by domain
☐ by objective
☐ by item

☐ for individual students
☐ for individual classrooms
☐ by grade level within each school
☐ by grade level within the school system
☐ by grade level within the state

☐ percentage correct
☐ percentile
☐ stanine
☐ standard score
☐ other _____

B. OBJECTIVE MEASURES OF WRITTEN COMPOSITION

8. Is an objective test used to measure writing ability?

☐ YES ☐ NO

***If you answered NO, do not complete this section, but go to section C.*

***If you answered YES, go to the next question.*

9. If an objective test is used, which of the following domains are measured at the designated grade levels?

PRIMARY	MIDDLE	JUNIOR HIGH	SENIOR HIGH
☐ Content	☐ Content	☐ Content	☐ Content
☐ Organization	☐ Organization	☐ Organization	☐ Organization
☐ Style	☐ Style	☐ Style	☐ Style
☐ Spelling	☐ Spelling	☐ Spelling	☐ Spelling
☐ Punctuation	☐ Punctuation	☐ Punctuation	☐ Punctuation
☐ Capitalization	☐ Capitalization	☐ Capitalization	☐ Capitalization
☐ Syntax	☐ Syntax	☐ Syntax	☐ Syntax

10. If an objective test is used, do items measure specified minimum standards?

☐ YES ☐ NO

C. WRITING SAMPLE AS MEASURES OF WRITTEN COMPOSITION

11. Is a writing sample used to measure writing ability?

☐ YES ☐ NO

***If you answered NO, do not complete this section, but go to question 23.*

***If you answered YES, go to the next question.*

12. Does the writing sample measure specified minimum standards?

☐ YES ☐ NO

13. Who developed the items for the writing sample?

☐ State Department Personnel
☐ University Personnel
☐ Private Contractor
☐ Classroom Teachers
☐ LEA Administrators
☐ Other_____

14. If a writing sample is utilized is the test administered to:

☐ entire population tested ☐ sample of population

15. How many students are tested at each grade level with a writing sample?

PRIMARY	MIDDLE	JUNIOR HIGH	SENIOR HIGH
☐ less than 5,000	☐ less than 5,000	☐ less than 5,000	☐ less than 5,000
☐ 5,000 – 10,000	☐ 5,000 – 10,000	☐ 5,000 – 10,000	☐ 5,000 – 10,000
☐ 10,000 – 20,000	☐ 10,000 – 20,000	☐ 10,000 – 20,000	☐ 10,000 – 20,000
☐ 20,000 – 30,000	☐ 20,000 – 30,000	☐ 20,000 – 30,000	☐ 20,000 – 30,000
☐ 30,000 – 40,000	☐ 30,000 – 40,000	☐ 30,000 – 40,000	☐ 30,000 – 40,000
☐ 40,000 – 60,000	☐ 40,000 – 60,000	☐ 40,000 – 60,000	☐ 40,000 – 60,000
☐ more than 60,000	☐ more than 60,000	☐ more than 60,000	☐ more than 60,000

16. What types of writing are measured with a writing sample?

☐ narration
☐ exposition
☐ description
☐ persuasion
☐ other _____

17. Is the writing sample evaluated for mechanics? ☐ YES ☐ NO

If YES, which of the following mechanics are evaluated?

☐ spelling ☐ punctuation ☐ usage
☐ capitalization ☐ language structure ☐ handwriting

18. Are scorers trained to score the writing sample with a high degree of scorer agreement?

☐ YES ☐ NO

19. Statistically, how is scorer agreement determined? _____

20. Who scores the writing sample?

☐ classroom teachers
☐ graduate students
☐ test contractor
☐ state department employees
☐ other _____

21. Who trains the scorers?

☐ State department personnel
☐ test contractors
☐ consultants
☐ other _____

22. What system of scoring is used?

☐ holistic
☐ analytical
☐ writing scales (specify) _____
☐ primary trait
☐ other _____

23 Two schools of thought appear to be emerging concerning the assessment of writing. One school insists that the only true measure of writing skill is to have students write. The other school insists that most important mental processes, including writing, can be measured well by objective items. Explain why you feel that a writing sample either is or is not essential in the assessment of writing.

Table 3
States With a Writing Assessment

State	Action Initiating Assessment	Grades Tested	Do States Have Minimum Standards?	Is an Objective Test Used?	Does the Test Measure Minimum Standards?	Is a Writing Sample Used?	Does the Writing Sample Measure Minimum Standards?	Method of Scoring
AL	Policy	3,6,9	Yes	No	No	Yes	Yes	Holistic
DE	Mand./Policy	1-8,11	Yes	Yes	No	Yes	No	Primary Trait
CA	Mandated	3,6,12	No	Yes	No	No	No	
FL	Mandated	3,5,8,11	Yes	Yes	Yes	Yes	Yes	Analytical
HI	*Ad. Decision	4,8,11	Yes	No	No	Yes	Yes	Holistic
ID	Policy	9	Yes	No	No	Yes	Yes	Holistic
ME	Mand./Policy	4,8,11	No	No	No	Yes	Yes	Holistic
MD	Policy	9-12	Yes	Yes	Yes	Yes	Yes	Holistic
MA	Policy	7,8,9,12	Obj.	No	No	Yes	Yes	Hol./Analytic
MI	Mand./Policy	4,7,10	Yes	I.D.		Yes	Yes	
MN	*Ad. Decision	4,8,11	No	No	No	Yes	No	Primary Trait
NV	Mandated	3,6,9-12	Yes	3-6	Yes	Yes	Yes	Holistic
NC	Rec.	11	Yes	No	No	Yes	Yes	Hol./Analytic
NH	Policy	5,9,12	No	No	No	Yes	No	Holistic
NJ	Mandated	9	Yes	Yes	Yes	Yes	Yes	Holistic
NM	?	10	In process	No	No	Yes	No	Holistic

TABLE 3 (Cont.)
States With a Writing Assessment

State	Action Initiating Assessment	Grades Tested	Do States Have Minimum Standards?	Is an Objective Test Used?	Does the Test Measure Minimum Standards?	Is a Writing Sample Used?	Does the Writing Sample Measure Minimum Standards?	Method of Scoring
OH	Mandated	8,12	Yes	Yes	Yes	Yes	No	Holistic
OR	*Ad. Decision	4,7,11	Yes	Yes	No	Yes	No	Holistic
PA	Mandated	5,8,11	No	Yes	No	No	No	----
RI	Policy	4,6,8,10	In Process	Yes	Yes	Yes	Yes	Holistic
SC	Mandated	6,8,11	Yes	No	No	Yes	Yes	Hol. Analyt
TX	Mandated	3,5,9	Yes	Yes	Yes	Yes	Yes	Holistic
WY	*Ad. Decision	6,9	No	No	No	Yes	No	Holistic
LA	Mandated	3,7,10	Yes	Yes	Yes	Yes	Yes	Primary Trait

*Ad. - Administrative

TABLE 4
School Systems With a Writing Assessment

School District	Action Initiating Assessment	Grades Tested	Do Districts Have Minimum Standards?	Is an Objective Test Used?	Does the Test Measure Minimum Standards?	Is a Writing Sample Used?	Does the Writing Sample Measure Minimum Standards?	Method of Scoring
AR, Little Rock	Policy	1-11	No	Yes	No	No		
AZ, Phoenix	*Ad. Decision	9-12	Yes	Yes	Yes	Yes	Yes	Analytical
CA, Monterey	Policy	1-12	Yes	No	No	Yes	Yes	Holistic
FL, Tallahassee	Policy	1-8	Yes	Yes	Yes	Yes	Yes	
GA, Atlanta	Policy	1-12	Yes	Yes	Yes	No	No	
IA, Des Moines	*Ad. Decision	9	No	No	No	Yes	No	Hol./Analy.
IL, Chicago	Policy	9-12	Yes	No	No	Yes	Yes	
KS, Wichita	*Ad. Decision	K-12	Yes	Yes	No	Yes	No	Holistic
MA, Boston	Mandated	2,5,8	Yes	No	No	Yes	Yes	Holistic
MD, Baltimore	Policy	1-9	Yes	Yes	Yes	Yes	Yes	Analytical

*Ad. dec. - Administrative decision

TABLE 4 (Cont.)
School Systems With a Writing Assessment

School District	Action Initiating Assessment	Grades Tested	Do Districts Have Minimum Standards?	Is an Objective Test Used?	Does the Test Measure Minimum Standards?	Is a Writing Sample Used?	Does the Writing Sample Measure Minimum Standards?	Method of Scoring
MI, Detroit	Policy	10-12	Yes	Yes	Yes	Yes	Yes	Holistic
NC, Raleigh	*Ad. Decision	1-12	Yes	Yes	Yes	Yes	Yes	teacher judgment
NM, Albuquerque	*Ad. Decision	10-12	Yes	Yes - 4,6,9 No - 10	No	Yes	Yes	
NM, Santa Fe	Policy	7-12	No	No	No	Yes	No	Hol./Analy.
NY, New York	Policy	8,11	Yes	No	No	Yes	No	Holistic
OR, Portland		3-9	Yes	Yes	Yes	No	No	
TX, Austin	Mandated	3,9	Yes	Yes	Yes	Yes	Yes	Holistic
WI, Madison	Mandated	5,8,11	No	Yes	No	Yes	No	Hol./PT
WA, Seattle	Policy	3,6,9, 11	Yes	No	No	Yes	Yes	W. Scale
WY, Laramie		6,9	No	No	No	Yes	No	Holistic

*Ad. dec. - Administrative decision
Hol./PT - Holistic/Primary Trait

TABLE 5
The Development of Minimum Standards in Writing in States

State	How was the minimum standards document developed?							How is inform. from writing assessment utilized?				
	Clsrm. Teacher	Sch. Adm.	Uni. Pers.	SDE	Contract	Com. Rep.	Other	LEA	Sch.	Clsrm.	SDE	Other
AL				X				X	X	X		
CA							Dist. dev. own	X	X			SDE & Legislature
HI								X				
ID				X		X		X	X			
LA				X				X	X	X	X	
ME							Com. Review	X				NAEP Studies
MA								X	X			
MI						X						Under development
MN												Statewide Reporting
NH											X	Model by Local Dist.
NV	X	X	X			X	Com. rep. & sp. cons.	X	X	X		Students & parents
OH						X		X	X		X	Education org.
OR								X	X	X		Reported by SDE to leg. & media
RI	X	X	X	X	X	X					X	S.D. of Regents
SC						X	Dist. level cirr. sp.	X	X	X	X	
WY								X	X			Returned to school
NJ						X		X	X	X		Students

TABLE 6
The Development of Miminum Standards in Writing in School Systems

School System	How was the minimum standards document developed?							How is inform. from writing assessment utilized?				
	Clsrm. Teacher	Sch. Adm.	Uni. Pers.	SDE	Contract	Com. Rep.	Other	LEA	Sch.	Clsrm.	SDE	Other
AR, Little Rock								X	X	X		
AZ, Phoenix						X			X	X		
CA, Monterey	X								X	X		
FL, Tallahassee						X		X	X	X		
GA, Atlanta							Lang. arts sup. & Clsrm T.		X	X		
IA, Des Moines									X	X		
IL, Chicago	X	X						X				
OH, Cincinnati						X	Parents		X	X		Parents
MD, Baltimore						X			X	X		
MA, Boston							com. of t., ad., pub., & pers.	X	X	X		
MI, Detroit	X	X					Cen. Staff & Sup.					

