DOCTORAL DISSERTATION SERIES

PUBLICATION: 5922

AUTHOR: Gerald Lloyd Kincaid, Ed. D., 1953
Michigan State College

TITLE: SOME FACTORS AFFECTING VARIATIONS IN THE QUALITY OF STUDENTS' WRITING

University Microfilms, Ann Arbor, Michigan

SOME FACTORS AFFECTING VARIATIONS IN THE QUALITY OF STUDENTS' WRITING

By

Gerald L. Kincaid

A THESIS

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Science in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

School of Education

1953

SOME FACTORS AFFECTING VARIATIONS IN THE QUALITY OF STUDENTS' WRITING

By

Gerald L. Kincaid

AN ABSTRACT

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Science in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

School of Education

Year 1953

Approved Milosh Muntyan

THESIS ABSTRACT

Statement of the problem. The purpose of this thesis was to determine whether a single paper written by a student on a given topic at a particular time can be considered as a representative sample of his writing ability--and thus provide a valid basis for evaluating ability at any time in a writing course. The use of a single paper for evaluating a student's achievement under such conditions involves four basic assumptions.

1. that any given topic provides the same stimulus as any other topic,
2. that any given topic elicits constant responses at different times,
3. that the psychological pressure introduced by an examination situation has no adverse effect on the quality of writing, and
4. that the quality of writing is stable from topic to topic and from time to time for any student regardless of writing ability.

Methodology. The data for testing these assumptions were obtained by having four groups of students write two papers on different topics on each of two different days. Two groups, A and B, wrote on both days without the pressure of an examination situation. The other two groups, C and D, wrote their two papers on the first day without the examination pressure. On the second day, however, they wrote their two papers on the same two topics under the pressure of the examinations situation. Thus, groups A and B provided a basis for determining variations in the quality of writing due to topics on each day and due to efficiency variations from the first to the second day, and provided a control group for determining the effect of pressure on groups C and D.

Findings. From the analysis of the average writing performance by groups of students, it was found that: neither the content factors, due to the different topics assigned, variations in the efficiency of students from day to day on the same topics, nor the psychological pressure introduced by the final examination situation had any significant effect on the average quality of writing by a group of twenty or more students.

From the analysis of the writing performance by individual students in each group, it was found that:

1. content factors and variations in efficiency were about equally responsible for significant variations in the quality of writing by 47 of the 80 students involved in the study; and those variations occurred with significantly greater frequency for the strong students than for the weak students.
2. for the strong students, dissimilar topics were responsible for no more significant variations than were the similar topics. For the weak students, however, dissimilar topics did seem to result in a significantly greater frequency of variations in the quality of writing than did similar topics.
3. the psychological pressure of the examination did not result in a significantly greater frequency of gains and/or losses for individual students than when that pressure was absent.

Implications. From the results of this study, it seems reasonable to conclude that for the purpose of evaluating student achievement at any time, or for the purpose of evaluating improvement of individual students in a writing course, samples of writing should be obtained on different topics on the same day and on the same topics on different days.

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to his adviser, Dr. Milosh Muntyan, for his encouragement and critical assistance throughout the development of this study.

He is also indebted to Dr. Harry W. Sundwall for his helpful suggestions in planning the study and in analyzing the results.

The writer deeply appreciates the cooperation of Mr. Paul D. Bagwell, Dr. Clyde W. Dow, and Mr. Judson M. Perkins in making available the students for this study.

Grateful acknowledgment is also due Drs. John Schmid and Willard G. Warrington for their suggestions and advice relative to the statistical treatment of the data.

****
****
****
*****
***
*

VITA

Gerald Lloyd Kincaid

candidate for the degree of

Doctor of Education

Final examination, May 20, 9:00 A.M., 202A Morrill Hall.

Dissertation: Some Factors Affecting Variations in the Quality of Students' Writing

Outline of Studies

Major fields of emphasis: Educational Psychology, Guidance and Counseling

Cognate field: Communication Skills

Biographical Items

Born, March 1, 1914, Palestine, Illinois

Undergraduate Studies, Eastern Illinois State Teachers College, 1935-39

Graduate Studies, University of Illinois, 1940-41, Michigan State College, 1948-53

Experience: Speech teacher and debate coach, Pekin Community High School, Pekin, Illinois, 1945-46, Instructor in Communication Skills, Michigan State College, 1946-50, Educational Research, Board of Examiners, Michigan State College, 1950-53

TABLE OF CONTENTS

CHAPTER I

STATEMENT OF THE PROBLEM AND ITS IMPORTANCE

Statement of the problem. Teachers of English composition courses have been faced with two major problems in evaluating student achievement. First has been the problem of obtaining reliable and valid ratings for each sample of writing. Second has been the problem of deciding what sample, or samples, of a student's writing should be used for a particular evaluation purpose.

When the teacher's purpose is to assign a grade best representing achievement in a writing course, the sample, or samples, of a student's writing should be obtained at the end of the course. But when the teacher wants to evaluate improvement in the quality of writing, a sample, or samples, should be obtained at the beginning and at the end of the course.

To accomplish these purposes, teachers frequently use only one sample of writing in each case. The reason for this is the amount of time and effort necessary to obtain reliable and valid ratings of English compositions. The question has been asked whether a single paper provides a representative sample of writing ability. Does the quality of a student's writing vary considerably from topic to topic and from time to time? Thus far, no studies have been reported which provide a definitive answer to that question.

The purpose of this thesis was to determine whether a single paper written by a student on a given topic at a particular time can be

considered as a representative sample of his writing ability--and thus provide a valid basis for evaluating ability at any time in a writing course. The use of a single paper for evaluating a student's achievement under these conditions involves the following basic assumptions:

1) that, for practical purposes, any given topic provides the same stimulus as any other topic,

2) that such a stimulus will elicit constant responses at different times,

3) that, since such a paper would be a final examination for the course, the psychological pressure introduced by the final examination situation would have no adverse effect on the quality of writing, and

4) that the quality of writing is stable from topic to topic and from time to time, with or without the psychological pressure of a final examination, for all students regardless of individual variations in writing ability.

Therefore, in order to determine whether a single paper written on a given topic at a particular time provides a valid basis for evaluating student achievement in a writing course, it was necessary to test the four assumptions stated above. This thesis is a report of a study planned and conducted for the purpose of testing those assumptions.

Importance of the problem. Perhaps the importance of obtaining a reliable sample of student writing for evaluation purposes can be better understood by examining 1) the relation of training in language skills to other educational objectives, 2) the relation of training in written expression

to the training in the language skills, and 3) the inability of composition teachers to adequately answer the criticism that too many students do not develop satisfactory writing skills.

Training in the use of the English language has been accepted as one of the most important objectives in American schools--as demonstrated by the universal requirement that all students, from the elementary to the college level, successfully complete courses which involve such training. The report of the President's Commission on Higher Education is typical of the emphasis commonly placed on the importance of training in language expression. According to that report, "Few of the abilities men possess are of greater human significance than their power to order ideas clearly and to set these before their fellows by tongue or pen."[1]

Although training in both oral and written expression has been universally recognized as an important educational objective, written expression has received the greatest amount of attention. This is demonstrated by the universal emphasis on writing in American public schools, and by a continuous stream of criticism to the effect that too many young people haven't been taught to write. The results of training in this area seem to have been subjected to more criticism than the results of training in any other area.

Greene has summarized the criticisms of training in written expression. He states that

> High school and college graduates are said to be unable to write legibly, spell accurately, or compose an acceptable

---

1. President's Commission on Higher Education, "Establishing the Goals", vol. 1, Higher Education for American Democracy, U. S. Government Printing Office, Washington, D. C., 1947, pp. 52-53.

> letter. Particularly are they criticized as being unable to think clearly and logically and to express their thoughts in well-chosen, properly enunciated words arranged in interesting and clean-cut sentences.[2]

Many teachers of English composition have come to accept such criticism as an inevitable concomitant to their profession. From the writer's observation, however, composition teachers resent the implication that their teaching is inferior to that in other areas. They maintain that incorrect language habits are difficult to change--that progress in changing such habits is necessarily slow. Several studies tend to support this claim. Lyman states that "It is now generally accepted that about .5 of a step on a ten-point scale is the normal composition improvement in one year."[3] Huddelson reported an average improvement per year of about four per cent.[4] Anderson and Traxler reported an average improvement per year for high school students of 3.3 points when the papers were rated on a sixty-point scale.[5] An unpublished study by the writer showed an average improvement of four points, at the end of a year's training of college freshmen, when the papers were rated on a fifty-point scale.

These averages are important only to the extent that they indicate small increments in writing improvement. Naturally, some students make

---

2. H. A. Greene, "English--Language, Grammar, and Composition", Encyclopedia of Educational Research, Macmillan, New York, 1950, p. 385.

3. R. L. Lyman, "Investigations in the Field of Written Composition", Summary of Investigations Relating to Grammar, Language, and Composition, Supplementary Educational Monographs No. 36, The University of Chicago, 1929, p. 196.

4. Earl Huddelson, "The Effect of Objective Standards Upon Composition Teachers' Judgments", Journal of Educational Research, vol. 12 (December, 1925), pp. 329-40.

5. H. A. Anderson, and A. E. Traxler, "The Reliability of the Reading of an English Essay Test," a second study, School Review, vol. 48, 1940, pp. 521-30.

more than average improvement; but, at the same time, a similar number of students make less than average improvement during any given training period. It is also true that for any group of students entering a writing course, usually a number of them will be considerably below average in writing ability. Unless those students make average or better than average improvement, the quality of their writing is likely to be fairly low even at the end of a year's training.

Lack of cooperation by teachers in other areas is also given as a reason for student failure to make the desired improvement in writing. Lange found that students were not held responsible for the quality of composition in other courses. He found that when students were asked to proof-read their papers written for courses other than English, they were able to correct one-third of their spelling errors and one-half of their punctuation errors. He also found that assignments were at fault, that some topics were too complex for clear and adequate answers in the time and space provided, that sketchy and ambiguous questions invited careless answers.[6] Thus, the students' experiences in other courses tended to counteract the efforts of the composition teachers.

Finally, there is the question of adequate teaching methods and procedures for bringing about the desired improvement in student writing. Composition teachers have made numerous claims for certain methods and procedures, but have been unable to present sufficient evidence to make their claims convincing, even to their fellow teachers. In his summary

---

6. Phil C. Lange, "A Sampling of Composition Errors of College Freshmen in a Course Other Than English", Journal of Educational Research, vol. 42 (November 1948), pp. 191-200.

of "Investigations in Methods of Teaching" (English composition), Lyman states that most of the studies were definitely lacking in certain respects: "1) they exercised inadequate control over teaching conditions; 2) they failed to isolate single variables for measurement; and 3) many of them employed unsatisfactory criteria for improvement. The experimenters themselves frankly admitted the imperfections thus enumerated."[7]

Although careful planning of a research study on methods of teaching composition may provide for adequate control over teaching conditions and for the isolation of single variables, such a study cannot be successful unless a satisfactory criterion for improvement in writing is employed. And the employment of a satisfactory criterion is dependent upon valid and reliable evaluations of the quality of student writing.

Tyler offers two assumptions basic to the educational process and important to the present argument: 1) Education is a process which seeks to change behavior patterns of human beings, and 2) evaluation of the educational program is a process for finding out to what degree these changes are taking place.[8] Certainly, the composition teacher seeks to change the language patterns of human beings. Even though this may be difficult, some methods may be better than others for bringing about the desired changes. Yet research on such methods can not be successful unless it is possible to find out to what degree those changes have taken place, i.e., unless student improvement in writing can be evaluated reliably and validly.

---

7. Lyman, op. cit., p. 253.

8. Ralph W. Tyler, "Purposes and Procedures of the Evaluation Staff", Appraising and Recording Student Progress, Harper and Brothers, New York, 1942, pp. 11-12.

Even though the lack of cooperation by teachers of other courses may tend to counteract the efforts of the composition teacher, research studies to determine the effects of obtaining such cooperation cannot be successful unless student improvement can be evaluated reliably and validly. Thus, reliable and valid evaluation of student improvement in writing seems to be the key to successful research on better teaching methods.

Considerable research has been conducted to improve procedures and techniques for evaluating writing ability. Such research has been on two types of evaluation--direct and indirect. Direct evaluation has consisted of determining the quality of a particular composition and assigning a value to that quality. Indirect evaluation has consisted of constructing and administering an objective-type test composed of test items concerning various elements of the writing process. The pertinent research literature pertaining to these two types of evaluation is reviewed in the following chapter.