TABLE 6 (Cont.)
The Development of Minimum Standards in Writing in School Systems

School System	How was the minimum standards document developed?							How is inform. from writing assessment utilized?				
	Clsrm. Teacher	Sch. Adm.	Uni. Pers.	SDE	Contract	Com. Rep.	Other	LEA	Sch.	Clsrm.	SDE	Other
NC, Raleigh	X	X					Coll. of Research ideas		X	X		Parents/students
NM, Santa Fe									X			
NY, New York	X			X		X		X	X	X	X	
TX, Austin				X		X		X	X	X	X	
WA, Seattle	X						Committee		X			
WY, Laramie								X	X		X	

Explanation of Abbreviations

How was the minimum standards document developed?
Clsrm. - Classroom
Sch. Adm. - School Administrators
Uni. Pers. - University Personnel
Com. Rep. - Committee Representative

How is information from writing assessment utilized?
Sch. - School
Clsrm. - Classroom

TABLE 7
Writing Sample as Measures of Written Composition in States

State	Is it used?	M.S.	WHO IS TESTED?							
			<5,000	5,000–10,000	10,000–20,000	20,000–30,000	30,000–40,000	40,000–60,000	>60,000	Entire Pop.
AL	Yes	Yes						P,M,J		Yes
CA	Yes'75	No	S							No
HI	Yes	Yes	M,J,S							No
ID	Yes	Yes			J,S					Yes
LA	Yes	Yes	P,J,S							No
ME	Yes				S					
MA	Yes	Yes							J	Yes
MI	Yes	Yes	P,M,J							No
MN	Yes	No	M,J,S							No
NH	Yes	No	M,J,S							No
NJ	Yes	Yes							S	Yes
NM	Yes	Yes								
NV	Yes	Yes		S						Yes
OH	Yes	No	J,S							No
OR	Yes	No	M,J,S							No
RI	Yes	Yes	P,M,J,S							Yes
SC	Yes	Yes							M,J,S	Yes
NY	Yes	No	M,J							Yes

TABLE 7 (Cont.)
Writing Sample as Measures of Written Composition

State	Is it used?	M.S.	WHO IS TESTED? <5,000	5,000-10,000	10,000-20,000	20,000-30,000	30,000-40,000	40,000-60,000	>60,000	Entire Pop.
MD	Yes	Yes							S	Yes
FL	Yes	Yes	P,M,J,S							No
DE	Yes	No	J							No
PA	No									
TX	Yes	Yes							P,M,S	Yes
NC	Yes	Yes	S							No

Explanation of Abbreviations

M.S. - Minimum Standards
P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 8
Writing Sample as Measures of Written Composition in School Systems

School System	Is it used?	M.S.	WHO IS TESTED?							
			<5,000	5,000-10,000	10,000-20,000	20,000-30,000	30,000-40,000	40,000-60,000	>60,000	Entire Pop.
AR, Little Rock										
AZ, Phoenix	Yes	Yes		S						Yes
CA, Monterey	Yes	Yes	P,M,J,S							Yes
FL, Tallahassee	Yes	Yes		M,J						Yes
GA, Atlanta	No									
IA, Des Moines	Yes	No	J							Yes
IL, Chicago	Yes	Yes							S	
OH, Cincinnati	Yes	Yes	P,M,J,S							Yes
MD, Baltimore	Yes	Yes	M			S		J	P	Yes
MA, Boston	Yes	Yes		P,M,J,S						Yes
MI, Detroit	Yes	Yes			S					Yes

TABLE 8 (Cont.)
Writing Sample as Measures of Written Composition in School Systems

School System	Is it used?	M.S.	WHO IS TESTED? <5,000	5,000-10,000	10,000-20,000	20,000-30,000	30,000-40,000	40,000-60,000	>60,000	Entire Pop.
NC, Raleigh	Yes	Yes	P,M,J,S							No
NM, Santa Fe	Yes	No	J,S							Yes
NY, New York	Yes	No							J,S	Yes
TX, Austin	Yes	Yes	P,M,S							Yes
WA, Seattle	Yes	Yes	P,M,J,S							Yes
WY, Laramie	Yes	No	M,J							

Explanation of Abbreviations

P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 8 (Cont.)
Writing Sample as Measures of Written Composition in School Systems

School System	Is it used?	M.S.	WHO IS TESTED? <5,000	5,000-10,000	10,000-20,000	20,000-30,000	30,000-40,000	40,000-60,000	>60,000	Entire Pop.
WI, Madison	Yes	No	M,J,S							No
OR, Portland	No									
NM, Albuquerque	Yes	Yes		P,M,J,S						Yes
KS, Wichita	Yes	No			P,M,J,S					Yes

Explanation of Abbreviations

M.S. - Minimum Standards
P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 9
Key Participants
Development of the Writing Task in State Assessment Programs

State	Who developed the items for the writing sample?						Type					Eval.	Eval. Mechanics					
	SDE	Uni. Pers.	Con.	Teacher	LEA	Other	N.	E.	D.	P.	O.		S.	C.	P.	L.	U.	H.
AL	X	X		X	X	Committee	X					Yes	X	X	X	X	X	X
CA	X	X		X	X	Committee		X	X			Yes	X	X	X	X	X	X
HI						Spec. Task	X	X	X	X		Yes	X	X	X	X	X	
ID	X						X	X	X	X		No						
LA				X		Committee	X	X	X	X		Yes	X	X	X	X	X	X
ME						NAEP	X	X	X			No						
MA	X			X		Commitee			X	X		No						
MI	X	X				NAEP	Xud	Xud	Xud			Yes						
MN	X	X		X	X	Committee	X	X	X	X	X	Yes	X	X	X	X	X	
NH	X						X78	X80				Yes	X	X	X	X	X	
NJ			X				X					No						
NM	X			X	X	Committee			X	X	B.L.							
NV	X			X		Committee		X	X	X	B.L.	Yes	X	X	X	X	X	X
OH	X	X		X		Committee	X	X	X	X		Yes	X	X	X	X	X	X
OR	X						X	X	X	X		No						
RI			X				X			X		Yes	X		X	X		
SC	X	X	X	X	X	Rev. Group	X	X	X	X		Yes	X	X	X	X	X	X
WY	X	X		X			X		X			Yes	X	X	X	X	X	X

nder development

TABLE 9 (Cont.)
Development of the Writing Task in State Assessment Programs

State	Who developed the items for the writing sample?						Type					Eval.	Eval. Mechanics					
	SDE	Uni. Pers.	Con.	Teacher	LEA	Other	N.	E.	D.	P.	O.		S.	C.	P.	L.	U.	H.
MD	X		X	X	X		X	X	X		X	Yes			X		X	
FL						State, Clsrm teacher & Pri. Con.					Spec Task	Yes	X	X	X	X	X	X
DE	X				X	NAEP	X			X		Yes	X	X	X	X	X	
TX	X	X	X	X			X		X	X		No						
NC			X									Yes	X	X	X	X	X	

Explanation of Abbreviations

Who developed the items for the writing sample?
Uni. Pers. - University Personnel
Con. - Contractor

Type
N. - Narration
E. - Exposition
D. - Description
P. - Persuasion
O. - Other (B.L. - Business Letter, Mess. - Message)

Evaluation Mechanics
S. - Spelling
C. - Capitalization
P. - Punctuation
L. - Language Structure
U. - Usage
H. - Handwriting

TABLE 10
Development of Writing Tasks in School Systems

School System	Who developed the items for the writing sample?						Type					Eval.	Eval. Mechanics					
	SDE	Uni. Pers.	Con.	Teacher	LEA	Other	N.	E.	D.	P.	O.		S.	C.	P.	L.	U.	H.
AR, Little Rock																		
AZ, Phoenix				X	X		X	X	X			Yes	X	X	X	X	X	X
CA, Monterey				X				X	X			Yes	X	X	X	X	X	X
FL, Tallahassee		X		X	X		X	X	X			Yes	X	X	X	X	X	X
IA, Des Moines				X	X					X		Yes	X	X	X	X	X	X
IL, Chicago				X			X	X	X	X	Pro/ Supp	Yes	X	X	X	X	X	X
OH, Cincinnati				X		Super./ Review P.					X	Yes	X	X	X	X	X	X
MD, Baltimore				X	X			X		X		Yes	X	X	X	X	X	X
MA, Boston	X			X			X	X	X			Yes/H						
MI, Detroit			X			Dept./L.A.		X				Yes	X	X	X	X	X	X
NC, Raleigh			X	X	X		X	X	X			Yes	X	X	X	X	X	X
NM, Santa Fe				X					D		Mess B.L.							

TABLE 10 (Cont.)
Development of Writing Tasks in School Systems

School System	Who developed the items for the writing sample?						Type					Eval.	Eval. Mechanics					
	SDE	Uni. Pers.	Con.	Teacher	LEA	Other	N.	E.	D.	P.	O.		S.	C.	P.	L.	U.	H.
NY, New York	X							X		X	B.L.							
TX, Austin	X	X	X	X			X	X		X	Var.	Yes	X	X	X	X	X	X
WA, Seattle						Clsrm T./ Curr. Sp.					Let.	Yes	X	X	X	X	X	X
WY, Laramie						Com. of Local and State Uni. members		X	X									

Explanation of Abbreviations

Who developed the items for the writing sample?
Uni. Pers. - University Personnel
Con. - Contractor

Type
N. - Narration
E. - Exposition
D. - Description
P. - Persuasion
O. - Other (B.L. - Business Letter, Mess. - Message)

Evaluation Mechanics
S. - Spelling
C. - Capitalization
P. - Punctuation
L. - Language Structure
U. - Usage
H. - Handwriting

TABLE 10 (Cont.)
Development of Writing Tasks in School Systems

School System	Who developed the items for the writing sample?						Type					Eval.	Eval. Mechanics					
	SDE	Uni. Pers.	Con.	Teacher	LEA	Other	N.	E.	D.	P.	O.		S.	C.	P.	L.	U.	H.
WI, Madison	X	X		X		Parent/Bus. People	X	X		X	Bus. Voc.	No						
NM, Albuquerque	X			X				X	X	X		Yes	X	X	X	X	X	X
KS, Wichita				X		Coord. of Lang. Arts		X				Yes	X	X	X	X	X	X

Explanation of Abbreviations

Who developed the items for the writing sample?
Uni. Pers. - University Personnel
Con. - Contractor

Type
N. - Narration
E. - Exposition
D. - Description
P. - Persuasion
O. - Other (B.L. - Business Letter, Mess. - Message)

Evaluation Mechanics
S. - Spelling
C. - Capitalization
P. - Punctuation
L. - Language Structure
U. - Usage
H. - Handwriting