CHAPTER II

RESEARCH ON METHODS FOR EVALUATING WRITING ABILITY-- REVIEW OF THE LITERATURE

During the past fifty years, considerable research has been conducted in an attempt to improve the reliability of evaluating both the writing ability of students and the quality of specific papers written by students. Attempts have been made to develop objective-type tests for evaluating students' writing ability and to develop more objective and reliable methods for evaluating the quality of the specific papers written.

The objective English test provides only an indirect evaluation of a student's writing ability. That is, the student is presented some written material. His activity consists of selecting the best of several suggested alterations of that material and then indicating that selection by making a mark in the appropriate place. He does no actual writing. Fairly high reliabilities have been obtained for such tests. Huddleston reports reliabilities ranging from .88 to .94 for a group of objective English tests.[1] Such reliabilities indicate that student performance on such tests does not vary greatly from time to time. However, some variation does occur. And if such a test is to be successful in measuring student improvement, then the improvement by each individual must be greater than the student's

---

1. Edith M. Huddleston, Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Techniques (Ph. D. thesis), New York University, 1952; published as Research Bulletin RB-52-7, Educational Testing Service, Princeton, N. J., 1952, pp. 9-10.

variation in performance on the test, otherwise no real improvement will be shown by the pre-test and post-test scores.

The validity of an objective English test is a different matter. The validity of a test is determined by comparing the performance of persons on the test with some outside criterion. As Adkins states, "One can resort to all sorts of statistical maneuverings with scores on a test and never fully establish its validity if an independent criterion is lacking."[2]

If an objective English test is supposed to evaluate a student's writing ability, then its validity for that purpose must, in the final analysis, be based on some other kind of evaluation of the student's actual writing performance. On this point Adkins states that

> . . . the best and most logical way to determine what a selection test predicts is to make a comparison between the performance of persons on the test and their performance on the job or on actual work assignments. If those who score high on the test are also the most successful on the job, the test is said to be valid for the purpose of predicting performance on that particular job.[3]

Few attempts have been made, however, to establish the validity of objective English tests for predicting the actual writing performance of students. In most validity studies, English grades have been used as the independent criterion. Huddleston reports validity coefficients ranging from .11 to .73, when English grades were used as the criterion in each case.[4] Edmiston and Gingerich obtained a validity of .55 for the "English

---

2. Dorothy C. Adkins, Construction and Analysis of Achievement Tests, U. S. Government Printing Office, Washington, D. C., 1947, pp. 161.

3. Ibid.

4. Huddleston, op. cit., pp. 10-11.

Usage Test of the Ohio State Every Pupil Tests", when the scores on a composition test were used as the independent criterion.[5]

Lockwood found that validity coefficients of intelligence tests with composition scores are about the same as those of objective English tests. He found the correlation between general intelligence (using the Otis test) to be .67 for 57 boys, and .76 for the 43 girls in the study. At the same time he found a correlation of .77 between composition scores and semester grades,[6] which is slightly higher than the highest validity coefficient reported by Huddleston for an objective English test with a similar criterion.

A part of the difficulty in obtaining estimates of objective English test validities can be attributed to the lack of a reliable criterion. Both course grades and ratings of individual compositions have been grossly unreliable in the past. However, the validity of objective English tests has been questioned on logical grounds. Greene concludes that "most objective tests measure only a few of the more obvious and mechanical skills. Important elements of style and quality undoubtedly lie beneath the surface of those mechanical factors."[7] Greene also states that the experimental

---

5. R. W. Edmiston, and C. N. Gingerich, "The Relation of Factors of English Usage to Composition", Journal of Educational Research, vol. 36, 1942, pp. 269-71.

6. H. R. Lockwood, Correlation of the Mental Maturity of One Hundred College Freshmen and Their Ability to Write English Composition (Master's Thesis), University of Chicago, 1925. (Seen in abstract only) R. L. Lyman, "Investigations in the Field of Written Composition", Summary of Investigations Relating to Grammar, Language, and Composition, Supplementary Educational Monographs No. 36, The University of Chicago, 1929, p. 175.

7. H. A. Greene, "English--Language Grammar, and Composition", Encyclopedia of Educational Research, Macmillan, New York, 1950, p. 394.

evidence reveals a lack of relationship between knowledge of those mechanical factors and a better utilization of them when writing compositions. Furthermore, he reports that in 1936, "the Curriculum Commission of the Teachers of English recommended that all teaching of grammar separate from the manipulation of sentences be discontinued, since every scientific attempt to prove that knowledge of grammar is useful has failed."[8]

Writing requires more than a recognition of correct mechanical factors; it requires a proper use of them. Likewise writing involves more than a recognition of valid interpretations of data; it involves the expression of original interpretations by the writer. The difference between making original interpretations of data and the recognition of such interpretations was illustrated in a study reported by Hartung, and others. Their study was conducted to determine a valid method for measuring students' ability to interpret data correctly. An objective test was constructed for that purpose. The student was presented a collection of data, together with a list of interpretations. He was to select those interpretations which the data justified. To provide a criterion for the objective test, the students had been provided the same data at an earlier time and had been asked to write free essay responses following such general directions as: "'Write five statements that you are sure are true according to the facts given in these data,' and 'Write three statements based on the data which you are not quite sure are true according to these data'."[9]

---

8. Ibid., p. 392.

9. M. L. Hartung, L. Weisman, H. G. McMullen, and H. C. Trimble, "Aspects of Thinking", Appraising and Recording Student Progress, Harper and Brothers, New York, 1942, pp. 65-67.

A comparison of the responses for the same individuals on the two test forms produced the following patterns:

a. The student reacts similarly on corresponding items of the two forms,

b. The student is overcautious on an item in judging interpretations made by others but goes beyond the data on the corresponding item in making his own interpretations. The reverse pattern also appears.

c. The student is either very cautious or goes beyond the data in judging interpretations made by others but is accurate when making his own interpretations. Here also the reverse pattern appears.[10]

The only claim made by the authors for the objective test was that it can be used as an index of the general accuracy with which a group can make original interpretations. On the other hand, the authors admit that its validity as an index for predicting the accuracy of original interpretations by individuals is not high.[11] The different patterns of behavior in dealing with ideas on the two test forms used in this study indicate rather clearly that a student's treatment of ideas presented in an objective test provides a very poor basis for predicting how he will treat those same ideas in a written composition.

In summary, it seems that the student's knowledge or recognition of mechanical and grammatical factors involved in writing has a low relationship with his actual use of them in his own writing. Also, it seems that the student's recognition of correct interpretation of data made by others has a low relationship with his own interpretation of the same data.

---

10. Ibid., p. 73.

11. Ibid., p. 72.

Therefore, on logical grounds, it seems reasonable to question the validity of the scores on an objective English test as an index either of specific aspects of a student's writing ability, or of his over-all writing ability at a particular time. At best, it seems, the objective English test scores can be used with some validity only for predicting the liklihood of a student's success or failure in a writing course. For that purpose, it has about the same predicting efficiency as a general intelligence test. It would seem that the objective English test would provide little or no basis for evaluating student achievement in a writing course, which could be used, in turn, as a basis for evaluating the effectiveness of teaching methods for individual students.

On the other hand, the traditional (direct) method of evaluating a student's writing has been subjected to more severe criticism than has the objective test. In fact, the objective test had its origin in an attempt to overcome the gross unreliability of scoring the essay test.[12] Stalnaker states that ". . . the most recurrent criticism of the essay test, and the one about which most has been written, concerns the unreliability of evaluating essay answers . . . . the typical essay test as typically handled, whether by the classroom teacher or by 'experts' . . . is not reliably graded and, therefore, cannot stand alone as a good measuring instrument."[13] Huddelson reports a study in which the average deviation of the scores assigned certain papers by ten teachers was more

---

12. Adkins, op. cit., pp. 6-7.

13. John N. Stalnaker, "The Essay Type Examination", Educational Measurement, American Council on Education, Washington, D. C., 1951, pp. 498-501.

than two years of normal pupil growth. Of eight themes, teacher A would have failed six while teacher B would have passed all of them.[14]

Stalnaker's statement relative to "experts" failure to grade essays reliably is partially supported by a recent study by Huddleston on "measurement of Writing Ability at the College-Entrance Level." The study was conducted at the request and under the partial supervision of the College Entrance Examinations Board. Yet Huddleston reported a reader reliability of only .62 for the essays used in the study.[15]

On the other hand, Stalnaker states that "Reliable readings of essay papers is possible, however, where the questions are carefully framed with the problem of evaluation in mind and where the readers are trained in the techniques of consistent reading."[16] Lyman also indicates that reliable readings of essays can be brought about by extensive training in the use of objective techniques.[17]

Reliable readings, although not typical, have been reported occasionally. Traxler and Anderson reported one study with a reader reliability of .94 on papers written on Topic A, and of .85 on papers written on Topic B.[18] In a later study, the same authors reported reliabilities by the same reader on different days of .89 and .94 for the same two topics,

---

14. Earl Huddelson, "The Effect of Objective Standards Upon Composition Teachers' Judgments", Journal of Educational Research, vol. 12 (December 1925), pp. 329-40.

15. Huddleston, op. cit., p. 73.

16. Stalnaker, op. cit., p. 503.

17. Lyman, op. cit., p. 196.

18. Arthur E. Traxler and Harold A. Anderson, "The Reliability of an Essay Test in English," School Review, vol. 43 (September 1935), pp. 534-39.

respectively. They also reported reliabilities, for two readers, of .86 and .89 on the same two topics, respectively.[19] Of eight reading periods reported by Huddleston for the College Entrance Examinations Board, four periods had reading reliabilities ranging from .82 to .89.[20] In a reliability study made by the writer, of ten pairs of readers, two pairs had reliabilities above .90.

Even when reliable readings of English compositions have been obtained, the problem of reliable evaluation of student writing ability has not been solved. In addition, it is necessary to consider the reliability of the student's performance from paper to paper. Diederich states that a correlation of only .55 can be expected between scores of papers by the same students on similar topics on different days--even when reliable readings have been obtained.[21] Traxler and Anderson reported a correlation of .60 between scores of papers by the same students on different topics, when the reader reliabilities were .94 on Topic A and .85 on Topic B.[22]

Thus it appears that the quality of a student's writing may vary noticeably either from topic to topic or from time to time, or both. No studies are reported, however, in which an attempt has been made to differentiate between these two sources and their possible effects on the quality of a student's writing.

---

19. H. A. Anderson, and A. E. Traxler, "The Reliability of the Reading of an English Essay Test", School Review, vol. 48, 1940, pp. 521-30.

20. Huddleston, op. cit., p. 7.

21. Paul B. Diederich, "The Measurement of Skill in Writing," School Review, vol. 55 (April 1948), pp. 375-84.

22. A. E. Traxler and H. A. Anderson, op. cit.

Since the reliability of an objective-type test usually is not influenced by difficulties in obtaining accurate scoring (which is a matter of clerical accuracy), considerable attention has been directed toward the sources of unreliability for that type of test. For the most part, those sources are directly related to variations in student performance on the test. Adkins states that

> The test itself often falls further short of perfection than is usual with physical instruments. But the variations in the persons measured--variations due to such influences as fatigue, previous testing with the same or a similar test, and lack of interest or effort--and variations in conditions and methods of test administration from time to time or from examiner to examiner often contribute most to the unreliability of test results. These latter factors, which originate outside the test itself, can be controlled to a considerable extent by careful attention to procedures of test administration.[23]

Adkins' description of "variations in the persons measured" indicates that a student's performance may vary from time to time as a result of psychological variations in behavior which are temporary in character. R. L. Thorndike indicates a similar source of variation in test performance. Under "temporary factors affecting performance on many or all tests at a particular time," he lists health, fatigue, motivation, emotional strain, external conditions, momentary set, fluctuations and idiosyncrasies of human memory, etc.[24] Thorndike, however, considers another source of variation which is more lasting or permanent. He calls this source "the chance element determining whether the individual does or does not know

---

23. Adkins, op. cit., p. 149.

24. R. L. Thorndike, "Reliability", Educational Measurement, American Council on Education, Washington, D. C., 1951, p. 568.

a particular fact."[25] According to Thorndike,

> There will be a certain amount of variation in specific bits of knowledge or skill, so that even the individual who has a high over-all ability in the area in question will lack specific items of knowledge or skill and the individual low in general performance will succeed on isolated items not known by his generally more proficient fellow.[26]

The sources of test variation described by Adkins as "variations in the persons" and by Thorndike as "temporary factors affecting performance" are described by Loevinger as "transitory variations in efficiency."[27]

On the other hand, Thorndike's more permanent source of variation concerning accidental knowledge (or lack of it) is included and extended somewhat by Loevinger under what she calls "accidental content factors".[28] To illustrate her extension of this source of variation, Loevinger states:

> We may consider a test containing a number of problems in verbal deductive reasoning. The problems would differ in subject matter, and apart from the difficulty of the relationships involved, some topics would be easier for a particular individual, relative to other individuals, than other topics, whether because of familiarity, congeniality, or specific emotional factors.[29]

(For convenience, these two sources of variation in student performance will be referred to, henceforth, as "content factors" and "variations in efficiency.")

---

25. Ibid.

26. Ibid.

27. Jane Loevinger, "A Systematic Approach to the Construction and Evaluation of Tests of Ability", Psychological Monographs, vol. 61, No. 4, 1947, p. 5.

28. Ibid.

29. Ibid.

The effect of these two sources on the reliability of tests is considered by Cronbach. He classifies reliability coefficients under three types: the coefficient of equivalence, the coefficient of stability, and the coefficient of stability and equivalence. "The coefficient of equivalence indicates how precisely the test measures the person's performance at a particular time."[30] It is obtained by correlating the scores on two tests designed to measure the same abilities, administered to the same students during the same period, or by what is known as the "split-half" method. This method usually consists of correlating the scores on the odd items with those on the even items. The coefficient of equivalence, then, is concerned with variations in student performance when the subject-matter (content factors) has been changed but while the "efficiency level" remains fairly constant.

"The coefficient of stability shows the extent to which scores on the particular test items are stable over a period of time. It indicates whether a sample of behavior taken at one time is typical of behavior at other times."[31] This coefficient is obtained by correlating scores on one test administered to the same students at different times. Thus the coefficient of stability is concerned with variations in student performance due to variations in efficiency when the content factors have been held constant.

The coefficient of stability and equivalence is obtained by correlating the scores on two comparable tests administered to the same students

---

30. Lee J. Cronbach, Essentials of Psychological Testing, Harper and Brothers, New York, 1949, pp. 65-69.

31. Ibid.

at different times. Consequently, both content factors and variations in efficiency affect the results. There is no way of knowing, however, which of the factors may be responsible for such variations as occur in the scores of any individual.

Loevinger discusses these two sources more specifically in relation to both the coefficient of equivalence and the coefficient of stability. She states that when the reliability coefficient (of equivalence) is computed by the split-half method, using the scores on the odd and even items to make up the two halves, the content factors act to lower the coefficient obtained, while the efficiency level (entering both scores in the same way) acts to raise the coefficient obtained.[32] In other words, the content factors are primarily responsible for the variations which occur between the students' scores on the two halves of the test. Since the efficiency level of the student enters into both sets of scores in the same way, there is no indication of how much variation might occur if the test were administered at another time.

But when the test-retest is used to obtain the coefficient of stability, Loevinger indicates that variations in efficiency level act to lower the coefficient, while the content factors (being repeated) may act to raise the correlation.[33] In other words, variations in efficiency seem primarily responsible for the variations which occur between the students' scores on the two administrations of the same test at short intervals. Since the

---

32. Loevinger, op. cit., pp. 5-6.

33. Ibid.

content factors remain the same, there is no indication of how much variation would occur if two comparable tests (differing only in content) were administered during the same period.

Thus, to determine the relative effects of these two sources on individual students, it would be necessary to administer two comparable tests during the same period and then to repeat both tests at another time. Variations in performance on the two tests during the same periods would provide an estimate of the effect of content factors; while the variations in performance on the same tests at different times would provide an estimate of the effect of variations in efficiency.

Studies of intelligence tests provide some concrete evidence supporting these contentions about the effects of content factors and variations in efficiency. Cronbach reports a split-half coefficient of .91 from an administration of the "Detroit Beginning First-Grade Intelligence Test", Administration of the same test four months later produced a test-retest coefficient of stability of .76.[34] The period between the two administrations of the test did, of course, provide an opportunity for those who were unfamiliar with the content of the examination at the time of its first administration to gain that familiarity before the second administration. In this case the effect of the content factors may have been as great as though a different test had been administered. In other words, the split-half coefficient of .91 provided an estimate of variation due to content factors at that particular time. The test-retest coefficient of

---

34. Cronbach, op. cit., p. 69.

.76 provided an estimate of the effect not only of variations in efficiency but also of content factors as well.

Stroud reports that even when intelligence tests are administered only two or three days apart, "the median variability is normally expected to be about 5 IQ points. That is, one-half of the pupils change 5 IQ points or more; a few change markedly."[35]

Not only has a variability in student performance on intelligence tests been noted from day to day, but a difference in the amount of variation has been noted for students with different abilities. Stroud reports that the probable error of an IQ score ranges from 1.49, for scores below 70, to 3.54, for scores above 130.[36] Merrill reports a similar range for the standard error of IQ scores--from 2.2, for scores below 70, to 5.2, for scores above 130.[37] Since such tests are carefully administered, and since the variations are both plus and minus, these findings would suggest that, insofar as such tests denote ability, the degree of variations in test performance is in direct proportion to the ability of the individuals.

An unpublished study by the writer suggests that variations in the quality of student writing may closely parallel the variations according to ability found in student performance on intelligence tests. This study concerned evaluating improvement in the quality of student writing over a

---

35. James B. Stroud, Psychology in Education, Longmans, Greene, and Company, New York, 1946, p. 300.

36. Ibid., p. 301.

37. Maud A. Merrill, "The Significance of IQ's on the Revised Stanford-Binet Scales", Journal of Educational Psychology, vol. 29 (December 1938) pp. 641-651.

period of a year's training in composition. Each student wrote a paper at the beginning and another at the end of the year. Both papers were on the same topic. Twenty-six and eight-tenths per cent of the 198 students involved received lower scores on the final paper than on the initial paper. Eighty per cent of these students who received lower scores on the final paper, had received scores on their initial papers which were above the mean for that group of papers. Insofar as the score on the initial paper indicated the student's writing ability, the fact that this eighty per cent were above the initial mean suggests that students above average in writing ability may vary more in the quality of their writing than those with less than average ability.

Although much of our educational measurement has tended to ignore the variations in student performance noted thus far, such variations in human behavior are consistent with current psychological theory and findings.

However, the use of a student's score on a single test or on a single paper as indicative of his achievement at that particular time is to operate on the basis of the mechanistic (reflex-arc) psychology in vogue at the turn of the century and shortly thereafter. As Mowrer states, this psychology was based on two abstractions, S (stimulus) and R (response), "with only a thin, equally abstract, arrow connecting them."[38] In answer to Fearing's repeated reference "to the quality of 'invariability' which reflexes are alleged to possess," Mowrer writes:

> . . . from the outset this property was an embarrassment to psychologists who were attempting to make the reflex arc the

---

38. O. H. Mowrer, "Learning Theory", Review of Educational Research, vol. 22, No. 5 (December 1952), p. 476.

> cornerstone of a new psychology, embarrassing chiefly for the reason that anything approximating an adequate theory of behavior must provide for the occurrence of change, plasticity, learning. The classical conception of the reflex was in this respect singularly deficient.[39]

E. L. Thorndike noted the failure of the mechanistic psychology to account for variability in human behavior. In 1907, he wrote:

> . . . for any human being's thought and conduct, depending as they do upon the action of his nervous system, will sometimes show mysterious alterations--behavior unexplainable by the laws of instinct, association, and dissociation. The nervous system is influenced not only by the factors accounted for in these three laws, but also by fatigue, drugs, sickness, the decay of old age, shock, the chance variations of blood pressure, metabolism and the like.[40]

Several years later, Thorndike reported a study which supported the view that human behavior is variable from time to time. His study was concerned with plotting the work curve of five different graduate psychology students on five different days. The work consisted of adding numbers for one and one-half to two hours, recording the time, in seconds, at the completion of each row of 16 examples--thus obtaining the length of time required to complete each row of addition problems. Thorndike concluded that "the variation in the form of the work curve on different days is so great as to require careful consideration of the unreliability of any determination based on only a few days' records."[41]

---

39. Ibid.

40. E. L. Thorndike, The Elements of Psychology, The Mason-Henry Press, Syracuse, New York, 1907, p. 222.

41. E. L. Thorndike, Mental Work and Fatigue and Individual Differences, Teachers College, Columbia University, New York, 1923, p. 53.

Thorndike gave the complete data for one subject on the time required to complete each row of examples on each day. Computing the average time required to complete each of the nine rows of examples used on five different days provided a check on the variation of over-all efficiency on different days. The following variations were found in the average number of seconds required to complete each row of examples:

1st Day	2nd Day	3rd Day	4th Day	5th Day
142.7	129.7	150.3	119.4	104.8

The maximum difference in the average time required to complete each row of addition problems was 45.5 seconds--between the third and fifth days. In other words, the subject spent 43.4 per cent more time per row on the third day than on the fifth day. The minimum difference was 7.6 seconds--between the first and third days. The subject spent 5.3 per cent more time per row on the third day than on the first day.[42]

On the other hand, Gates reported a study on variations in efficiency of performing addition problems at different times during the same day. The greatest variation noted during the day was 4.2 per cent,[43] which is less than the minimum difference noted between days. Thus it appears that variations in performance on different days is likely to be greater than such variations at different times during the same day.

Gates also departs from the basic concept of the mechanistic psychology in his discussion of human behavior, classifying the sources of

---

42. Ibid., p. 49.

43. Arthur I. Gates, Psychology for Students of Education, The Macmillan Company, New York, 1931, p. 471.

variation as external and internal:

> Most of the complex reactions with which psychology is concerned moreover are made not to a single and simple stimulus but to a combination of forces. Prominent among the forces are the activities going on within the person at the time. The activity of any bodily mechanism serves as a partial cause or stimulus for further activity. Thus behavior of a man is determined by the combined and coordinated effects of what we may for convenience divide into external and internal activities, including in the latter both conscious and unconscious activities.[44]

Although Gates made this statement in 1931, it is closely paralleled by Mowrer's recent summary of "learning Theory":

> . . . It is now apparent that these two forms of learning exist and function, not side-by-side, but end-to-end: sign learning is the process whereby external events come to produce internal drive states, and solution learning is the process whereby internal drive states produce external, overt behavior. We thus advance from a simple, and pretty clearly inadequate, S-R psychology to an S-R:S-R psychology. By drawing a circle (O) around the R:S part of this sequence, we not only rediscover the 'organism' but we redeem it from the state of 'emptiness' to which extreme behaviorism condemned it, and thus begin, realistically, to examine the organism-as-a-whole.[45]

Thus, the topic assigned at a particular time, or the subject matter of a particular test item, may be considered as an external factor, while the activities going on within a person at a particular time may be considered as internal factors varying from time to time. Inasmuch as the topic assigned, or the subject matter of a particular test item, may affect the internal activities of the individual, the two sources of variation are inseparable. In general, however, the internal factors (the efficiency level of the individual) may be considered as the primary variable when

---

44. Ibid., p. 65.

45. Mowrer, op. cit., p. 492.

the same content factors are repeated a few days apart. Likewise, when different content factors are presented during a short period, the efficiency level should remain fairly constant--thus the content factors may be considered as the primary variable.

SUMMARY

As previously indicated, the teaching of written expression is considered as one of the most important educational objectives. At the same time, student achievement in this area has been subjected to considerable criticism. The failure of students to attain the desired goals has been explained as follows:

1) Improvement in language habits is a slow process.

2) Lack of cooperation by teachers in other areas tends to counteract the work of the composition teacher.

3) Teaching methods in use may be inadequate.

Research related to these factors has been limited in its value by the lack of valid and reliable evaluation techniques for determining student improvement in writing. Although objective English tests have proved to be reliable instruments, they have not been validated for evaluating improvement in the quality of student writing. On a logical basis, it seems doubtful that objective tests can be developed that will accomplish that purpose.

On the other hand, the direct evaluation of the quality and the improvement of student writing has been hindered by two major obstacles: 1) Reliable ratings of samples of student writing have been difficult to

obtain, and 2) the reliability of a student's writing performance has been questioned.

Through the intensive training of raters in the use of writing scales and various types of score cards, reliable ratings of student writing have been obtained. Also, there is considerable evidence to indicate that "content factors" and "variations in efficiency" are the primary sources of variation in student performance on tests. Yet most studies conducted to evaluate student improvement in writing have used a single composition for the pre-test and another composition on a different topic for the post-test. Even though such evaluation procedures have shown a small over-all improvement in the quality of writing, the results were practically meaningless for evaluating individual improvement. In two studies, the scores reported were lower on the final papers than on the initial papers for a considerable portion of the students. It seems reasonable to assume in each case that those students did not become poorer writers during the year's training period. Thus it seems reasonable to conclude that the evaluation procedure was at fault. If the evaluation procedure was responsible for the apparent regression in writing ability for many students, it seems reasonable to assume that the evaluation procedure also resulted in apparent gains which were exaggerated for others.