TABLE 11
Objective Testing in State Assessment Programs

State	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
AL	No								
CA	Yes	No	M	M	M,S	P,M,S	P,M,S	P,M,S	P,M,S
HI	No								
ID	No								
LA	Yes	Yes	P,J,S	P,J,S	P,J,S	P,J,S	P,J,S	P,J,S	P,J,S
ME									
MA	No								
MI	U.D.								
MN	No								
NH	No								
NJ	Yes	Yes		S		S	S	S	
NM	No								
NV	Yes3,6 No9-12	No No				P,M	P,M	P,M	
OH	Yes	Yes				J,S	J,S	J,S	
OR	Yes	No		M,J		M,J	M,J	M,J	
RI	Yes	No				P,M	P,M	P,M	
SC	No								
WY	No								

TABLE 11 (Cont.)
Objective Testing in State Assessment Programs

State	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
NC	No								
TX	Yes	Yes				P,M,S	P,M,S	P,M,S	P,M,S
PA	Yes	No	M,J,S	M,J,S	M,J,S		M,J,S	M,J,S	M,J,S
DE	Yes	No				P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
FL	Yes	Yes		P,M,J,S		P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
MD	Yes	Yes	S	S					

Explanation of Abbreviations

M.S. - Minimum Standards

P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 12
Objective Testing in School Systems

School System	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
AR, Little Rock	Yes	No				P,M,J,S	P,M,J,S	P,M,J,S	
AZ, Phoenix	Yes	Yes	S	S		S	S	S	S
CA, Monterey	No								
FL, Tallahassee	Yes	Yes	J	M,J	P,M,J	P,M,J	P,M,J	P,M,J	
GA, Atlanta	Yes	Yes		M,J,S		P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
IA, Des Moines	No								
IL, Chicago	No								
OH Cincinnati	Yes	Yes		P,M	M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
MD, Baltimore	Yes	Yes				P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
MA, Boston	No								
MI, Detroit	Yes	Yes				S	S	S	S
NC, Raleigh	Yes CRT					P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S

TABLE 12 (Cont.)
Objective Testing in School Systems

School System	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
NM, Santa Fe	No								
NY, New York	No								
TX, Austin	Yes	Yes	P,M,S	P,M,S		P,M,S	P,M,S	P,M,S	P,M,S
WA, Seattle	No								
WY, Laramie	No								

Explanation of Abbreviations

M.S. - Minimum Standards

P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 13
Reporting of Writing Assessment Results by State Agencies

How are results reported?

State	Subj.	Domain	Obj.	Item	Student	Clsrm	Grade Sch.	Grade LEA	Grade SDE	Perc. Co.	Percent	Stanine	Sta. Score	Other
AL			X		X					X				
CA		X					X			X	X	X		Latent Trait Theory
HI			X					X						Percentage 4-1
ID	X				X		X	X	X	X				
LA	X	X	X	X	X	X	X	X	X	X				
ME			X						X	X				
MA	X				X		X	X	X					Holistic
MI			Xud.	Xud.										
MN				X					X	X				Other analyses
NH	X		X						X					Des. of high of low papers
NJ	X		X	X	X	X	X	X	X	X			X	
NM					X									Pass/Fail
NV	X	X	X	X	X		X	X	X	X	X	X	X	Holistic
OH	X		X	X					X					% of student & categories
OR		X	X	X	X	X	X		X	X				Writing ex. 1-4
RI			X	X					X	X8,10	X4,6			
SC	X		X		X	X	X	X	X	X				Raw Score
WY		X			X			X	X					

TABLE 13 (Cont.)
Reporting of Writing Assessment Results by State Agencies

How are results reported?

School System	Subj.	Domain	Obj.	Item	Student	Clsrm	Grade Sch.	Grade LEA	Grade SDE	Perc. Co.	Percent	Stanine	Sta. Score	Other
MD		X			X					X			X	Holistic
FL			X						X					% mastery of standards
DE	X	X	X	X	X	X	X	X	X		X		X	
PA	X			X			X	X	X	X	X			
TX	X		X		X		X	X	X					% mastery
NC		X	X		X		X	X	X					undecided

Explanation of Abbreviations

Subj. - Subject area
Obj. - Objective
Clsrm - Classroom
Perc. Co. - Percentage correct
Sta. score - Standard score

TABLE 14
Reporting of Writing Assessment Results in School Systems

How are results reported?

School System	Subj.	Domain	Obj.	Item	Student	Clsrm	Grade Sch.	Grade LEA	Grade SDE	Perc. Co.	Percent	Stanine	Sta. Score	Other
AR, Little Rock	X	X			X		X	X		X	X	X		Growth Scale
AZ, Phoenix			X		X	X	X	X		X				No of skills passed
CA, Monterey		X			X		X	X						Student rating pro by comp. domain
FL, Tallahassee	X	X	X		X	X	X	X						% ach. of objective
GA, Atlanta	X		X		X	X	X	X		X				
IA, Des Moines			X	X	X	X				X				Item
IL, Chicago	X				X					X				
OH, Cincinnati	X				X	X	X			X				
MD, Baltimore			X	X	X	X	X			X				
MA, Boston	X				X									Holistic
MI, Detroit	X		X		X		X	X		X				

TABLE 14 (Cont.)
Reporting of Writing Assessment Results in School Systems

How are results reported?

School System	Subj.	Domain	Obj.	Item	Student	Clsrm	Grade Sch.	Grade LEA	Grade SDE	Perc. Co.	Percent	Stanine	Sta. Score	Other
NC, Raleigh			X		X									Objective mastered
NM, Santa Fe	X							X						Pass/Fail
NY, New York					X		X			X				
TX, Austin	X		X	X	X			X	X	X			X	
WA, Seattle				X	X									Satisfactory/Unsa
WY, Laramie	X				X			X						Holistic

TABLE 14 (Cont.)
Objective Testing in School Systems

School System	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
NM, Santa Fe	No								
NY, New York	No								
TX, Austin	Yes	Yes	P,M,S	P,M,S		P,M,S	P,M,S	P,M,S	P,M,S
WA, Seattle	No								
WY, Laramie	No								

Explanation of Abbreviations

M.S. - Minimum Standards

P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

TABLE 14 (Cont.)
Objective Testing in State Assessment Programs

School System	Obj. Test	M.S.	If an objective test is used, which of the following domains are measured at the designated grade levels? Content	Organization	Style	Spelling	Punctuation	Capitalization	Syntax
KS, Wichita	Yes	No				P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
NM, Albuquerque	Yes 4,6,9 No 10		P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
OR, Portland	Yes	Yes		P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S	P,M,J,S
WI, Madison	Yes	No	P,J,S			P,J,S	P,J,S	P,J,S	

71

Explanation of Abbreviations

M.S. - Minimum Standards

P. - Primary
M. - Middle
J. - Junior High
S. - Senior High

Summary of National Conference on Writing Assessment

In order to conduct a deeper probe into the nature of large-scale assessment in the Nation, the investigators planned a conference to which individuals who were involved in large assessment programs were invited. The responses were classified according to the assessment programs which are described below:

1. States and Local Agencies that use a Writing Sample

   a. Holistic Scoring
   b. Primary Trait
   c. Analytical

2. States and Local Agencies that do not use a Writing Sample

After the responses were classified a random sample was drawn which was representative of the description of writing assessment programs as indicated from the survey. A representative from each program identified in the random selection was invited to attend a Writing Assessment Conference in New Orleans on July 10. Each representative accepted the invitation. Listed below are the systems which were represented at the conference. After the name of the system is a description of the type of assessment program used in that system.

Delaware	Primary Trait
Pennsylvania	No Writing Sample
North Carolina	Holistic/Analytical
Florida	Analytical
Texas	Focused Holistic
Maryland	In process of developing the program
Madison, Wisconsin	Holistic
Wichita, Kansas	Holistic
Portland, Oregon	No Writing Sample
Albuquerque, Nex Mexico	Holistic

In addition to the identified participants other persons attended representing the states of Arkansas, Louisiana, Texas and South Carolina. A complete listing is provided in the Appendix.

The conference was planned using a discussion format. No planned speeches were presented. The investigators led the discussion following the questions indicated on the agenda provided on the next pages. Following the agenda is a summary of the ensuing discussion.

A G E N D A

NIE - LOUISIANA TECH WRITING CONFERENCE

FEASIBILITY OF ASSESSING WRITING USING MULTIPLE TECHNIQUES

Hotel Marie Antoinette
New Orleans, Louisiana
July 10, 1981

8:30 a.m. - Conference Overview

This conference is planned to facilitate discussion. There are no planned speeches. The following questions have been presented for consideration.

1. What are the purposes of a large-scale writing assessment?

2. What information is obtained by a direct measure of writing that is not available from indirect measures?

3. How are writing samples scored efficiently and reliably?
   a. What are the advantages of the various methods of assessing written composition (holistic, primary trait, and analytical)?
      1. What information is yielded by each scoring technique?
      2. How is the information reported and utilized?
      3. Can this information be utilized to improve instruction? How?
      4. How does this information compare with information yielded by indirect measures?
   b. How is mechanics scored?
   c. How are writing tasks developed?
      1. What kinds of writing tasks are developed?
      2. How many tasks are needed to secure a reliable and valid sample of writing?
      3. How are tasks developed?
      4. What types of prompts elicit the best responses?
      5. How much time is allotted for students to write?

4. The Scoring Process
   a. How are scorers selected and trained?
   b. Describe the scoring process.
   c. How is scorer reliability maintained?
   d. What is time involved in scoring using each of the techniques?
   e. What is the cost involved in scoring using each of the techniques?

5. How is the scoring of large numbers of writing samples managed?
   a. What management plans have been used successfully to score the entire population?
   b. What sampling techniques yield the best results in describing the status of writing in a given population?
   c. How does an agency maintain reference point from year to year when scoring writing samples?

Conclusions

Considering time, cost, and information yielded, is the scoring of a writing sample in large-scale assessment cost effective?

How does an agency make the cost/benefit decision on whether to test writing by direct or indirect measures?

How can both approaches be integrated?

Discussion Summary

This section presents a summary of discussion stimulated by each question on the agenda.

1. *What are the purposes of a large-scale writing assessment?*
   a. *Who are audiences?*

Practitioners agreed that the major purposes of large-scale writing assessment are four-fold as follows:

1. improvement of instruction;
2. assurance that each child has acquired the basic skills in writing;
3. determination of the status of writing in a given state or system;
4. provision of quality education.

2. *What information is obtained by a direct measure of writing that is not available from indirect measures?*

This question generated a great deal of discussion which included a number of opinions. There appeared to be a concensus that certain qualities of writing ability can be measured through the use of objective tests. Research studies were cited which demonstrate a high correlation between direct and indirect measures. Lutz indicated that a multiple-choice item can be constructed to measure a given writing skill; however, a problem inherent in the construction of the item is the establishment of the parameters of the item. Apparently, this is an area of needed research.

Systems utilizing direct measurement indicated that while their assessment programs were mandated, the decision to use direct measures of writing resulted from recommendations by professionals. There seemed to be a concensus among participants that certain qualities can be measured only by use of a writing sample. Purves asserted that the general public does not tend to accept the results of objective tests as an indication of writing ability. One participant indicated that a writing sample offers more "legal evidence" of writing ability than do objective tests. On the other hand, a participant from a large city school system argued that for an indirect measure to be "worth anything" the results must get back to the school level. The cost of scoring a writing sample for every student is prohibitive. Therefore, the school system elected not to make use of a writing sample rather than resort to the use of a random sample.

The participants agreed that the ideal situation would be one in which every paper is scored. However, time and cost prohibit the scoring of a writing sample for an entire population of students. Most participants tended to agree that although the information yielded by a writing sample could only be generalized to a population, the scoring of writing from a sample of the population is essential to the writing assessment program

primarily because of the instructional message to teachers. If teachers know that students' writing ability will be evaluated by means of direct measure, teachers will require more writing production from students.

3. *How are writing samples scored efficiently and reliably?*
   a. *What are the advantages of the various methods of assessing written composition (holistic, primary trait, and analytical)?*
      1. *What information is yielded by each scoring technique?*
      2. *How is the information reported and utilized?*
      3. *Can this information be utilized to improve instruction? How?*
      4. *How does this information compare with information yielded by indirect measures?*
   b. *How is mechanics scored?*
   c. *How are writing tasks developed?*
      1. *What kinds of writing tasks are developed?*
      2. *How many tasks are needed to secure a reliable and valid sample of writing?*
      3. *How are tasks developed?*
      4. *What types of prompts elicit the best responses?*
      5. *How much time is allotted for students to write?*

4. *The Scoring Process*
   a. *How are scorers selected and trained?*
   b. *Describe the scoring process.*
   c. *How is scorer reliability maintained?*
   d. *What is time involved in scoring using each of the techniques?*
   e. *What is the cost involved in scoring using each of the techniques?*

To introduce the discussion of this question, each of the three consultants reviewed the theory and rationale underlying each of the three major scoring techniques as well as the procedures and guidelines to be followed in implementing each system. Following the discussion of the three consultants, various participants discussed the scoring procedures as implemented in his-her system. A summary of that discussion follows:

A. One state representive reported a scoring technique identified as "focused-holistic." The discourse model used included the following modes of writing: informational, persuasive, and expressive. Focused-holistic scoring was defined as being a "criterion-referenced form of holistic scoring. A five point scoring guide ranging from 0 (not scorable) to 4 (excellent) was defined for each writing mode and was based on statewide purposes, a specified audience, and the constraint of the text. The participant displayed a 100 page scoring document which contained scoring guides, papers illustrating each score point, and directions for use. The scoring was accomplished by a contractor who employed scorers who met qualifications specified by the SDE. To score the state assessment it takes 100-155 readers seven to eight weeks working full time. Approximately 40 papers are read per hour. Total cost was estimated to be 12.71 per student. However, the scoring of the writing sample was a part of the larger state assessment contract.

B. A second state is involved in the developmental stage of a procedure identified as analytical holistic. Students were given a choice of two options. The writing task was defined in the prompt. Papers were scored holistically following procedures outlined by ETS. Following the holistic scoring of all items on a paper, analytic scoring was conducted on an item-by-item basis. Each item was scored analytically by one scorer. Scorers were teachers employed by the contractor. Scorers were able to score one paper per minute holistically and one paper every three minutes analytically. cost was estimated to be approximately $10 per pupil.

C. A third state participant described scoring procedures which paralleled the primary trait scoring procedures implemented by NAEP. Items developed by NAEP were used. General comparisons between the national performance and the state performance could then be made.

In addition to primary trait scoring, an expressive item was scored for secondary traits and the persuasive writing essay for writing mechanics. All scoring was accomplished by a contractor. No cost estimate was provided by the participant.

D. A fourth state participant described an analytical scoring procedure. A unique feature of this state's plan was the fact that selected teachers from all over the state were trained as scorers. One scorer could score approximately 50 papers. The cost of the program was estimated to be about $10 per student.

E. Local school systems in attendance also used a variety of methods. One LEA indicated that a form of "holistic" scoring was used. All teachers were dismissed one-half day to score papers. A second school system represented was in the developmental stage. The participant's comment was simply to indicate that after having the other reports that he had to go home and plan some more. A third LEA used a contractor.

5. *How are writing tasks developed?*
   a. *How many tasks are needed to secure a reliable and valid sample of writing?*
   b. *How are tasks developed?*
   c. *What types of prompts elicit the best responses?*

The number and kind of writing tasks appeared to vary with the scoring technique. Those systems utilizing primary trait scoring tended to develop writing tasks which were narrow, clearly defined, and audience-oriented. Since primary trait tasks are dependent on the writing mode and since such task tend to stimulate a short writing sample, usually more than one task was assigned.

On the other hand, tasks which were developed for holistic scoring tended to be of more open-ended nature. Such tasks were generally designed to stimulate a twenty-minute essay. Therefore, usually only one task was presented.

Most systems described similar procedures for developing the writing task which paralleled procedures indicated by the results of the

questionnaire. Once the method of evaluation was determined, tasks were designed by committees. In most cases these tasks were subjected to field-testing before being utilized in a state assessment program. However, the methods employed in analyzing the results from the field test differed from system to system. Many looked at the percentage of agreement obtained by scorers as well as the nature of the responses from students.

A variety of topics had been used by the participants. There appeared to be a general agreement that topics for the younger child must require very little information to be generated, whereas topics for the older student required that more information be generated. Participants did not agree if an audience should be specified or not. As to be expected, those favoring primary trait scoring also favored the specification of an audience. Opinions varied among those favoring holistic scoring. Some felt that the specification of the audience was nothing more than a "fabrication." The student was aware that the "real" audience was the teacher or the scorer as the case may be. One state defined the audience for the writer simply as that person who would score the writing.

Dr. Purves questioned if topics or scoring dealt with cultural differences. He was especially concerned that in the field test, if a topic was found to show undue bias, it was discarded. The culture of the population should be considered in designing the task.

6. *How is the scoring of large numbers of writing samples managed?*
   a. *What management plans have been used successfully to score the entire population?*
   b. *What sampling techniques yield the best results in describing the status of writing in a given population?*

A number of procedures have been used to manage the scoring of large numbers of papers. Several systems depend upon a contractor. One state brings together teachers selected from throughout the state to score the writing samples. Only a description of the status of students' writing could be generalized to the total population and a classroom report is impossible.

All agreed that to effectively impact instruction, the information must be returned for each school, each classroom, and each student. Dr. Purves suggested that a random sample of writing responses be scored at the state level as an anchor with the balance of the writing samples being returned to the school system from which they originated. At the system level a sample might be scored and the rest returned to individual schools for classroom teachers to score. With this plan all teachers would be involved in the scoring process thereby creating a situation which might impact instruction. The sample serving as an anchor at the state level would enable comparisons to be made between system samples and state samples, school and system samples, and school and state samples.

CHAPTER 4

CONCLUSIONS AND RECOMMENDATIONS

A central question underlying this research has been related to the necessity of a writing sample in a large-scale writing assessment. Apparently, numerous authorities in the field of English contend that a writing sample is necessary. It is the opinion of these proponents of a writing sample that objective tests do not measure the higher order thinking skills which are reflected in the student's ability to select and explore content, achieve a style and tone appropriate for a designated audience and organize content logically. Concerning mechanics, the proponents argue that objective tests only measure the student's ability to proofread and do not measure his ability to use mechanics correctly. The findings in this research indicated that there was no relationship between the scores of Louisiana students on the Louisiana State Assessment Test in Writing (an objective test) and the scores received on the writing sample which was scored by a variation of the Primary Trait System of scoring. However, a comparison of the skills measured on the objective test and the skills rated on the writing sample revealed that the two measures were not evaluating the same skills. Therefore, no conclusion can be made from these findings.

The scoring of mechanics appears to pose a problem for individuals implementing a large-scale writing assessment. For holistic scoring, a general impression of the student's ability to use the mechanics correctly is a consideration in assigning a score. A similar technique is applied in analytical scoring except that a separate score for mechanics is assigned. NAEP scores mechanics by determining the percentage of errors. The administrators in Louisiana, however, attempted to develop a criterion referenced system for scoring mechanics. The minimum competencies were defined and the students' use of mechanics was scored in terms of these minimum competencies. A separate score was assigned for each of the following: punctuation, capitalization, spelling, and syntax. If a student violated one convention specified as a minimum skill the student was scored "below minimum." The problem with this system was that a writing task does not necessarily stimulate all students to attempt the same mechanical conventions. Therefore, a "below minimum" score does not necessarily mean the same thing for all students. One student may write only a few sentences which require no internal punctuation. This student is assigned a minimum score. On the other hand, another student may write a lengthy, well-developed passage, attempt a number of internal punctuation marks, and missuse one of them. This student is assigned a "below minimum" score. Is the student really "below minimum" in his ability to use punctuation marks? How does he compare with the student who attempted no punctuation marks?

The question remains: can mechanics be scored on a writing sample in a large-scale assessment program by a system other than one dependent on scorer impressions?

A review of the literature did not appear to indicate that sufficient research has been conducted to either support the need for a writing sample or to indicate the power of the objective test for predicting writing ability. The classic study conducted by ETS did indicate that a forty-five minute objective test provided as much information as three forty-five minute essays (Godshalk, Swineford and Coffman, 1966). Dr. William Lutz, Professor of English at Rutgers and consultant for this study contended that it, was his opinion that items could be developed that might predict a students ability to write. At this time, more research is needed to determine the power of a score on an objective measure of writing skills as a predictor of writing ability. Until this question is satisfactorily addressed, the question concerning the necessity of the writing sample can not be answered.

The consensus among the participants at the conference seemed to be that the major reason for including a writing sample with any assessment of ability in written composition was political. The fact that a writing sample is required would, most agreed, cause more writing to be included in the curriculum. However, doubt was expressed concerning the amount and value of information yielded by a writing sample given present reporting practices. For example, if primary trait scoring is used, the report of scores indicates simply that a certain percent scored minimum on a particular item at a particular grade level, a certain percent scored above minimum and a certain percent below minimum. Unless the classroom teacher is thoroughly familiar with the nature and directions of the item and with details of the scoring guide, very little is yielded which can be translated into instructional improvements. This situation creates a certain irony in that the primary trait scoring system was conceived on the premise that specificity of writing objective and corresponding writing stimulus and rubric would produce information easily interpreted by the practitioner. Indeed if doubt exists that primary trait is yielding information which can be translated into improved instructional procedures, support of a pure holistic method of scoring becomes very dubious. By its very nature, holistic scoring does not easily yeild specific information for an item. Analytical scoring presents certain limitations in that the scale is not as appropriate to some writing tasks as it is to others.