At best, it would seem that the usefulness of current evaluation procedures in this area has been limited to providing the teacher with some evidence that a group of students, as a whole, has made some improvement, and that such improvement is rather small. Evaluation procedures have failed to provide the teacher the pertinent information about individual

improvement which he needs before he can intelligently approach the task of improving his teaching methods and procedures. As Tyler states, one of the important purposes of evaluation ". . . is to provide information basic to effective guidance of individual students . . . . Merely the judgment that he is doing average work in a particular course is not enough. We need to find out more accurately where he is progressing and where he is having difficulties."[46]

For this purpose, we need, first of all, reliable and valid ratings of the student's writing. Second, we need to have a reliable sample of a student's writing at the time his achievement is being evaluated. And, if a reliable sample cannot be obtained, then enough samples should be obtained to provide a reliable estimate of writing ability. If, for example, the quality of a student's writing varies from topic to topic, then we would need samples of his writing on different topics; and, if the quality of his writing varies from time to time, then we would need samples of his writing at different times; and, if the psychological pressure of the final examination situation has an adverse effect on the quality of a student's writing, then we would need samples of his writing under other conditions; and, if the quality of student writing varies in amount from topic to topic and from day to day in direct proportion to student ability in writing, then we would need more samples of writing for those with more ability than for those with less ability--in order to obtain a valid estimate of each student's writing ability.

---

46. Tyler, op. cit., pp. 8-9.

The purpose of this thesis was to test these basic assumptions on the stability of a student's writing performance--to determine whether a single paper written by a student on a given topic at a particular time can be considered as a representative sample of his writing ability, and thus provide a valid basis for evaluating a student's ability at any time in a writing course.

CHAPTER III

DESIGN AND PROCEDURE FOR CONDUCTING THE STUDY

As previously stated, four basic assumptions are inherent in the use of a single paper written by a student on a given topic at a particular time as a basis for evaluating his achievement at any time in a writing course. Such a practice assumes that:

1. any given topic provides the same stimulus as any other topic,
2. any given topic elicits constant responses at different times,
3. the psychological pressure introduced by the examination situation has no adverse effect on the quality of student writing, and
4. the quality of student writing is stable from topic to topic and from time to time, with or without the pressure of an examination, regardless of variations in student writing ability.

Assignments and Design

Since the purpose of this thesis was to test these four assumptions, specific writing assignments and specific conditions relative to each assumption were provided for groups of students enrolled in the first term of the Written and Spoken English[1] course at Michigan State College. This is a required three-term (quarter) course in the Basic College, which is a general education program. Students receive training in four communication

---

1. In 1952, the name of the course was officially changed to "Communication Skills."

skills: writing, speaking, reading and listening. They attend classes five hours each week--one hour of lecture, two hours of recitation, and two hours of laboratory. All papers are written in class during these two-hour laboratory periods.

Requirements for the first assumption. To test the first assumption--that any given topic provides the same stimulus as any other topic--it was necessary to have students write papers on two or more topics during one writing period, so that the general efficiency level of the students would enter into the writing for each topic in about the same manner. The selection of the topics presented two problems, one concerning the nature of the topics and the other concerning the number of topics to be assigned.

Although Diederich[2] has recommended that similar topics be used in order to obtain the best estimate of a student's writing ability, frequently the training in a writing course involves more than one type of writing, or one type of topic. For example, in the Written and Spoken English Course at Michigan State College, the two assignments considered near the end of the first term are concerned with 1) the development of an idea by the use of two or more methods, and 2) the giving of directions or the description of a process. The question arises whether the quality of a student's writing would be the same on two similar topics related to one or the other of those assignments. And, would the variations between

---

2. Paul B. Diederich, "The Measurement of Skill in Writing," School Review, vol. 55 (April 1948), pp. 375-84.

two papers on similar topics related to one assignment be any different than on dissimilar topics related to the two assignments?

In order to resolve this, it was necessary to assign both similar and dissimilar topics. By combining two similar topics with one dissimilar topic, only three topics were necessary to provide a satisfactory test for the first assumption.

The specific selection of the topics was made in collaboration with the two instructors from the Written and Spoken English Department, whose students wrote the papers used in this study. Since the papers were to be written near the end of the term, the selection of the topics was limited to the nature of the two course assignments during that period. It was agreed that the two similar topics could best be selected from the course assignment concerning "the development of an idea by the use of two or more methods," and that the dissimilar topic could be selected from the assignment concerning "the giving of directions or the description of a process." Thus, the two similar topics could be developed by the use of similar methods, with the content providing the major variable. On the other hand, the dissimilar topic would differ from the other two topics not only in content but, also, in the method needed for development.

Insofar as possible, topics were selected concerning material common to all students enrolled in the course. The two similar topics assigned were: 1) "It Is (Is Not) Too Far Between Classrooms for Students to Travel During the Ten-Minute Period Allowed," and 2) "Textbooks for Freshmen at Michigan State College Are (Are Not) Too Expensive." Each student was provided a brochure of information pertaining to each topic.

The dissimilar topic assigned was: "Give Directions to a Stranger at the Union Building Enabling Him to Get to Building A-6 on South Campus." For this topic, each student was provided a map of the campus. (For convenience, henceforth, these three topics will be referred to in abbreviated form as: "Distance Between Classes", "Cost of Textbooks", and "Giving Directions", respectively.)

These three topics provided the necessary variations in stimuli to test the assumption that "any given topic provides the same stimulus as any other topic." It would have been desirable to have each student write on all three topics--two similar and one dissimilar--during one writing period. Such a procedure, however, would have introduced the possibility of fatigue developing to the extent that the efficiency level of the students would be reduced considerably before the task was completed. Such an effect would make it impossible to determine variations in the quality of writing due to the topics assigned. Therefore, it was considered advisable to restrict the writing assignments to two topics during a single two-hour writing period.

With such a restriction, it was necessary to use two groups of students in order to compare variations in the quality of writing between papers on similar topics with variations between papers on dissimilar topics. Group A was asked to write a paper on each of the two similar topics. Group B was asked to write a paper on one of the similar topics, "Distance Between Classes", and another paper on the dissimilar topic, "Giving Directions". This procedure provided a basis for comparing both the effect of similar topics and the effect of dissimilar topics on the quality of

student writing--thereby providing a basis for testing the assumption that "any given topic provides the same stimulus as any other topic". (Provision for testing whether fatigue entered into the writing of the second paper is explained later under "Control of the fatigue factor".)

Requirements for the second assumption. To test the second assumption--that any given topic elicits constant responses at different times--it was necessary to have students write on the same topic on different days. Thus, by repeating the assignments described for testing the first assumption, a basis was provided for testing the second assumption. In fact, the assignment of the same topics to the same students on a second day provided sets of papers on the three different topics as a basis for comparing the quality of student writing on different days--a basis for determining whether variations in efficiency from day to day have any effect on the quality of student writing.

Requirements for the third assumption. To test the third assumption--that the psychological pressure of the examination situation has no adverse effect on the quality of student writing--it was necessary to have students write without that pressure at one time and to write on the same topic with that pressure at another time.

Two other groups of students, groups C and D, were used for this purpose. They became the pressure group, while groups A and B served as a control--writing on both days without the psychological pressure of the examination situation.

It was necessary to have the students used for testing the third assumption (the pressure groups C and D) write their first set of papers

under the same conditions and with the same assignments used for the control group. Thus, group C was assigned to write on the two similar topics: "Distance Between Classes" and "Cost of Textbooks", and group D was assigned to write on the two dissimilar topics: "Distance Between Classes" and "Giving Directions". Then by adding the examination situation as the only variable for these pressure students on the second day, a basis was provided for determining the effect of the examination situation on the quality of student writing.

Requirements for the fourth assumption. To test the fourth assumption--that the quality of student writing is stable from topic to topic and from time to time, with or without the pressure of an examination, regardless of variations in writing ability--it was necessary to rank the students according to their writing ability and then to analyze the degree and frequency of variations in the quality of writing in relation to the rank of individuals. (The method of ranking is described in Chapter V.)

Summary. The following writing assignments provided the bases for testing each of the first three assumptions:

1. For the assumption that any given topic provides the same stimulus as any other topic, two papers were written during a single writing period by each student in Group A on similar topics, and in Group B on dissimilar topics.

2. For the assumption that any given topic elicits constant responses at different times, the initial assignments for Groups A and B were repeated.

3. For the assumption that the psychological pressure introduced by the examination situation has no adverse effect on the quality of student writing, two papers were written by each student in Group C on similar topics and in Group D on dissimilar topics, without the examination pressure on the first day and with that pressure on the second day.

Design in tabular form. The similarities and variations in the writing assignments for the four groups of students are shown below in tabular form:

Key to tabulation: Topic 1 - "Distance Between Classes"
Topic 2 - "Cost of Textbooks"
Topic 3 - "Giving Directions"
W/O - Writing without the examination pressure
W - Writing with the examination pressure

	First Day Similar Topics		Second Day Similar Topics	
	Topic 1	Topic 2	Topic 1	Topic 2
Group A	W/O	W/O	W/O	W/O
Group C	W/O	W/O	W	W

	Dissimilar Topics		Dissimilar Topics	
	Topic 1	Topic 3	Topic 1	Topic 3
Group B	W/O	W/O	W/O	W/O
Group D	W/O	W/O	W	W

Controls Provided

Although the arrangement of assignments and conditions shown above provided the basic design necessary for testing the four assumptions involved in this study, certain precautions and controls were necessary to preclude the intrusion of extraneous factors into the writing situations

which would invalidate the findings. It was considered advisable to select groups of students whose schedule of classes corresponded with the design of the study so that the necessary assignments could be made without creating an artificial situation. Controls were provided to: 1) prevent fatigue from counteracting the effect of different topics assigned during the same writing period, 2) provide the appropriate motivation for each writing situation, and 3) provide the appropriate length of time between the two writing periods to prevent undue distortion from new learning experiences or from memory of the first writing experience.

Selection of Groups. Through the cooperation of the head of the Written and Spoken English department, and two staff members of that department, four groups of students were provided for writing the appropriate number of papers, and under the appropriate conditions, necessary for this study.

Each staff member was teaching two groups of beginning students in the course. Both instructors met each of their two groups on the same days, Tuesdays and Thursdays, for two hours each day. Each group contained a sufficient number of students so that, barring an unusual occurrence of absences, a minimum of twenty students would be present on both days the writing assignments were to take place. Thus, these four groups would provide eighty students, each of whom would write four papers under the conditions described above.

Since the students in these groups were in the habit of writing papers in class during a two-hour laboratory, the writing assignments for this study were made to coincide with those laboratory periods--thus providing for a minimum departure from the normal classroom procedure.

Control of Fatigue Factors: Having students write two papers during a two-hour period could easily result in a certain amount of fatigue during the second half of that period. Such fatigue, if it occurred, might well bring about a lowering of the efficiency level of the students, thereby lowering the quality of the writing during the second hour. Thus, such fatigue would invalidate the findings in relation to the effect of the topics on the quality of the writing for that period.

To reduce the possibility of fatigue, the students were instructed to complete the first paper during the first hour and to take a ten-minute break before starting the second topic.

To provide a check on the effect of the fatigue factor, every other student was instructed to write on the first topic during the first hour and on the second topic during the second hour. The other students were instructed to write on the second topic during the first hour and on the first topic during the second hour. This procedure provided a basis for determining whether the students writing on a specific topic during the first hour did better than those writing on the same topic during the second hour.

Control of Motivation Factors: Having a student write an essay as a part of an examination which is to determine his grade for the term is usually considered as a motivating factor, as well as a psychological pressure factor which may disturb him. That is, the reward of a good grade or the fear of a poor grade will usually stimulate the student to try to do good work. On the other hand, the routine writing assignment in the Written

and Spoken English course also is a factor, although a minor one, in determining the student's end term grade.

In order to better determine the effect of the psychological pressure of the examination, it was decided to eliminate the grade factor (as a possible disturbance) from the writing situation on the first day for all of the students involved, and to eliminate the grade factor on the second day for the non-pressure students, groups A and B, who served as the control group in relation to the pressure factor.

Elimination of the grade factor raised a motivation problem for all of the students on the first day's assignments, and for the non-pressure students on the second day's assignments--since it was desirable that all students should attempt to write good quality papers for all assignments.

It was decided to replace the grade factor with an explanation of the assignments, indicating how the writing assignments could be of value to the students. On the second day the grade factor was involved only for the pressure students, groups C and D. For those students on that day, the grade factor was intensified by providing an examination situation.