Validity of information yielded is affected also by scorer reliability which is very difficult to maintain at a level over .70. Reliability figures are affected by, among other factors, the nature and complexity of the item, and the amount of writing produced. With reliability hovering on an average at the .70 to .80 range, the question must be raised of the validity of information yielded.

At best only a limited random sample can be scored in a statewide assessment. Therefore impact on specific LEA's and their individual schools is obviously limited to non-existent. However, suggestions were made of ways to filter assessment results down to the individual school level. For example, since every child at designated grade levels is tested, pull a random sample to satisfy the legislative mandate of a writing assessment. Then return all papers to individual school systems to be scored either by designated teachers or by all teachers at certain grade levels. Scorer reliability would not be relevant and teachers could

observe, first-hand, the performance of their students. As a result teachers would be in a better position to address student problems related to a specific mode of writing. Then, perhaps, instruction would be impacted.

The question of the feasibility of including a writing sample in a state-wide or system-wide assessment depends upon first, how the writing sample is scored and results reported and second, how the results are translated into improved instructional procedures. If a sample is scored and reported in order to satisfy a legislative mandate but does not result in targeted in-service for teachers with follow-up to check for implementation, then scoring of a writing sample would be difficult to support. However, if resulting data is sufficiently wrung for information and if this information is translated into improved instructional practices, scoring a writing sample may indeed be worth the cost involved in both time and money. It should be noted that no research has been conducted to determine the impact of assessment programs on instructional procedures practiced in the classroom.

Throughout the nation, local and state education agencies are concerned with the problem of determining what students can and cannot do in the basic skill areas. Writing is one of those concerns. While many agencies are attempting to measure the ability of students to write, the findings of this study indicate that there is little uniformity in how writing ability is measured. Most of the education administrators responding to the survey in this study indicated that they believed a writing sample was essential. However, the methods used to score the writing appeared to vary from agency to agency. Although, most of the administrators reported using holistic scoring, many interviews with several of the individuals indicated that holistic scoring techniques varied. After hearing Dr. Lutz describe the holistic technique as applied by ETS, one administrator remarked that he just thought he was using holistic scoring. Apparently, what he used and called holistic scoring was nothing like the procedure developed by ETS. The same was true with the Primary Trait System. While several agencies reported using the Primary Trait System, when the systems were described, each was different from the other and from the original system as developed by NAEP. Other agencies tended to report such systems as "focused holistic" and "analytical-holistic."

As administrators described the system which was unique to their respective agencies, they did so with a great amount of pride and a sense of ownership. The procedures described by other participants apparently offered no appeal. When violations to basic assumptions were suggested by the consultants, administrators tended to ignore these warnings.

The methods and techniques used to score writing samples in large scale assessment are simply too varied to draw any but the most general conclusions about their value. A number of questions, both technical and metaphysical, remain unanswered. The nature of the writing tasks which are designed to promote writing presents a major area of needed research. Very little is known at this point about how carefully norms are determined, how stable the results are, how the results might advantage one group over another, nor how accurate their predictive value is. Because competency

testing programs are being used to make important decisions about people's lives, it is imperative that administrators of testing programs, as well as the users, scrutinize their programs with care.

Project Abstract

The purpose of this study was to determine the "state of the art" of large-scale writing assessment in the nation. With the onset of the accountability movement, many state and local education agencies have been mandated to assess all of the basic skills, with written composition presenting the most formidable problems. These problems have resulted in the establishment of many varied approaches to the assessment process. In order to ascertain the status of large-scale assessment of writing in the nation, all fifty state education agencies and selected large city agencies were sent a questionnaire related to on-going practices being used in assessment programs. From those responding to the questionnaire, ten participants representing varying assessment philosophies were selected to attend a conference on writing assessment in New Orleans. The purpose of this conference was to provide both verification and clarification of questionnaire findings. As a result of both the questionnaire and the conference, the researchers drew conclusions concerning such problems as the selection of scoring procedures, development of items, selection and training of scorers, value and utilization of information yielded by a writing sample, scoring of mechanics, and other problems associated with the implementation of a large-scale writing assessment utilizing both objective and applied procedures.

APPENDIX

List of City School Systems Receiving a Questionnaire

Louisiana Scoring Guides for Writing Samples (Mechanics)

Key Participants and Consultants in National Writing Conference

List of City School Systems Receiving a Questionnaire

Atlanta, GA
Mr. Alonzo Crim, Superintendent
Ind. School District 203
224 Central Avenue, S.W.
Atlanta, GA 30303

Miami, FL
Mr. E. L. Whigham, Superintendent
1410 N.E. Second Ave.
Miami, FL 33132

Chicago, IL
Mr. James F. Redmond, Superintendent
228 North La Salle Street
Chicago, IL 60601

Houston, TX
Mr. George G. Garver
Superintendent of Houston ISD
3800 Richmond
Houston, TX 77027

Dallas, TX
Mr. Nolan Estes
Superintendent of Dallas ISD
3700 Ross Ave.
Dallas, TX 75204

Detroit, MI
Mr. Charles J. Wolfe, Superintendent
5057 Woodward
Detroit, MI 48202

Cincinnati, OH
Mr. Donald R. Waldrip, Superintendent
230 E. 9th St.
Cincinnati, OH 45202

Des Moines, IA
Mr. Dwight M. Davis, Superintendent
1800 Grand Ave.
Des Moine, IA 50307

Phoenix, AZ
Mr. Gerald DeGrow, Superintendent
2526 W. Osborn Rd.
Phoenix, AZ 85017

Jackson, MS
Dr. Brandon Sparkman, Superintendent
Box 2338
Jackson, MS 39205

Birmingham, AL
Mr. Henry C. Sparks, Superintendent
Board of Education Drawer 10007
Birmingham, AL 35202

Philadelphia, PA
Mr. Matthew W. Costanzo, Superintendent
Parkway at 21st St.
Philadelphia, PA 19103

Boston, MA
Mr. William J. Leary, Superintendent
15 Beacon St.
Boston, MA 02108

Louisville, KY
Dr. Newman Walker, Superintendent
Fourth at Broadway
Louisville, KY 40202

Tulsa, OK
Mr. Tom Summers, Superintendent
Tulsa, OK 74101

Raleigh, NC
Mr. C. L. Hooper
City Superintendent
601 Devereux St.
Raleigh, NC 27605

Augusta, GA
Mr. H. M. Duncan, Superintendent
Richmond County Schools
2083 Heckle St.
Augusta, GA 30904

Omaha, NE
Mr. Owen A. Knutzen, Superintendent
3902 Davenport
Omaha, NE 68131

Santa Fe, NM
Mr. Phillip Bebo, Superintendent
610 Alta Vista
Santa Fe, NM 87501

Seattle, WA
Mr. Forbes Bottomly, Superintendent
Seattle School District No. 1
815 Fourth Ave. N.
Seattle, WA 98109

Kansas City, KS
Mr. O. L. Plucker, Superintendent
Kansas City Wyandotte Unfd. Dist. 500
Library Bldg.
Kansas City, KS 66101

Wichita Falls, KS
Dr. Alvin E. Morris, Superintendent
Wichita Sedgwick Unfd. Dist. 259
428 S. Broadway
Wichita Fall, KS 67202

Madison, WI
Mr. D. S. Ritchie, Superintendent
545 W. Dayton
Madison, WI 53703

Baltimore, MD
Mr. Roland N. Patterson, Superintendent
3 E. 25th St.
Baltimore, MD 21218

Charleston, West Virginia
Dr. K. E. Underwood
Charleston, West Virginia 25311

Columbia, SC
Dr. Calud E. Kitchens, Superintendent
1616 Richland St.
Columiba, SC 29201

Richmond, VA
Dr. Thomas C. Little, Superintendent
312 N. 9th St.
Richmond, VA 23219

Newark, NJ
Mr. Edward Pfeffer, Superintendent
Education Bd. 31 Green St.
Newark, NJ 07102

Hartford, CT
Medill Bair, Superintendent
249 High St.
Hartford, CT 06103

Mobile, AL
Mr. Harold Collins
Mobile, AL 36601

Tallahasse, FL
Mr. F. W. Ashmore, Superintendent
P. O. Box 246
Tallahassee, FL 32302

Austin, TX
Dr. Jack L. Davidson
Superintendent of Austin ISD
6100 N. Guadalupe
Austin, TX 78752

Albuquerque, NM
Mr. E. Stapleton, Superintendent
Box 1927
Albuquerque, NM 87103

Cheyenne, WY
Dr. Joe Lutjeharms
Superintendent, Dist. 1
Cheyenne, WY 82001

New York, NY
Mr. Calvin E. Gross
Superintendent of School
110 Livingston St.
Brooklyn, NY 11201

Denver, CO
Dr. Allan M. Hosler
Supervisor
Development & Evaluation
Denver Public Schools
3800 York Street
Denver, CO 80205

Montclair, NJ
Mrs. Judi Granick
Director, Planning, Research & Evaluation
Montclair Board of Education
22 Valley Road
Montclair, NJ 07042

Denver, CO
Mr. [illegible]ry Beal
Supervisor, Department of Development & Evaluation
Denver Public Schools
900 Grant Street
Denver, CO 80203

Nashville, TN
Dr. Edward Binkley
Director
Department of Research & Evaluation
Metropolitan Public Schools
2601 Bransford Avenue
Nashville, TN 37204

New Orleans, LA
Dr. Constance C. Dolese
Director of Secondary Education
New Orleans Public Schools
4100 Touro Street
New Orleans, LA 70122

St. Louis, MO
Ms. Sandra Edelman
Administrative Assistant
St. Louis Public Schools
1517 South Theresa
St. Louis, Mo 63104

Downey, CA
Dr. Gordon E. Footman
Director, Division of Program Evaluation, Research, and
Pupil Services
Los Angeles County Superintendent of Schools
9300 East Imperial Highway
Downey, CA 90242

Portland, OR
Dr. Walter Hathaway
Portland Public Schools
P. O. Box 3107
Portland, OR 97208

Madison, WI
Dr. Darwin Kaufman
Evaluation Coordinator
Department of Public Instruction
126 Langdon Street, Rm 308
Madison, WI 53702

Spokane, WA
Ms. Sandra Meacham
Evaluation and Measurement Specialist
Central Valley School District #356
123 South Bowdish Road
Spokane, WA 99206

Monterey, CA
Dr. Lloyd Swanson
Director of Evaluation
Monterey Peninsula Unified
School District
P. O. Box 1031
Monterey, CA 93940

Lancaster, PA
Dr. John Tardibuono
Director, Project 81
School District of Lancaster
225 West Orange Street
Lancaster, PA 17604

Pontiac, MI
Dr. William Veitch
Assistant Director of Systematic Studies
Oakland Schools
2100 Pontiac Lake Road
Pontiac, MI 48054

Little Rock, AR
Dr. Carolyn Weddle
Assistant-Superintendent of
Program Implementation
Little Rock School District
West Markham and Izard
Little Rock, AR 72201

Fourth Grade

Scoring Guide

Sample

Secondary Trait - Spelling

Level 1 = If a paper does not qualify for a two, a score of one is given.

2 = Spelling is generally reasonable and indicates some concept of sound-letter relationship. The following words are spelled coreectly:

High frequency words (53)

a	are	boy	girl	it	the
am	at	can	he	of	they
and	be	for	in	on	to
was	you	I	is	that	

plurals of nouns by adding s (36)

basic color words (31)

number names through 10 (32)

3 = The writer adheres to the following conventions:

spells initial and final consonant sounds (B. 15 & 16)

spells short vowel sound in a word (B. 17)

spells long vowel sound in a word (B. 18)

spells phonetically regular words with CVC pattern (B. 19)

spells one syllable words with VC final e pattern (B. 20)

spells words with variant sounds of /c/ and /g/ (B. 21)

spells the initial sound in words using consonant blends (B.22)

spells number names through one hundred, days of week and months of year (B. 33, 34, 35)

spells plurals of nouns by adding <u>es</u>

spells words with final <u>y</u> changed to <u>i</u> before adding <u>es</u>.

spells verbs with <u>ing</u>

adds <u>ing</u> suffix:
directly to root word (B 39)
doubles final consonant (B. 40)
drops final <u>e</u> (B 41)

spells high frequency words

am	this	her	said	
all	with	his	what	
but	get	like	who	
by	how	little	she	
these	had	we	some	
will	have	why	their	
do	up	not	them	
down	write	one	when	
each	from	out	your	(B. 54)

spells holidays, seasons of the year, and frequently used school and community words

Fourth Grade

Scoring Guide

Secondary Trait - Capitalization

Level 1 = If the paper does not qualify for a two, a score of one is given.

2 = Errors in capitalization are present. Shows some concept of capitalization. Capitals are always present in the following instances:

Proper nouns (1)

Beginning of sentences (2a)

Pronoun "I" (2b)

Abbreviations (Mr., months, St., Rd., Ave., days of week, post office) (2c)

Initials (2d)

3 = The paper adheres to the following conventions:

Titles (books, poems, reports stories) (2e)

Titles of persons (Mother, Father, Aunt, Uncle) (2f)

Titles of address as a part of proper nouns (2g)

Heading, salutation, and closing of letters (4)

Fourth Grade

Scoring Guide

Secondary Trait - Punctuation

Level 1 = If a paper does not qualify for a two, a score of one is given.

2 = Errors are present, but the responses adhere to the following conventions:

Appropriate end punctuation (1a, 2, 3)

Uses comma correctly between days of month and year, after greeting and closing of a letter, between names of cities and states (4d1)

Uses colon appropriately in time of day (5a)

3 = The paper adheres to the following conventions:

Uses comma in words in a series (4,a.1)

Underlines titles of books (7a)

Uses apostrophe with possessive singular nouns (10a)

Indents and paragraphs:
Heading and closing of letter (11,a,1)

Beginning of a paragraph (11,a,2)

Fourth Grade

Scoring Guide

Secondary Trait - Syntax

Level 1 = If a paper does not qualify for a two, a score of one is given.

2 = The response contains at least one complete sentence. The writer adheres to the following conventions:

uses appropriate subject pronouns (I, we, he, she, it, you, they) E. 12

uses connecting words (and, but, or) E.13

uses a, an, the, appropriately (E. 19)

uses high frequency words correctly:

a	are	boy	girl	it	the
am	at	can	he	of	they
and	be	for	in	on	to
was	you	I	is	that	

3 = The response contains at least one complete sentence which adheres to all of the following conventions :

uses appropriate verb tense (E. 7)

uses appropriate noun form (E. 9)

uses appropriate form of singular possessive (E. 11)

uses correctly formed negative statements (E. 14)

uses appropriate word order (E. 17)

uses appropriate object pronouns (E. 16)

uses appropriate form of plural possessive

uses simple predicate to agree with simple subject (E. 18)

uses plural possessive nouns (E. 22)

uses appropriate helping and main verb combinations (E. 23)

uses comparative and superlative forms of adjective (E. 24)

uses appropriate demonstrative pronouns (E. 25)

uses appropriate inflectional endings to express correct verb tense and number (E. 26)

Key Participants and Consultants
National Writing Conference
New Orleans
July 10

Carol Robinson
Albuquerque Public Schools

Jim Hertzog
Pennsylvania State Department of Education

Kenneth Loewe
Florida State Department of Education

Gail J. Ames
Delaware Department of Public Instruction

Walter Hathaway
Portland Public Schools

Ray Crisp
Wichita Public Schools

Mary L. Crovo
Maryland State Department of Education

William J. Brown
North Carolina State Department of Education

Vicki Fredrick
Wisconsin Department of Public Instruction

Carol Ann Greenhalgh
Texas State Department of Education

Alan Purves
University of Illinois, Urbana

William Lutz
Rutgers University

Ina Mullis
NAEP

Other Participants
National Writing Conference
New Orleans
July 10

Louise A. Cobb
Louisiana State Department of Education

Marvin Zimmerman
Little Rock School District

Donna L. Nola
Louisiana State Department of Education

Cornelia B. Barnes
Louisiana State Department of Education

Joseph Williams, Jr.
Louisiana State Department of Education

Jean Halsell
Ouachita Parish Schools - LA

Ruth Berlin
Ouachita Parish Schools - LA

Hugh Peck
Louisiana State Department of Education

Margaret M. Ruska
Austin Independent School District

Rebecca Christian
Louisiana State Department of Education

Jimmie Steptoe
Louisiana State Department of Education

Sandra Konrad
Arkansas State Department of Education

BIBLIOGRAPHY

Anderson, C.C. "The New Step Essay Test as a Measure of Composition Ability." Educational and Psychological Measurement, (Spring, 1960), 95-102.

Arnold, Lois V. "Writer's Cramp and Eyestrain-Are They Paying Off?" In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 87-89.

Authur, Russell, John Gretencord, Sandra Johnson, and Robert Hunting. In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM, Urbana, IL: NCTE, 1965, 100-12.

Association of English Teachers of Western Pennsylvania. Suggestions For Evaluating Junior High School Writing. Urbana, IL: NCTE, n.d.

--------. Suggestions For Evaluating Senior High School Writing. Champaign, IL: NCTE, n.d.

Bader, Arno L., and William R. Steinhoff. "Ratings and Analysis of Student Themes." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 113-19.

Baily, Richard W. "Measuring Student Writing Ability." Manuscript. University of Michigan. Ann Arbor, MI.

Barnes, Douglas, James Britton, and Harold Rosen. Language, The Learner and the School, rev. ed. Baltimore, MD: Penguin Books, 1971.

Bata, E.J. "A Study of the Relative Effectiveness of Marking Techniques on Junior College Freshman English Composition." Ph.D. Dissertation. University of Maryland, 1973. DAI, (1973), 73-17028.

Bateman, Douglas, and Frank Zidonic. The Effect of a Study of Transformational Grammar on the Writing of Ninth and Tenth Graders. Urbana, IL: NCTE, 1966.

Beach, R. "Self-Evaluation Strategies of Extensive Revisers and Non-revisers." College Composition and Communication, 27 (1976), 160-64.

Beaven, Mary H. "Individualized Goal Setting, Self-Evaluation, and Peer Evaluation." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell. Urbana, IL: NCTE, 1977, 135-56.

Berger, Irwin, "Eleven Common Sense Principles About Language, and Student Writing Examined For Logic and Clarity." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 6-16.

Bernstein, Rusy S., and Bernard Tanner. The California High School Proficiency Examination: Evaluating the Writing Samples. (ED 147 806). Washington, D.C.: ERIC, 1977.

Bracht, G.H., and K.D. Hopkins. "The Community of Easy and Objective Tests of Academic Achievement." Educational and Psychological Measurement, 30 (1970), 359-64.

Braddock, R., R. Lloyd-Jones, and L. Schoer. Research in Written Composition. Champaing, IL: NCTE, 1963.

Brekke, Alice. "The Impact of Testing on One California University Campus: What the EPT Has Done To Us and For Us." Journal of the Council of Writing Program Administrators, 3, no 3 (Spring, 1980), 23-26.

Breland, H.M. "Can Multiple-Choice Tests Measure Writing Skills?" The College Board Review, no 103 (Spring, 1977), 11-13, 32-33.

---------. A Study of College English Placement and the Test of Standard Written English (College Board Research and Development Report RDR-76-77-4). Princeton, NJ: The College Board, 1977.

---------, and Judith L. Gaynor. "A Comparison of Direct and Indirect Assessment of Writing Skill." Journal of Educational Measurement, 16, no 2 (Summer, 1979), 119-28.

Britton, James. Language and Learning. Baltimore, MD: Penguin Books, 1970.

---------, et al. The Development of Writing Abilities(11-18). London: Macmillan, 1975.

Britton, J.N. Multiple Marking of English Composition: An Account of an Experiment. (ED 059 194). London: n.p., 1966.

Brown, Rexford. "What We Know Now and How We Could Know More About Writing Ability in America." Journal of Basic Writing, 1, No 4 (Spring/ Summer, 1978), 1-6.

Buxton, E.W. "An Experiment to Test the Effects of Writing Frequency and Guided Practice Upon Students' Skill in Written Expression." Ph.D. Dissertation, Stanford University, 1958.

California Association of Teachers of English. "California Essay Scale." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 147-60.

---------. A Scale For Evaluation of High School Student Essays. Champaign, IL: NCTE, 1960.

CEEB Commission on English. "Poetry Analysis From End-of-Year Exams." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 120-23.

Chase, Clinton I. "The Impact of Achievement Expectations and Handwriting Quality on Scoring Essay Tests." Journal of Educational Measurement, 16, No 1 (Spring, 1979), 39-42.

---------. "The Impact of Some Obvious Variables on Essay Test Scores." Journal of Educational Measurement, 5, No 4 (1968), 315-18.

Christensen, M. "The College-Level Examination Program's Freshman English Equivalency Examinations." Research in the Teaching of English, 11 (Fall, 1977), 186-92.

Clark, Michael. "There is No Such Thing as Good Writing (So What are We Looking For?)" Manuscript. University of Michigan. Ann Arbor, MI.

Clemson, E. A Study of the Basic Skills Assessment Direct and Indirect Measures of Writing Ability. Princeton, NJ: ETS, 1978.

Cleveland Heights-University Heights City School District. "Composition Writing Scale." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 159-60.

Coffman, W. E. "On the Reliability of Ratings of Essay Examinations in English." Research in the Teaching of English, 5 (1971), 24-36.

---------. "On the Validity of Essay Tests of Achievement." Journal of Educational Measurement, 3 (1966), 151-56.

---------, and Judith Papachristou. "Experimental Objective Tests of Writing Ability for the Law School Admission Test." Journal of Legal Education, (July, 1955), 388-94.

Cohen, Arthur M. "Assessing College Students' Ability to Write Composition." RTE, 7 (1973), 356-71.

Coleman, V.B. "A Comparison Between the Relative Effectiveness of Marginal, Interlinear, and Terminal Commentary, and of Audiotaped Commentary in Responding to English Compositions." Ph.D. Dissertation, University of Pittsburgh, 1972. DAI (1973), 73-04121.