Instructions: The non-pressure students, groups A and B, received the following instructions to provide the necessary motivation in place of the grade factor on the first day:

> It is believed, as a result of several studies, that in order for a student to have a fair opportunity to demonstrate his writing ability, he should write on two different topics on the same day. Your grade for the term will not be affected by the papers you write today. However, your papers will be graded by a team of raters. You will be informed of the grade given your papers so that you can compare those grades with those you have received on previous papers. Also, those

> grades should give you a better idea of your actual writing ability at present than does a grade on any single paper.

On the second day, the non-pressure students received the following explanation:

> As a result of several studies, it is believed that in order for a student to demonstrate his writing ability, he should write on two different topics on the same day and then repeat the performance a few days later. Today you are asked again to write on the same topics assigned at the last writing period. Again your grade for the term will not be affected by the grade you receive on these papers. However, your papers will be graded by a team of raters to discover whether you can write better on one topic than on another, and whether you can write better on one day than on another. This information should be valuable to you during your next two terms of work in the course. The results of the grading will be posted on the bulletin board early next term.

For the pressure students, groups C and D, the explanation for the first day's assignment was the same as for the non-pressure students. On the second day, in order to provide the examination pressure, the following explanation was placed in the assignment:

> During the last writing period you wrote two themes which did not affect your grade for the term. As promised, your papers will be graded by a team of raters to determine whether you wrote better on one topic than on the other topic. These results will be posted on the bulletin board as soon as possible. It is felt, however, that after having written on a topic once you should be able to do a better job of writing if given another opportunity to write on the same topic. Today you are to write on the same two topics again. Since you have had a practice period for writing these two papers, the themes you write today will be considered as your final and best effort. They will be considered as a part of your final examination, and the grades on these two papers will be counted in with your speech grades and your final examination grade to make up your grade for the term.

(These explanations were effective to the extent that students began to inquire about the results before they were posted.)

It was felt that the above explanations for the non-pressure situations would provide the students with a reason for completing the

assignments, but that such explanations would provide a minimum of student concern about the grades assigned their papers. Thus, if the final examination pressure has an adverse effect on the quality of student writing, that effect should be apparent from the variations in the quality of student writing noted from day to day for the pressure students in comparison with those noted from day to day for the non-pressure students.

Control of the time factor: The amount of time elapsing between the writing of the first and second set of papers on the same topics was considered in relation to two problems. First, if the time between the writings is too short, the students may remember rather clearly how they wrote the first paper. As a result, the second paper on the same topic may be merely a second edition of the first paper without a re-examination of the problem of developing the topic assigned--which, otherwise, would be necessary for the writing of a new paper on the same topic.

On the other hand, if too much time elapses between the writing of the first and second paper on the same topic, new learning experiences may well be the cause for many of the differences which may be noted--instead of such differences being caused by variations in efficiency.

A compromise had to be made between these two extremes in order to reduce to a minimum the effects of memory and the effects of new learning experiences. Since the students were enrolled in the Written and Spoken English course, the compromise had to favor the shorter time. Otherwise, the effect of new learning experiences might result in variations in the quality of writing for a considerable number of students. Thus, it was considered advisable to space the two writing periods without any intervening

writing periods for regular class work. The maximum time between two such periods was one week. Therefore, it was arranged to have the two sets of papers on the same topics written during the writing laboratory periods of each of the last two weeks of the term. This arrangement provided a natural time to introduce the final examination situation into the writing assignment for the pressure students on the second day.

Summary. The design for this study consisted of having four groups of students write two papers during a single writing period on each of two days. These four groups were given specific assignments under the appropriate conditions to provide a basis for testing the four assumptions which are the concern of this study.

Groups A and B wrote on similar and dissimilar topics, respectively, under the same conditions on each of two different days--thus providing a basis for testing the effect of topics and the effect of variations in efficiency from day to day on the quality of student writing, as well as providing a control group in relation to testing the pressure factor.

Groups C and D also wrote on similar and dissimilar topics, respectively, on each of two different days--with the examination situation on the second day providing the variable to be tested in relation to its effect on the quality of student writing.

In addition, controls were provided, insofar as possible, to isolate the variables being tested. These controls were concerned, primarily, with providing groups which could write under uniform conditions, providing an equal amount of time for writing each paper, providing uniform

motivation when the pressure factor was absent, and providing controls to reduce to a minimum the effects of fatigue, of memory, and of new learning experiences.

CHAPTER IV

METHOD FOR EVALUATING THE QUALITY OF STUDENT WRITING

As explained in the preceding chapter, assignments and conditions were provided for isolating the variable in each assumption to be tested. Yet, in the final analysis, those assumptions could not be tested unless the error in evaluating the quality of writing was less than the variations in the quality of writing from topic to topic and from time to time. Therefore, obtaining reliable and valid ratings of the papers involved in this study was considered a crucial factor in successfully testing the four assumptions.

In order to obtain the most valid and reliable ratings possible, it was considered important to select the best available method for rating themes, to determine what procedures have been most successful in using that method, and to establish controls consistent with the assumptions being tested.

Selecting the Rating Method

In order to determine whether the theme rating method employed by the Written and Spoken English Department during the past seven years, or some other method, should be used for this study, the pertinent research literature on theme rating methods was reviewed.

Writing scales: Shortly after the turn of the century, considerable attention was devoted to the development of writing scales in order to improve the reliability of rating English compositions.

Hillegas[1] was the first to attempt a scientific development of a writing scale, which was to provide for accurate comparisons between the quality of writing at the same school at different times or at different schools at the same time. In the development of the scale, no attempt was made to define different qualities of writing. Merit was a term used to indicate that quality which competent judges agreed upon.

The theory submitted as the basis for developing the scale was that: "Differences that are equally often noticed are equal, unless the differences are either always or never noticed."[2] The unit in this scale was defined as that difference which exactly seventy-five per cent of the judges are able to distinguish. All that was required to derive that unit was a set of samples that varied from each other by small degrees of quality. When two samples were found on which seventy-five per cent of the judges agreed in calling one better than the other, the difference between the two samples was one unit in the scale.[3]

Trabue worked out two supplements to the Hillegas scale in order to: 1) supply a need for a supplementary scale composed of compositions of the same general type as those written by Nassau County pupils, and

---

1. Milo B. Hillegas, "Scale for the Measurement of Quality in English Composition by Young People," Teachers College Record, vol. 13 (September 1912), p. 339.

2. Ibid., p. 344.

3. Ibid., p. 347.

2) supply some tentative standards indicating the quality of English compositions to be expected from the pupils of any given school grade.[4] These statements suggest that compositions must be judged in relation to the type of writing involved. Yet Hillegas, in answer to an objection raised relative to comparing narrative and descriptive writing, stated merely that the judges involved had not offered any such objections.[5]

Since the sample compositions used in constructing the Hillegas and Trabue scales were comparatively short (from approximately 50 to 250 words in length), it seems doubtful that judgments were made relative to any aspect of writing other than form--that judgments could have been made relative to supporting materials or to the organization of them. Lyman concludes that: "Objective and specific categories of excellence are generally inadequate in composition scales."[6] Also, studies by Dolch[7] and Leonard[8] indicate that there is little or no correlation of excellence between the content, the organization, and the mechanics of a composition--that such characteristics should be evaluated separately.

Inasmuch as the Written and Spoken English course places as much emphasis on the use of relevant supporting materials and the organization

---

4. M. R. Trabue, "Supplementing the Hillegas Scale," Teachers College Record, vol. 18, 1917, p. 51.

5. Hillegas, op. cit., p. 384.

6. R. L. Lyman, "Investigations in the Field of Written Compositions," Summary of Investigations Relating to Grammar, Language, and Compositions, University of Chicago, 1929, p. 195.

7. E. W. Dolch, Jr., "More Accurate Use of Composition Scales," English Journal, vol. 11 (November 1922), pp. 536-44.

8. S. A. Leonard, "Building a Scale of Purely Composition Quality," English Journal, vol. 14 (December 1925), pp. 760-75.

of them as on writing form, any rating method which fails to evaluate such characteristics would be unsuitable for evaluating the papers written for this study.

Also, as Thurstone[9] has demonstrated, the theory on which the Hillegas scale and the Trabue supplements are based is likely to be false when applied to specimens of English compositions--that when we have complex, overlapping stimuli involved, we cannot expect differences which are equally often noticed to be equal.

Thus, it seems doubtful that either the Hillegas or the Trabue scale contains units approaching equality--as has been claimed. In fact, such inequality of units seems to have been supported by Trabue's observation of judgments concerning two compositions which previously had been equated. He stated that "Such large differences in the form of surfaces of distributions of judgments were certainly unexpected. They tend to make one critical of the assumption that the variability of judgment on one composition is equal to the variability on any other."[10]

Neither Hillegas nor Trabue reported any reliability studies on the use of their scales by two or more people on the same set of compositions. Darsie reported a correlation of .88 between two sets of ratings from the use of the Willing scale.[11] According to Lyman, the evidence is about

---

9. L. L. Thurstone, "Equally Often Noticed Differences," Journal of Educational Psychology, vol. 18 (May 1927), pp. 292-93.

10. Trabue, op. cit., p. 51.

11. M. L. Darsie, "The Reliability of Judgments Based on the Willing Composition Scale," Journal of Educational Research, vol. 5 (January 1922), pp. 89-90.

equally divided as to whether the use of composition scales is effective in reducing the disparity of teacher's marks. He does state, however, that extensive training in the use of such scales is effective in that direction.[12]

Sorting methods: In a series of studies conducted in the early 1930's, Sims attempted to improve the reliability of theme ratings by having the readers sort the papers into five stacks according to certain standards. Utilizing this method, he obtained ratings with a coefficient of contingency of .77, which he interpreted as an estimation of the reliability of these ratings.[13]

Sims also used the same procedure for rating essay examinations: rating by sorting the discussion questions, one at a time, into five different stacks according to certain standards. Again he obtained a reliability of .77.[14]

Percentage basis: Sims then tried another approach, which consisted of having the readers mark the papers on a percentage basis--with a maximum score of 100 as a basis for marking. He then determined letter grades by using the mean and standard deviation for determining the divisions between

---

12. Lyman, op. cit., p. 196.

13. Verner M. Sims, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," Journal of Educational Research, vol. 24 (October 1931), pp. 216-23.

14. Verner M. Sims, "Reducing the Variability of Essay Examination Marks Through Eliminating Variations in Standard of Grading," Journal of Educational Research, vol. 26 (May 1933), pp. 637-47.

grades. He concluded that this procedure was effective in reducing variations caused by different standards, but did not draw any conclusions regarding the relative merits of the two systems he had employed.[15]

Use of standards: Stalnaker used still another approach. Over a period of four years, attempts were made to reach agreement on standards. Each rater continued to mark papers by assigning letter grades. At the end of four years, the reliability of the ratings was only .55. He concluded that, in order to improve the reliability of theme ratings, certain characteristics should be agreed upon by the teaching profession. Then each characteristic should be rated separately on the basis of five points--5 to 0. He suggested such characteristics as organization, coherence, and mechanical perfections, with more to be added.[16]

Rating characteristics: Stalnaker's suggestion has been utilized to a considerable extent in recent years. As early as 1917, Sackett reported the use of a score card in a study comparing different methods of rating compositions. His score card consisted of three major categories, each having three sub-divisions. The total possible points for each category is shown below:

Correctness--30
- Spelling and capitalization-- 15
- Paragraphing-- 8
- Punctuation-- 7

---

15. Verner M. Sims, "Improving the Measuring Qualities of an Essay Examination," Journal of Educational Research, vol. 27 (September 1933), pp. 20-31.

16. John M. Stalnaker, "Question IV, The Essay," English Journal (College Edition), vol. 26 (February 1937), pp. 133-40.

Sentence Structure--20	
Syntax--	10
Simplicity--	7
Diction--	3
Thought-Content--50	
Originality--	25
Unity--	15
Coherence--	10

Sackett evaluated four methods of grading compositions: teacher's judgment, the Hillegas scale, the Ballou narration scale, and the score card. He found the score card to be somewhat superior to the other three methods.[17]

Anderson and Traxler, using a similar rating method, obtained reliabilities ranging from .86 to .94. They used eight separate categories in their rating card: Accuracy, 6; Completeness, 6; Spelling, 6; Punctuation, 6; Language errors, 6; Coherence between main divisions, 10; Organization of paragraphs, 10; and Organization of sentences, 10.[18]

Diederich reported a study comparing the reliability of the rating method described above (which he calls the atomistic method) with the rating of themes by assigning a single score to each paper--based on a ten-point scale (which he calls the wholistic method). He found the atomistic method more reliable, but more time-consuming, than the wholistic method. He suggests that the wholistic approach might become as reliable were as much time devoted to the rating of each paper.[19]

---

17. Leroy W. Sackett, "Comparable Measures of Composition," School and Society, vol. 5 (February 1917), pp. 233-39.

18. H. A. Anderson, and A. E. Traxler, "The Reliability of the Reading of an English Essay Test," School Review, vol. 48, 1940, pp. 521-30.

19. Paul B. Diederich, "Readers' Methods Scrutinized in English Essay Scoring," Educational Testing Service Developments, No. 1 (October 1951), p. 2.

Gerber, after using the atomistic method for rating both themes and speeches, recommended the following characteristics for the rating of speeches and themes: purpose, content, organization, language, and presentation (special problems of oral or written delivery).[20]

For the past seven years, the atomistic approach has been used for the rating of comprehensive examination themes of students completing their work in the Written and Spoken English course at Michigan State College. During that time a rating form, similar to that described by Gerber, has been used which consists of five categories. Each category is rated on a ten-point scale, as shown below, (See Appendix A for definition of each category):

THEME RATING SCALE

Name and/or Number_____ Date_____

	Superior 10 9 8 7 6 5 4 3 2 1 Unsatisfactory
Conventions of Grammar	
Sentence Structure	
Diction	
Organization	
Content	

Rater_____ Total_____

In a reliability study of ten pairs of raters using the form shown above, the present writer found reliabilities ranging from .36 to .94; two pairs of raters had reliabilities above .90. Thus, it appears that, if the theme raters are properly selected and trained, the score card or rating form provides a method for rating themes which is as reliable as any method yet reported.

---

20. John C. Gerber, "Testing and Evaluation in the Skills of Communication," College English, vol. 9 (April 1948), pp. 375-84.

It should be noted, however, that this method is subject to one of the same criticisms made in relation to the Hillegas and Trabue writing scales. That is, there is no assurance that this method of rating English compositions produces scores consisting of equal units. Thus, statistical treatment of scores obtained in such a manner may be open to question.

Another possible weakness of the rating method used by the Written and Spoken English department was indicated in a study by Starring. His study concerned the actual use of the rating form by staff members when certain elements in the themes had been weakened. He had five sets of themes rated, and then weakened a different element in each set of themes. Starring discovered that, when the weakened themes were rated again, the raters did not make a distinction between weakness in sentence structure and weakness in conventions of grammar; nor did they make a distinction between weakness in content and weakness in organization. In other words, although the rating form is a list of five categories to be regarded as discrete and of equal weight, it was not so regarded by the raters. In actual practice, the raters scored the themes on three categories: 1) a combination of conventions of grammar and sentence structure, 2) diction, and 3) a combination of organization and content.[21]

These findings are valid, perhaps, only for the raters used in Starring's study, or for a similar group of raters. It should be noted that Starring had twenty-four members of the Written and Spoken English Department rate the papers used in his study. No analysis was made to

---

21. Robert W. Starring, A Study of Ratings of Comprehensive Examination Themes When Certain Elements Are Weakened (Unpublished Ed. D. Thesis), Michigan State College, August 1952, pp. 104-105.

discover whether any of the raters departed from the patterns noted for the group of raters employed.

The findings of Dolch and Leonard (reported earlier) that there is little correlation in excellence between the organization and the content of papers indicate that such elements can be distinguished by raters in relation to differences in excellence. Thus, if accurate evaluations of student writing are to be made in relation to the separate elements designated on a rating form, it would seem important to have the raters carefully selected and trained for that purpose.

Since it has been demonstrated that the rating method employed by the Written and Spoken English Department can produce highly reliable ratings when proper precautions are taken, and since the rating form employed by that department was designed for evaluating the objectives of that course, that method and that form seemed to provide the best available means for evaluating the papers involved in this study.

Obtaining Reliable Ratings

In order to insure the appropriate use of the rating method and the rating form selected, it was considered important to determine the number and types of raters needed for that purpose, and to use the most valid method for determining the score for each paper.

Determining number of raters: Gerber[22] and Diederich[23] both agree that,

---

22. Gerber, op. cit., pp. 375-84.

23. Diederich, "The Measurement of Skill in Writing," op. cit., pp. 585-92.

in order to obtain the best evaluation of a theme, two raters should be employed and that the average of their ratings should be used. Furthermore, when the disparity between the two raters is excessive, a third rater should be employed. If the third rater agrees with one of the first two raters, the average of those two ratings is used. If there is no close agreement between any two of the three ratings, the average of the three ratings should be used.

This procedure has been followed for the rating of comprehensive examination themes at the completion of the Written and Spoken English course at Michigan State College for the past seven years. It has been the practice at this institution to use a third rater when the disparity between the first two ratings was in excess of five points on the total score--the maximum of which is 50 points for each rating.

Selecting raters: Diederich points out that to obtain significant differences, when evaluating student improvement in writing, only the most sensitive and reliable raters should be used.[24] Following this suggestion, the records of ratings by more than forty staff members of the Department of Written and Spoken English were examined. Nine raters were selected who appeared to have demonstrated a high degree of sensitiveness and reliability. Their average ratings were about the same and the distributions of their ratings were fairly uniform--that is, the ratings of each staff member formed a distribution approaching the normal curve.

---

24. Ibid.

From these nine raters it was possible to obtain three who were willing and able to devote the time necessary to complete the task of rating each of the 320 papers involved during the period between Fall term, 1951, and Winter term, 1952. Thus, instead of having to call in a third rater when there was a disparity between the first two raters, three raters judged each of the 320 papers; all three raters were rating the same papers at the same time and under the same external conditions. Such a procedure should provide for more valid and reliable ratings than would otherwise be possible.

Determining scores: Although not customary for the rating of themes, the use of three raters at a time has been the customary procedure for rating the comprehensive examination speeches in the Written and Spoken English course. It has been the practice to utilize the ratings of the two of the three raters who were in closest agreement. Two exceptions to this rule have been permitted. When no two of the three raters had arrived at ratings within five points of each other, the average of all three of the ratings were used. Also, when the differences among the ratings were equal (such as 22, 24, and 26), the average of all three ratings was used.

It has been the consensus of the 40 to 50 staff members involved in the above procedure that such a practice is valid. Almost invariably the rater who has varied considerably from the other two raters has recognized that some distractive factor caused him to rate too high or too low. Since these reactions are by staff members who have a high sense of responsibility to the student whose grade is being determined for three terms

of his work, there is good reason to believe that the distractions pointed out, in most cases, were real and that the ratings by the other raters were valid.

The procedure described above has resulted in agreement, within five points, by two of the three raters for 94 per cent of the speeches rated. The record on the rating of themes has been about the same. This procedure is fundamentally the same as the one used for the rating of comprehensive examination themes, and the same as that recommended by Diederich. Under this system, however, the third rater does not have to be called in when a discrepancy occurs. His rating has already been made and is available for tabulation and use along with the ratings of the other two raters.

Control of Factors Affecting Raters

Effect of variations in the efficiency level of raters. Since this study is concerned with the assumption that the efficiency level of the writer varies significantly from day to day, a safeguard should be provided against the possibility of a variation in the efficiency level of the raters from day to day.

To provide such a safeguard, three raters were employed to read each of the 320 papers involved; and all three were to rate the same papers at about the same time. That is, each paper was rated by Rater 1, who passed it on to Rater 2, who, in turn, passed it on to Rater 3. In addition, the papers were so organized that all four papers by each student were rated during the same rating period.

Effect of accidental content factors on raters. Since this study is concerned with the assumption that accidental content factors (the topics) provide a source of variation in the writing proficiency of students, the study should provide a safeguard against the effect of accidental content factors on the quality (the reliability) of the raters' work. If, for example, all of the papers on one topic were rated on one day and all of the papers on another topic were rated on another day, there would be no way of determining whether the observed differences were due to variations in the efficiency level of the students or of the raters, or due to the effect of accidental content factors on the raters--or a combination of all of these factors.

A partial safeguard for this problem was provided by having all of the papers by each student rated on the same day by all three of the raters--since each student wrote on two different topics. Half of the students, however, had a third topic for their second paper. To provide a safeguard against the danger of having the effect of accidental content factors of this topic combined with variations in efficiency level (of the raters), the papers were arranged so that the topics would be rated in the following order: Topic 1, Topic 2, Topic 1, Topic 3--with this series repeated throughout the rating of the 320 papers. Since the effect of the content factors on the raters has been isolated, variations between the reliability of ratings on different topics can be attributed to the topics.

Arranging the topics in this order automatically intersperses two students' papers, so that it would be more difficult for the raters to compare different papers by the same student. Although there is no

evidence that such a comparison would seriously affect the objectivity of the ratings, the papers were arranged so that any two papers by a single student were separated by at least two papers by other students.

Effect of fatigue on the raters. Since the fatigue factor was recognized as a possible disturbing influence on the efficiency level at which the students would be writing (a rest period being provided between the writing of the two papers on each day to counteract such an influence), it seemed advisable to take a similar precaution in regard to the raters. Consequently, a six-hour working day was planned. Each day was divided into approximately four equal parts of ninety minutes each. Two such periods were held each morning and two each afternoon, with a 15-20 minute rest period in the middle of the morning and afternoon sessions. This procedure was followed for rating all of the 320 papers involved in the study.

In view of the arrangement of the papers (described above), the effect of any fatigue factors on the raters should be evenly distributed among the papers in such a way as to preclude any serious effect on the factors being tested.

Criteria for Rating Themes

In view of the fact that the three raters employed had been using the same theme rating scale for the past four years to evaluate comprehensive examination themes, it did not seem advisable to make any alterations in that scale--since such alterations might result in a lowering of the reliability of the ratings.

On the other hand, Diederich points out that specific criteria for judging papers relative to a specific assignment should be set after reading at least ten papers written on that assignment.[25] It is important for the raters to reach an agreement on what is expected of the students in developing a paper on a particular assignment. Such agreement concerns the content, primarily, and the organization of the content. Criteria for the sentence structure, grammar and mechanics, and the diction, which make up the remainder of the general criteria listed on the rating scale in use, remain fairly constant for all assignments made in the Written and Spoken English course.

Although a specific assignment for writing a paper on a definite topic may influence the quality of the sentence structure, grammar and mechanics, and diction, the desired quality for these elements of writing remains fairly constant. In contrast, the quality of the supporting material, and the organization of it, have to be judged according to the topic, the materials provided, and, to some extent, the additional knowledge and ideas possessed by the students.

The papers which, by random procedure, had been eliminated from the study in order to equalize the number of students in each group were utilized to determine more specifically the criteria for rating the organization and content for each of the three topics involved.

---

25. Ibid.

Summary

Since it has been demonstrated that the rating method employed by the Written and Spoken English Department can produce highly reliable ratings when proper precautions are taken, and since the rating form employed by that department was designed for evaluating the objectives of that course, that method and that form seemed to provide the best available means for evaluating the papers involved in this study.

Three raters, whose records indicated ability in making sensitive discriminations, were employed to rate each of the 320 papers involved in this study. The papers on different topics were arranged so that the raters received them in alternate order--thus insuring that the efficiency level of the raters entered into the rating of each topic in about the same way. Rest periods were provided at regular intervals in order to reduce the possibility of excessive fatigue affecting the rater's work.

After reading papers which had been eliminated from the study, the raters formulated their criteria for evaluating the quality of writing on different topics--especially concerning organization and content. The three ratings for each paper were tabulated. The score for each paper was determined by computing the average of the two ratings which were in closest agreement, when the discrepancy between those two ratings was not in excess of five points. Otherwise, the average of the three ratings was used.

Thus, the data was provided for determining whether a <u>single paper</u> written on a <u>given topic</u> at a <u>particular time</u> provides a valid basis for evaluating a student's writing ability at any time in a writing course.

CHAPTER V

ANALYSIS OF THE DATA

The purpose of this study was to test four assumptions in order to determine whether a single paper written by a student on a given topic at a particular time can be considered as a valid basis for evaluating his achievement at any time in a writing course.

Since the 80 students involved in this study were only a small sample (about three per cent) of the entire student population enrolled in the course, and since this study was concerned with variations of individual performance of the 80 students who were equally divided among four groups, it was necessary to provide answers to the following questions before definitive tests could be provided for the four assumptions:

1. Were the students involved in this study representative of the entire Freshman population enrolled in the Written and Spoken English course?

2. Were the students in each group representative of the entire population involved in the study?

3. Did fatigue affect the quality of writing on the papers written during the second hour of the two-hour writing periods? And,

4. Were there any factors which affected the average quality of writing for the different groups of students--in relation to topics, days, or pressure?

Background Analysis

Eighty students representative of Freshman population. For the findings of this study to be generally applicable, it must be assumed that the students involved in the study were representative of the entire Freshman population enrolled in the Written and Spoken English course. Although there was no satisfactory method of testing this assumption, the enrollment procedure in use operated to insure that the selection of any four groups of students would provide a representative cross-section of the Freshman population.