Conlan, Gertrude. How the Essay in the College Board English Composition Test is Scored. An Introduction to the Reading of Readers. Princeton, NJ: ETS, 1976.

Cooper, Charles R. "The Holistic Evaluating of Writing." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell. Urbana, IL: NCTE, 1977, 3-32.

---------. "Measuring Growth in Writing." English Journal, 64, No 3 (1975), 111-20.

---------, and Lee Odell. Research on Composing: Points of Departure. Urbana, IL: NCTE, 1978.

Coward, Ann F. "The Method of Reading the Foreign Service Examination in English Composition." Research Bulletin, RB-50-57. Princeton, NJ: ETS, 1950. Out of print.

Davis, Ken. "Significant Improvement in Freshman Composition as Measured by Impromtu Essays: A Large-Scale Experiment." Research in the Teaching of English, 13, no 1 (February, 1979), 45-48.

Diederich, Paul B. Definitions of Ratings on the ETS Composition Scale. (ED 145 454). Washington, D.C.: ERIC, 1977.

---------. "How to Measure Growth in Writing Ability." EJ, 55 (April, 1966), 435-49.

---------. "In Praise of Praise." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 38-40.

---------. Measuring Growth in English. Urbana, IL: NCTE, 1974.

---------. "The 1950 College Board English Validity Study." Research Bulletin, RB-50-58. Princeton, NJ: ETS, 1950. Out of print.

---------. "Reading and Grading." In Improving English Composition. Eds. Arno Jewett, and Charles E. Bish. Washington, D.C.: National Education Association, 1965, Chapter 11.

---------, John W. French, and Sydell T. Carlton. "Factors in Judgments of Writing Ability." Research Bulletin, RB-61-15. Princeton, NJ: ETS, 1961. Out of print.

Dusel, William J. "Some Semantic Implications of Theme Corrections." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 44-47.

Earisman, Del. "Holistic Reading in the Writing Class." Manuscript. Upsala College, East Orange, NJ.

Elley, W. B. "The Role of Grammar in a Secondary School English Curriculum" New Zealand Journal of Educational Studies, 10 (May, 1975), 26-42. Rpt. in RTE, 10 (Spring, 1976), 5-21.

Emig, Janet. The Composing Process of Twelfth Graders. Urbana, IL: NCTE, 1971.

Euster, S. D. "Utilization of the Cloze Procedure as a Measure of Writing Skill of College Students." DAI, 39 (1979), 5900A.

Finalyson, Douglas S. "The Reliability of the Marking of Essays." British Journal of Educational Psychology, 21 (June, 1951), 126-34.

Finn, Patrick J. "Computer-Aided Description of Mature Word Choices in Writing." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell. Urbana, IL: NCTE, 1977, 69-88.

Follman, John G., and James A. Anderson. "An Investigation of the Reliability of Five Procedures for Grading English Themes." Research in the Teaching of English, 1-2 (Fall, 1967), 190-200.

Ford, B.W. "The Effects of Peer Editing/Grading on the Grammar-Usage and Theme-Composition Ability of College Freshman." Ed.D. Dissertation, University of Oklahoma, 1973. DAI (1973), 73-15321.

Fowles, M.E. Manual For Scoring the Writing Sample. Analytical Scoring, Holistic Scoring. Princeton, NJ: ETS, 1978.

Four Indiana Colleges. "Joint Statement on Freshman English in College and High School Preparation." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 23-28.

Frederick, V. Writing Assessment Research Report: A National Survey. Madison, WI: Wisconsin Department of Public Instruction, 1979.

Freedman, S. "The Evaluation of Student Writing." ED, 157 (1978), 79.

Freedman, S.W. "How Characteristics of Student Essays Influence Teachers' Evaluations." Journal of Educational Psychology, 71 (June, 1979), 328-38.

----------, and E. Nold. "A Multivariate Analysis of Readers' Responses to Essays." Research in the Teaching of English, 11 (1977).

French, J.W. Schools of Thought in Judging Excellence of English Themes. Princeton, NJ: ETS, 1962.

Godshalk, Fred I., Frances Swinefold, and William E. Coffman. The Measurement of Writing Ability. Princeton, NJ: College Entrance Examination Board, 1966.

Groff, P. "Does Negative Criticism Discourage Children's Compositions?" Elementary English, 52 (1975), 1032-34.

Grommon, Alfred, ed. Reviews of Selected Published Tests in English. Urbana, IL: NCTE, 1976.

----------. "Coordinating Composition in High School and College." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 90-95.

Grose, Lois M., Dorothy Miller, and Erwin R. Steinberg. "Suggestions for Evaluating Student Composition." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 90-95.

Hake, Rosemary. "With No Apology: Teaching to the Test." Journal of Basic Writing, 1, No 4 (Spring/Summer, 1978), 39-62.

Harris, Muriel. "Evaluation: The Process for Revision." Journal of Basic Writing, 1, No 4 (Sping/Summer, 1978), 82-90.

Hayadawa, S.I. "Linguistic Science and Teaching Composition." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 1-5.

Hilgers, Thomas L. "A Brief Note on Research Design and Reporting." Research in the Teaching of English, 13, No 3 (1979), 278-79.

Hillocks, George Jr. "Another Review of the Development of Writing Abilityes (11-18)." Research in the Teaching of English, 13, No 3 (1979), 284-88.

Hirsch, E.D. The Philosophy of Composition. Chicago, IL: University of Chicago Press, 1977.

Hogan, C. "Let's Not Scrap the Impromptu Test Essay Yet." Research in the Teaching of English, 11 (Winter, 1977), 219-25.

Hopkins, Thomas L. "The Marking System of the College Entrance Examination Board." Harvard Monographs in Education, Series 1, No 2 (October, 1921), 15pp.

Huddleston, Edith M. "Measurement of Writing Ability at the College-Level: Objective vs. Subjective Testing Techniques." Journal of Experimental Education, 22, No 3 (March, 1954), 165-213.

Hunt, Kellogg W. "Early Blooming and Late Blooming Syntactic Structures." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell. Urbana, IL: NCTE, 1977, 91-106.

Huntley, R.M., C.B. Scheiser, and R.J. Stiggins. "The Assessment of Rhetorical Proficiency: The Role of Objective Tests and Writing Samples." Paper presented at the annual meeting of the National Council on Measurement in Education, 1979.

Illinois English Bulletin. "Evaluating Ninth Grade Themes." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 96-100.

----------. "Evaluating Twelfth Grade Themes." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 130-38.

Jankinson, Edward, and Donald Seybold. Writing as a Process of Discovery. Bloomington, IND: Indiana Univeristy Press, 1970.

Jerabek, R., and D. Dieterich. "Composition Evaluation: The State of the Art." CCC, 26 (1975), 183-86.

Judine, Sister M., ed. A Guide For Evaluating Student Composition. Urbana, IL: NCTE, 1965.

Kentucky English Bulletin. "Two Types of Grading." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 138-46.

Klein, S.P., and F.M. Hart. "Chance and Systematic Factors Affecting Essay Grades." Journal of Educational Measurement, 5 (1968), 197-206.

Knapp, Thomas R. "Essay Topics and Modes, and Their Effects on Student Prose." Unpublished paper. Rochester, NY: University of Rochester, 1972.

Krupa, Gene H. "Primary Trait Scoring in the Classroom." College Composition and Communication, 30 (May, 1979), 214-15.

La Brant, Lou L. "Marking the Paper." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 29-37.

Lagana, J. R. "The Development, Implementation, and Evaluation of a Model for Teaching Composition Which Utilizes Individualized Learning and Peer Grouping." Ph.D. Dissertation. University of Pittsburgh, 1972, DAI (1973), 73-04127.

Larson, Richard, ed. Children and Writing in the Elementary School. New York: Oxford University Press, 1975.

---------. "Selected Bibliography of Writings on the Evaluation of Students' Achievements in Composition." Journal of Basic Writing, 1, No 4 (Spring/Summer, 1978), 91-100.

Leahy, Jack Thomas. "Objective Correlation and the Grading of English Composition." CCC, 25 (October, 1963), 35-38.

Lees, Elaine O. "Evaluating Student Writing." CCC, 30 (1979), 343-48.

Levine, Isidore. "From Guided Theme Reading to Improved Writing." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 48-60.

Lloyd-Jones, Richard. "Primary Test Scoring." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell. Urbana, IL: NCTE, 1977, 33-68.

Loban, Walter. The Language of Elementary School Children. Urbana, IL: NCTE, 1963.

Lynch, Catherine, And Patrica Klemans. "Evaluating Our Evaluations." College English, 40 (October, 1978), 166-70, 175-80.

Markham, L.R. "Influences of Handwriting Quality on Teacher Evaluation of Written Work." American Educational Research Journal, 13 No 4 (1976), 277-83.

Marshall, J.C. "Writing Neatness, Composition Errors, and Essay Grades Reexamined." The Journal of Educational Research, 65 (1972), 213-15.

Matthews, Roberta S. "The Evolution of One College's Attempt to Evaluate Student Writng." Journal of Basic Writing, 1, No 4 (Spring/Summer, 1978), 63-70.

McCleary, William J. "A Note on Reliability and Validity Problems in Composition Research." Research in the Teaching of English, 13 (October, 1979), 274-77.

McColly, W. "What Does Educational Research Say About the Judging of Writing Ability?" Journal of Educational Research, 64, No 4 (December, 1970), 148-56.

Mellon, J.C. National Assessment and the Teaching of English. Urbana, IL: NCTE, 1975.

Metzger, Elizabeth. "A Scheme for Measuring Growth in College Writing." Journal of Basic Writing, 1, No 4 (Spring/Summer, 1978), 71-81.

Michigan Newsletter. "Evaluating a Theme." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 75-86.

Miller, Peter. "An Analysis of Error-Types Used in the Interlinear Exercise of the College Entrance Examination Board's English Composition Test." Eleventh Yearbook, National Council on Measurements Used in Education (National Council on Measurement in Education), 1953-54, 19-20.

Morrison, R.L., and Phillip E. Vernon. "A New Method of Marking English Composition." British Journal of Educational Psychology, 11 (June, 1941), 109-19.

Moslemi, Marlene H. "The Grading of Creative Writing Essays." Research in the Teaching of English, 9 (1975), 154-61.

Mullis, I. The Primary Trait System for Scoring Writing Tasks. Denver, CO: National Assessment of Educational Progress, 1975.

Myers, Albert E., William E. Coffman, and Carolyn B. McConville. "Simplex Structure in the Grading of Essay Tests." Educational and Psychological Measurement, 1966 (in press).

National Assessment of Educational Progress. Explanatory and Persuasive Letter Writing(Writing Report No. 05-W-03). Denver, CO: NAEP, 1977(a).

---------. Expressive Writing: Selected Results From the Second National Assessment of Writing(Writing Report No. 05-W-02). Denver, CO: NAEP, 1976.