That procedure consisted of admitting students in alphabetical order throughout the three-day registration period. In order to equalize the number of students in each of the 93 groups (recitation sections) in which the students were being enrolled at that time, the enrollment in any one group was limited on each day. Thus, in each of the four groups used, the student names were distributed throughout the alphabet--In group A, from B to W; group B, from A to S; group C, from A to S; and in group D, from B to W. Thus, it seems reasonable to assume that the 80 students were representative of the entire Freshman population.

Each group representative of total group. These eighty students were equally divided among four groups of twenty students each. Since this study was to test assumptions concerning the effect of topics, of variations in efficiency, and pressure on the quality of student writing, and since those assumptions could be tested only by comparing the results for different groups, each group of twenty students should be a representative

sample (in relation to their writing ability) of the entire population of students involved in the study--or a method of equating the four groups must be utilized.

Since all eighty students wrote one paper on the same topic ("Distance Between Classes") under the same conditions, a basis was provided for determining whether or not each group of twenty students was representative of the total group. The method of analysis of variance[1] provides a test for the hypothesis that the differences in means for the four groups indicate real differences in the average quality of writing by these groups.

The mean scores of the papers written on Topic 1 ("Distance Between Classes") on the first day by each group are shown below:

Group A	Group C	Group B	Group D
24.5	23.4	25.4	22.4

Application of an analysis of variance produced an F ratio of .7457 (as shown below). This does not approach the F ratio of 8.57 needed for significance at the 5 per cent level. Thus, we can reject the hypothesis that there are significant differences among the four groups in the average quality of writing on Topic 1. Therefore it seems reasonable to conclude that each of the four groups is representative in their writing ability of the total group involved.

---

1. E. F. Lindquist, Statistical Analysis in Educational Research, Houghton Mifflin Company, New York, 1940, p. 91.

Results of Analysis of Variance

Source	d.f.	SS	Variance	F Ratio
Between groups of papers	3	101.5367	33.845	17457
Within groups of papers	76	3449.3100	45.385	

F ratio needed for significance at 5% level.....8.57
F ratio needed for significance at 1% level.....26.27

Effect of the fatigue factor. Although precautions were taken to prevent fatigue from producing variations in efficiency for the students writing papers during the second half of the writing period each day, a test was provided to determine, insofar as possible, whether such fatigue did alter the quality of papers written during the latter half of those periods. Since alternate topics, for each half of the two-hour writing periods, were passed out systematically to every other student in each section, it is reasonable to assume that the students receiving different assignments for each half of the writing period were evenly divided relative to their writing ability.

Thus, with only two groups of students involved (first-hour and second-hour), a t-test may be employed to test the hypothesis that the mean score for the papers written on a specific topic during the first hour is significantly higher than the mean for papers written on that topic during the second hour.[2] Since all students involved in the study wrote two papers on Topic 1, the test may be applied to the difference in means between all of the papers on this topic written during the first hour and

2. Ibid., p. 51.

those written during the second hour. The computation, however, revealed that the mean score for the papers written during the second hour was slightly higher than the mean of the papers written during the first hour (as shown below):

	Means
Papers on Topic 1, written during the second hour.....	23.925
Papers on Topic 1, written during the first hour.....	23.735
Difference.....	.19

Since the size of the means is just the opposite of what would be anticipated relative to the fatigue factors, it seems reasonable to assume that the fatigue factor did not result in a lowering of the quality of the writing during the second hour. Application of the t-test indicated that the difference in the two means would have happened by chance alone between 40 and 50 times out of a hundred cases.

Therefore, it seems reasonable to conclude that neither the fatigue factor nor any other factor relative to the first or second hour of the writing period had any significant effect on the average quality of the students' writing.

Effect of variations in efficiency and of content factors. Under normal conditions, it would be expected that the efficiency level at which a group of students would be writing at a particular time would vary from one individual to another. Some students would be at their peak; some would be at their low point. The remainder would be working at efficiency levels distributed between the two extremes. On another day, under normal conditions, the same type of distribution would be expected, but with many

individuals working at different levels of efficiency than they were on the first day. Thus, the average quality of the writing on both days should be approximately the same, and the variations in efficiency levels would not result in any significant variations in the mean scores of papers written on the same topics on different days.

On the other hand, it may be hypothesized, that if the "practice effect" or "new learning experiences" affected most of the writers in the same manner, then mean scores on different days would be significantly different.

Likewise, if the accidental content factors, provided by the different topics, affected most students in the same manner (that is, if one topic proved more difficult for most students than did the other topic), then that effect would result in a significant difference in the mean scores for the two topics.

In order to test the two hypotheses described above, again a test for a significant difference in means is required. Since each of the 20 students in each group wrote a paper on each of two different topics on each of two different days, four sets of papers were produced by each group: one set written on the first topic on the first day, one set on the second topic on the first day, one set on the first topic on the second day, and one set on the second topic on the second day.

Since groups A and B wrote their papers on both days under the same conditions, with no examination pressure added on the second day, the two factors, content and efficiency level described above, are the primary variables involved.

Although the first topic assigned was the same for both groups, the second topic assigned group B students was not only different from the first topic but was of a different type than the second topic assigned group A students. In order to distinguish between the possible effects of two dissimilar topics, as compared with the effects of two similar topics, an analysis of variance was applied to the four sets of papers for each of the two groups separately--to determine whether a significant difference existed between any of the means of the four sets of papers.[3]

The means for group A are shown below:

First Day		Second Day	
Topic 1	Topic 2	Topic 1	Topic 2
24.5	23.7	21.9	21.3

For group A, the F ratio obtained for 3 and 76 degrees of freedom was .7480 (shown below). This does not approach the F ratio of 8.57, which is needed for significance at the 5 per cent level. Therefore, it seems reasonable to conclude that there is no significant difference between the means of any two sets of papers written by group A.

Results of Analysis of Variance

Source	d.f.	SS	Variance	F ratio
Between sets of papers	3	136.9464	45.6488	.7480
Within sets of papers	76	4637.8100	61.0238	
F ratio needed for significance at 5% level				8.57
F ratio needed for significance at 1% level				26.27

3. E. F. Lindquist, loc. cit.

The means for group B are shown below:

First Day		Second Day	
Topic 1	Topic 3	Topic 1	Topic 3
25.4	21.4	24.1	21.5

For group B, the F ratio obtained for 3 and 76 degress of freedom was .9837 (shown below). Again this does not approach the F ratio of 8.57, which is needed for significance at the 5 per cent level. Therefore, it seems reasonable to conclude that there is no significant difference between the means of any two sets of papers written by group B.

Results of Analysis of Variance

Source	d.f.	SS	Variance	F ratio
Between sets of papers	3	221.4073	77.1350	.9837
Within sets of papers	76	5959.2900	78.4117	

F ratio needed for significance at 5% level..... 8.57
F ratio needed for significance at 1% level.....26.27

In view of the fact that no significant difference was found between any pair of means for either section, it seems reasonable to conclude that: 1) if there were variations in the efficiency level of individual writers from day to day which resulted in significant variations in the quality of their writing, those variations occurred in such a manner as to cancel out any significant effect on the mean scores for groups of writers; and 2) if the accidental content factors (due to different topics) resulted in significant variations in the quality of writing for individual students, again such variations occurred in such a manner as to cancel out any significant effect on the mean scores for groups of writers.

Effect of pressure on mean scores. The students in group C wrote on the same two topics (similar) on each day as did those in group A. On the second day, however, the students in group C had an additional psychological pressure factor added, by being informed that the papers written on that day would be considered as a part of their final examination. The students in group D wrote on the same two topics (dissimilar) on each day as did those in group B. On the second day, however, the students in group D wrote under the same additional psychological pressure as did those in group C.

Again, as with groups A and B, one of the primary variables involved in the writing of the four sets of papers by each of these two sections is the content factor (due to similar topics for group C, and to dissimilar topics for group D). Since it was demonstrated, by the preceding tests, that the variations in efficiency level from day to day did not affect the mean scores on papers written by a group of students, the additional psychological pressure factor of the examination situation may be considered as the second primary variable involved in the writing by groups C and D.

Again, since similar topics were utilized by group C and dissimilar topics by group D, it was deemed advisable to test the significance of the difference between means for each of the two groups separately in order to provide for the possibility of an interaction between the pressure and similar topics and between pressure and dissimilar topics having a significant effect on the results.

Thus, an analysis of variance again was used to determine whether there were any significant differences between the mean scores of each of the four sets of papers written by each group. The mean scores for group C are shown below:

First Day		Second Day	
Topic 1	Topic 2	Topic 1	Topic 2
23.4	23.9	25.0	24.6

For group C, the F ratio obtained for 3 and 76 degrees of freedom was .2700 (as shown below), which does not approach the F ratio of 8.57 needed for significance at the 5 per cent level. Therefore, it seems reasonable to conclude that there was no significant difference between the means of any two sets of papers written by group C.

Results of Analysis of Variance for Group C

Source	d.f.	SS	Variance	F ratio
Between sets of papers	3	34.8690	11.623	.2700
Within sets of papers	76	3271.3500	43.040	
F ratio needed for significance at 5% level.....				8.57
F ratio needed for significance at 1% level.....				26.27

The mean scores for group D are shown below:

First Day		Second Day	
Topic 1	Topic 3	Topic 1	Topic 3
22.4	19.2	23.9	20.9

For group D, the F ratio obtained for 76 and 3 degrees of freedom was 1.48 (as shown below). This does not approach the F ratio of 2.73

needed for significance at the 5 per cent level. Therefore, it seems reasonable to conclude that there is no significant difference between the mean scores of any two sets of papers written by group D.

Results of Analysis of Variance for Group D

Source	d.f.	SS	Variance	F ratio
Between groups of papers	3	245.4900	81.83	1.48
Within groups of papers	76	4199.1650	55.25	
F ratio needed for significance at 5% level.....				2.73
F ratio needed for significance at 1% level.....				4.06

Since the application of analysis of variance failed to show any significant difference in the means of the four sets of papers written by either group C or group D, it seems reasonable to conclude that neither the additional psychological pressure of the examination situation nor, again, the content factors (from the use of similar and dissimilar topics), nor a combination of such factors had any significant effect on the average quality of writing by these students.

There was, however, one other combination of students that could be used to test the effect of the additional psychological pressure on the average quality of writing by a group of students. The students of all four groups wrote on Topic 1 ("Distance Between Classes") on two different days, with the students in groups A and B writing without any additional psychological pressure on either day. The students in groups C and D, however, wrote on Topic 1 on the first day without additional pressure, but on the second day the psychological pressure of the examination situation was added.

Thus, by combining the scores on Topic 1 for groups A and B on each day and for groups C and D on each day, again four sets of papers are provided, the means of which can be tested for significant differences.

At the beginning of this chapter, it was demonstrated that there was no significant difference in the mean scores, among the four groups, for the papers written on Topic 1 on the first day. Also, it was demonstrated that there was no significant difference between any two mean scores of the four sets of papers written by each group. By combining the scores on Topic 1, for each day, by groups A and B and by groups C and D, the number of student papers in each group is double the number in previous groups tested--which should provide a more reliable test for the effect of the factors involved. The means for the combined sections on Topic 1 are:

First Day		Second Day	
Groups A and B	Groups C and D	Groups A and B	Groups C and D
24.9	22.9	22.5	24.5

Again, an analysis of variance was used to determine whether there was any significant difference between any two mean scores for these four sets of papers written on Topic 1. The F ratio for 3 and 156 degrees of freedom was .7957 (as shown below). This does not approach the F ratio of 8.55 needed for significance at the 5 per cent level. Therefore, it seems reasonable to conclude that there is no significant difference between any two of the four means--again demonstrating that the psychological pressure of the examination situation had no effect on the average quality of writing by the groups of students involved. Also (in relation to

groups A and B combined), it has been demonstrated again that variations in efficiency level from day to day did not have a significant effect on the average quality of writing by these students.

Result of Analysis of Variance for Combined Sections

Source	d.f.	SS	Variance	F ratio
Between sets of papers	3	130.6815	43.56	.7957
Within sets of papers	156	8539.4960	54.74	
F ratio needed for significance at 5% level.....				8.55
F ratio needed for significance at 1% level.....				26.25

Effect of Factors on Individuals

The tests thus far reported do not preclude the possibility of significant variations in writing proficiency for any individual or for any number of individuals involved in this study. All that has been demonstrated is that if such variations have occurred, they have occurred in such a manner as to cancel out any effect on the total or mean scores in each instance.