---------. Write/Rewrite: An Assessment of Revision Skills(Writing Report No. 05-W-04). Denver, CO: NAEP, 1977(b).

---------. Writing Mechanics, 1969-1974: A Capsule Description of Changes in Writing Mechanics(Writing Report No. 05-W-01). Denver, CO: NAEP, 1975.

---------. Writing Objectives. Denver, CO: NAEP, 1972.

Newcomb, J.S. "The Influence of Readers on Holistic Grading Essays." DAI, 38 (1977), 1133A.

Newsletter of the Michigan Council of Teachers of English. Evaluating a Theme. Ann Arbor, MI: MCTE, 1958.

Nold, Ellen, and Sarah Freedman. "An Analysis of Readers' Responses to Essays." Research in the Teaching of English, (Fall, 1977).

Noyes, Edward S., William M. Sale, and John M. Stalnaker. Report on the First Six Tests in English Composition. New York: College Entrance Examination Board, 1945, 72pp.

Nystrand, Martin. Assessing Written Communicative Competence: A Textual Cognition Model(ED 133 732). Washington, D.C.: ERIC, 1977.

---------. "Using Readability Research to Investigate Writing." Research in the Teaching of English, 13 (October, 1979), 231-42.

Odell, Lee. "Measuring Changes in Intellectual Processes as One Dimension of Growth in Writing." In Evaluating Writing: Describing, Measuring, Judging. Ed. C. Cooper, and L. Odell, Urbana, IL: NCTE, 1977, 107-34.

---------. "New Questions for Evaluators of Writing." Paper presented at the 4C's Meeting, march 1979, 15pp.

Olsen, Marjorie. "Summary of Main Findings on the Validity of the 1955 College Board General Composition Test." Statistical Report(SR-56-9). Princeton, NJ: ETS, 1956. Out of print.

———. "The Validity of the College Board General Composition Test." Statistical Report(SR-55-4). Princeton, NJ: ETS, 1955. Out of print.

Osterlund, B.L., and K. Cheney. "A Holistic Essay-Reading Composite as Criterion for the Validity of the Test of Standard Written English." Measurement and Evaluation in Guidance, 11 (1978), 155-58.

Page, Ellis B. "The Imminence of Grading Essays by Computer." Phi Delta Kappan, 47 (1966), 238-43.

———. "The Use of the Computer in Analysizing Student Essay." International Review of Education, 14 (1968), 210-25.

Pearson, Richard. "The Test Fails as an Entrance Examination." pp.2-9 in "Should the General Composition Test be Continued?" College Board Review, no 25 (Winter, 1955), 2-13.

Peterson, Edwin L. "A Magic Lantern for English." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 61-63.

Peterson, Stanley R. "Evaluating Expository Writing." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 65-74.

Pierson, H. "Peer and Teacher Correction: A Comparison of the Effects of Two Methods of Teaching Composition in Grade Nine English Classes." Ph.D. Dissertation, New York University, 1967.

Post, Winifred. "Five Compositions From a Tenth Grade College Preparatory Classroom," In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 124-29.

Rakes, Thomas A., and Lana McWilliams. "Bridging the Gap: Two Alternatives to Standardized Testing." English Journal, (October, 1978), 46-50.

Remondino, C. "A Factorial Analysis of Scholastic Compositions in the Mother Tongue." British Journal of Educational Psychology, 30 (1959).

Roody, Sarah I. "Managing Student Writing." In A Guide For Evaluating Student Composition. Ed. Sister M. Judine, IHM. Urbana, IL: NCTE, 1965, 41-43.

Sampson, Olive C. "Written Composition at 10 Years as an Aspect of Linguistic Development." British Journal of Educational Psychology, 34, Part 2 (June, 1964), 143-50.

Scannel, D.P., and J.C. Marshall. "The Effect of Selected Composition Errors on the Grades Assigned to Essay Examinations." American Educational Research Journal, 3 (1966), 125-30.

Schroeder, T.S. "The Effects of Positive and Corrective Written Teacher Feedback on Selected Writing Behaviors of Fourth-Grade Children." Ph.D. Dissertation. University of Kansas, 1973. DAI, (1973), 73-30869.

Sekera, C.R. 1978 Writing Assessment Report. Pueblo, CO: Department of Curriculum, School District No. 60, 1978.

Shale, D.G. "A Factorial Analysis of Essay Evaluations." DAI, 40 (1979), 553A-4A.

Shaughnessy, Mina. Errors and Expectations. New York: Oxford University Press, 1977.

Sheppard, E.M. "The Effect of Quality of Penmanship on Grades." Journal of Educational Research, 19 (1929), 102-05.

Slotnick, Henry B. "Toward a Theory of Computer Essay Grading." Journal of Educational Measurement, 9 (1972), 253-63.

---------, and John V. Knapp. "Essay Grading by Computer: A Laboratory Phenomenon?" English Journal, 60 (1971), 75-77.

Smith, Vernon H. "Measuring Teacher Judgment in the Evaluation of Written Composition." RTE, 3 (1969), 181-95.

Spandel, Vicki, and Richard J. Stiggins. Direct Measures of Writing Skill: Issues and Applications. Portland: Clearinghouse, 1980.

Stanton, B.E. "A Comparison of Theme Grades Written by Students Possessing Varying Amounts of Cumulative Written Guidance: Checklist, Instruction, and Questions and Feedback." Ed.D. Dissertation. Brigham Young University, 1973. DAI, (1974), 74-23646.

State of New Jersey Basic Skills Council. Scoring the Essays for the New Jersey College Basic Skills Placement Test. Princeton, NJ: ETS, 1979.

Steele, J.M. "The Assessment of Writing Proficiency via Qualitative Ratings of the Writing Samples." Paper presented at the annual meeting of the National Council on Measurement in Education, 1979.

Steller, N.A.D. "The Effects of Readers' Fatigue on the Grading of Essays." DAI, 39 (1978), 3541A-2A.

Stenberg, D.R. "A Comparison of Errors Made on the WEEPT and Those Found in Student Writing Samples as Bases for Placement in Freshman Composition Classes." DAI, 40 (1979), 2033A.

Stevens, A.E. "The Effects of Positive and Negative Evaluation on the Written Composition of Low Performing High School Students." Ed.D. Dissertation. Boston University, 1973. DAI, (1973), 73-23617.

Suhor, Charles. Report on "A Building-Centered Testing Program in Writing." Mass Testing in Composition. Is It Worth Writing Badly? New Orleans: Department of Resource Services, June, 1977.

Sutton, J.T., and E.D. Allen. "The Effect of Practice and Evaluation on Improvement in Written Composition." Cooperative Research Project, No. 1993. DeLand, FL: Stetson University, 1964.

Swineford, Frances. "College Entrance Examination Board General Composition Test J." Statistical Report SR-56-3. Princeton, NJ: ETS, 1956. Out of print.

---------. "Reliability of an Interlinear Test of Writing Ability." School and Society, 81, no 2050 (january, 1955), 25-27.

---------, and Marjorie Olsen. "Reliability and Validity of an Interlinear Test of Writing Ability." Research Bulletin RB-53-9. Princeton, NJ: ETS, 1953. Out of print.

Thomas, Macklin. "Construction Shift Exercises in Objective Form." Educational and Psychological Measurement, 16, No 2 (Summer, 1956), 181-86.

Thompson, R.F. Predicting Writing Quality. SN 1, No. 7. East Meadow, NY: English Studies Collection, 1976.

Underwood, D.J. "Evaluating Themes: Five Studies and An Application of One Study." M.A. Thesis. University of Illinois, 1968.

Vernon, P.E., ed. Secondary School Selection. London, England: Methuen, 1957.

Vernon, P.E., and G.D. Millican. "A Further Study of the Reliability of English Essays." The British Journal of Statistical Psychology, 7, Part 2 (November, 1954), 65-74.

Wallace, R.A. "Issues of Validity and Reliability in the Testing of Freshman Composition." DAI, 39 (1979), 4100A.

Weiss, Eleanor S. "The Interrelationships and Validities of Item Types in the College Board English Composition Test." Statistical Report SR-57-25. Princeton, NJ: ETS, 1957.

Whalen, Thomas E. "A Validation of the Smith Test for Measuring Teacher Judgment of Written Composition." Education, 93 (November, 1972), 172-75.

White, Edward M. "The California State University English Placement Test (EPT) - Purpose and Potential." Journal of the Council of Writing Program Administrators, 3 No 3 (Spring, 1980), 19-22.

----------. "Mass Testing of Individual Writing: The California Model." Journal of Basic Writing, 1 No 4 (Spring/Summer, 1978), 18-38.

Williams, Joseph. "Re-Evaluating Evaluating." Journal of Basic Writing, 1 No 4 (Spring/Summer, 1978), 7-17.

Wormsbecker, J.H. "A Comparative Study of Three Methods of Grading Compositions." M.A. Thesis. University of British Columbia, 1955.

ADDENDUM TO BIBLIOGRAPHY

Burros, Oscar K., ed. Eighth Mental Measurement Yearbook. Highland Park, New Hersey: Gryphon Press, 1980.

Braddock, Richard. Research in Written Composition. Champaign: NCATE, 1963.

Cooper, Charles. Evaluating Writing: National Council Teachers of English, 1977.

Diederich, P. B., French, J. W., and Carlton, S. T. "Factors in Judgments of Writing Ability." Research Bulletin RB - 61-15, Princeton, New Jersey: Educational Testing Services, August, 1961.

Diederich, P. B. "The Measurement of Skill in Writing," School Review, 54 (October, 1946), 584-592.

Godschalk and Swineford, and Coffman. The Measurement of Writing Ability, Princeton, New Jersey: College Entrance Examination Board.

Howerton, Mary Lou P. et. al. "The Relationship Between Quantitative and Qualitative Measures of Writing Skills," Educational Resources Information Center Ed. 137-410-416.

Lutz, William D. "How to Read 55,000 Essays a Year and Love It." Educational Resources Information Center Ed.

Pitts, M. "The Measurement of Students' Writing Performance in Relation to Instructional History." Paper presented atAERA, San Francisco, 1979.

"Resolution No. 1" CCCC (October, 1978), 309.

"Resolution No. 2" CCCC (October, 1974), 339.

Sanders, Sara E. and John H. Littlefield. "Perhaps Test Essays Can Reflect Significant in Freshman Composition: Report on a Successful Attempt."

Smith, Laura Spooner. "Measures of High School Students' Expository Writing: Direct and Indirect Strategies." Educational Resources Information Center Ed. 177.

Stanley and Hopkins. Educational and Psychological Measurement and Evaluation. Englewood, New Jersey: Prentice Hall (1972), 198.

Winters, L. "The Effects of Differing Response Criteria on the Assessment of Writing Competence." In E. Baker and E. Quellmalz. (Dirg) Studies in Measurement and Methodology - Work Unit 1: Design and Use of Tests (OB - NIE-G-78-0213), Los Angeles: University of California Center for the Study of Evaluation, November, 1978: also in "Alternative Scoring Systems for Predicting Criterion Group Membership." Paper presented at AERA, San Francisco, 1979.