Method of determining significant variations in the quality of writing by individuals. In order to study the effect of the various factors on the quality of the writing by individual students, it was necessary to examine the scores on pairs of papers by the same student and to determine whether the differences between those scores may have been due entirely to the unreliability of the ratings or whether those differences actually indicate variations in the quality of the student's writing.

Although the method adopted for rating the papers in this study is considered by the experts in the field as one of the most reliable and valid methods for evaluating the quality of a student's writing, the scores obtained by that method are not amenable to the usual statistical methods for determining the reliability and the error of the ratings. In most cases, the score assigned each paper was the average of only two ratings. Those two ratings, however, were not always by the same two people. And in some cases, three ratings were averaged in order to obtain the best estimate of the quality of a paper.

Ebel, with the assistance of Professors E. E. Cureton, Harold Gulliksen, and E. F. Lindquist, has developed a procedure for determining the reliability and error for sets of ratings--similar to those used in this study (See Appendix A). Ebel points out that if decisions are to be made in practice by comparing averages which come from different groups of raters, then the "between-raters" variance should be included as a part of the error term.[4] (These requirements are met by Ebel's procedure, which is illustrated in Appendix B.)

Since there was a possibility that accidental content factors due to the topics might influence the raters, as well as the writers, it seemed likely that the reliability of the ratings for each topic indicated that such variations did occur (as shown below). This means that the standard error of an observed score also would vary from topic to topic.

Since the reliability of the ratings varied from topic to topic, the over-all, or total, reliability and standard error of the ratings were

---

4. Robert L. Ebel, "Estimation of the Reliability of Ratings," Psychometrika, vol. 16, No. 4 (December 1951), pp. 407-424.

computed in order to make more accurate comparisons of a student's obtained scores on different topics. The reliability and the standard error of the ratings for each topic and for the total are shown below as computed by Ebel's procedure:

	Standard Error	Reliability
Topic 1	3.2031	.80
Topic 2	3.4583	.77
Topic 3	2.5416	.91
Total	3.1203	.83

In order to compare two obtained scores for the same student, it was necessary to compute the standard error of the difference between those two obtained scores--which, according to McNemar, is found by multiplying the standard error of the obtained score by the square root of two. Furthermore, this product multiplied by 1.96 provides the difference (at the 5 per cent level of confidence) between two obtained scores which is more than is to be expected on the basis of chance, i.e., on the basis of the unreliability of the scores.[5]

In order to obtain as accurate an estimate as possible of significant differences between the scores of a student's papers on two different topics, the standard error of difference between two obtained scores was computed for the total number of ratings involved in this study. And again, in order to obtain as accurate an estimate as possible, the standard error of the difference between two obtained scores on the same topic was computed for each topic separately--since there were some differences

---

5. Quinn McNemar, Psychological Statistics, John Wiley and Sons, New York, 1949, p. 130.

between the reliabilities of the ratings for the papers on different topics.

Each rater recorded a total score for each paper in whole units. The obtained scores for all but 14 of the 320 papers were computed by averaging the scores assigned by two raters; thus the obtained scores for most of the papers were either in whole or half units. There was a discrepancy of more than five points between the ratings for 14 of the 320 papers. For those papers the obtained scores were computed by averaging the ratings recorded by all three raters. For some of those papers the obtained scores were in one-third or two-thirds units.

Since the obtained scores for most of the papers were in whole or half units, and since the theoretical differences, necessary to denote a significant difference in the quality of writing, were in odd decimals, the minimum differences actually used to indicate a significant difference were somewhat greater than the theoretical differences in each case, as indicated below for each topic and for the total.

	Theoretical Difference	Minimum Difference Used
Topic 1	8.88	9.00
Topic 2	9.58	10.00
Topic 3	7.04	7.50
Total	8.65	9.00

Thus, when a difference between a student's scores for two papers on Topic 1 was found to be 9.00 points or more, it seemed reasonable to conclude that a real difference in the quality of his writing was indicated. Likewise, such a conclusion seemed reasonable when a difference in scores of 10.00 points or more was found between two papers on Topic 2, and of

7.50 points or more between two papers on Topic 3, and of 9.00 points or more between two papers on different topics.

Since each student wrote four papers (two on different topics on each of two different days), there were four possibilities for differences in the quality of a student's writing to be noted--differences between the papers on:

1. the first topic on different days,
2. the second topic on different days,
3. the first and second topics on the first day, and
4. the first and second topics on the second day.

With 80 students involved in this study, there were 320 pairs of papers--each of which provided an opportunity to note a significant difference, or variation, in the quality of writing.

The difference between the scores for each of the four pairings of papers by each student was computed and tabulated. Then the differences which equaled or exceeded the appropriate minimum difference, in each case, were checked (See Appendix C)--thus providing a basis for testing the first three assumptions posed in this study. To test the fourth assumption, it was necessary to rank the 80 students according to their writing ability.

Method of ranking students according to writing ability. The 80 students were ranked from high to low--using the highest score obtained by each student out of his set of four papers as a basis for the ranking. The present writer recognizes that to use the average of the four scores as an index for ranking the students would result in greater stability, tending to reduce the effect of the error in ratings.

On the other hand, if the scores on all four papers were averaged in order to obtain an index for ranking students, a number of students might fall at the same rank, even though their best papers were quite different in quality. Since the score on each paper was an average of the ratings reported by two or three raters, the process of averaging the scores on the four papers by each student, in order to obtain an index of his rank, would seem to restrict the usefulness of such ranking for this study. In fact, when all four papers by each student were averaged, as many as eight students fell at the same rank at two different points.

The method used for ranking the students should be consistent with the basic assumptions of this study. As previously indicated, the practice of utilizing a single theme as a representative sample of a student's writing ability for the purpose of evaluating student achievement assumes that there are no significant variations in the quality of a student's writing from day to day or from topic to topic. It also assumes that such stability of writing performance is true for all students, regardless of variations in ability.

If such assumptions were valid, it would make little difference whether students were ranked according to the highest score obtained, or according to the average of all four scores--since the only variations that would occur would be due entirely to errors in rating. Also, such variations would occur just as frequently for the weak students as for the strong students.

In this study, however, it is contended that such assumptions may be false--that significant variations in the quality of a student's writing

may occur not only from day to day and from topic to topic, but that such variations may occur more frequently for strong students than for weak students.

It is contended for example, that if only one topic is disconcerting to a student, and his efficiency level is at a low point on the first day but at a high point on the second day, on the latter day he may produce one paper of high quality and one somewhat lower in quality. His two papers on the first day might be quite low in quality in comparison with his best paper. The score on that best paper should be indicative of the student's writing ability under the most favorable circumstances.

On the other hand, the average of the scores on all four of his papers would provide an estimate of his achievement considerably below that provided by the score on his best paper--and somewhat above that provided by the score on his poorest paper.

Thus the use of the student's highest score, as a method of ranking, would likely produce some distortion due to error in the ratings; while the use of the average of all four scores would produce some distortion due to variations in performance. It seemed advisable, therefore, to rank the students by both methods in order to provide a check on both types of distortion.

Effect of Combined Factors

With the differences between the scores of the four pairings of papers by each student computed and tabulated (see Appendix C), and with the students ranked according to the two methods described above, it was possible to determine whether the use of a single paper written by a

student on a given topic at a particular time provides a valid basis for evaluating his achievement in a writing course at any time.

In order to arrive at that determination, it was necessary to examine the frequency and the degree of significant variations in the quality of writing for the total group of 80 students, which resulted from the combination of factors involved in this study--content, efficiency, and pressure.

Total group. In relation to frequency, there were 86 significant variations in quality of writing out of 320 opportunities for such variations to occur. Since the standard error of the difference between two obtained scores was computed at the 5 per cent level of confidence, no more than 16 such variations could be expected as a result of the unreliability of the ratings. These 86 variations were distributed among 47, or over 58 per cent, of the 80 students involved in this study.

In relation to degree, the average of the variations in excess of the difference attributed to error in ratings was four points--which was equal to the average improvement shown, in a previous study by the writer, for students during a year's training in the Written and Spoken English course. Also, a letter-grade distribution of the scores indicated that 50 variations of two or more letter grades probably were due to significant variations in the quality of writing. (See Appendix E).

Effect of combined factors on strong and weak students. The 80 students were ranked by two methods--by using the highest score for each student, and by using the average of the four scores for each student. After ranking the students by the first method, a re-ranking by the second method resulted in the shifting of only four students from the top 40 to the low 40. However, only 11 of the top 20 remained in that group. On the

other hand, 15 of the low 20 remained in the same group. These shifts in rank provide some evidence indicating greater variations in the quality of writing for the strong students than for the weak students.

Comparing the top 20 students with the low 20 students, by both methods of ranking, produced the following frequencies of significant variations:

	Top 20	Low 20
Using the highest score:	36	8
Using the average score:	22	11

Comparing the top 40 students with the low 40 students, by both methods of ranking, produced the following frequencies of significant variations:

	Top 40	Low 40
Using the highest score:	63	23
Using the average score:	55	31

It should be noted that when the average score is used for ranking the students, the difference between the frequency of significant variations for the top group and the low group is reduced. Apparently this reduction is the result of the averaging process--since the method of averages not only tends to reduce the range of scores for the entire group but it also tends to cancel out the real differences that may exist between papers by the same student. Therefore, it seems that the use of the highest score by each student to determine his rank is justified for the purpose of analyzing the differences between strong and weak students.

Thus, according to the data given above, significant variations in the quality of writing occurred about three times as often among the top 40 students as among the low 40. For the top 20 and the low 20, the ratio

is even larger in the same direction. (Appendix D shows the distribution for each group.)

A more meaningful interpretation is provided by examining the 50 variations of two or more letter grades. Of these, 43 were among the top 40 students while only seven were among the low 40. It seems reasonable to conclude that variations in the quality of writing occur not only more frequently but to a greater degree for the strong students than for the weak students.

Another meaningful approach is to determine the number of students having variations of two or more letter grades, and the frequency of such occurrences relative to the strong and weak students.

A total of 31 students had variations of two or more letter grades. Of these 31 students, 25 were among the top 40 students, while only six were among the low 40.

Of the 25 students in the top 40, 15 had two or more such variations out of four possibilities. Of the six students in the low 40, only two had two or more such variations.

In view of the variations noted in the quality of writing for over 58 per cent of the students involved, and in view of the fact that those variations appeared more frequently and to a greater degree for students who ranked above the median than for those below the median in writing ability, it seems reasonable to conclude that a single paper written by a student on a given topic at a particular time cannot be considered as a valid basis for evaluating his achievement in a writing course at any time, unless that student's writing ability was rather low; and, even

then, a single paper would not provide an infallible basis for such an evaluation.

Comparability of Content Factors and Variations in Efficiency

Although the evidence presented above contradicts the assumption that a single paper can be used for evaluating student achievement in writing, no evidence was presented concerning each of the first three assumptions involved. In each of those three assumptions, a possible source of variations in the quality of student writing was identified: 1) Content factors--from different topics, 2) variations in efficiency on different days, and 3) the psychological pressure of an examination situation.

Since all 80 students wrote on two different topics on each of two different days, a basis was provided for determining the comparability of content factors and over-all variations in efficiency as sources of the variations in the quality of their writing which were noted above. (The use of all four groups for this purpose necessarily included the pressure factor under the variations in efficiency and included both similar and dissimilar topics under content factors. Separate tests isolating the pressure factor and isolating similar and dissimilar topics are presented later.)

If it is assumed that different efficiency levels and different content factors have comparable effects on the variability in the quality of student writing, the expected frequency of significant variations in the quality of writing from each of these sources would be about the same.

Likewise, if it is assumed that high and low ranking students have comparable variations in the quality of their writing from day to day and from topic to topic, the expected frequency of significant variations in the quality of writing by high and low ranking students would be about the same.

Therefore, to test the assumption concerning sources of variation, it was necessary to determine whether the difference between the observed and the expected frequencies of significant variations from different efficiency levels and different content factors may have been due to chance or due to real differences between the effects of the two sources. Likewise, to test the assumption concerning high and low ranking students, it was necessary to determine whether the difference in the observed and the expected frequencies of significant variations for the high and low ranking students may have been due to chance or due to real differences in the variability of their writing.

By tabulating the observed frequency of significant variations in the quality of writing between days on the same topics and between topics on the same days, for the high and low ranking students separately, it was possible to apply the chi-square test[6] to determine the comparability of these two sources of variation in relation to student writing ability. The tabulation of those observed frequencies is shown below:

---

6. (The most appropriate statistical test for hypotheses concerning frequency distributions when each classification or category is dichotomous.) Lindquist, op. cit., pp. 33-47.

	Between Days	Between Topics	Combined Totals
High 20	17	19	36
Low 20	4	4	8
High 40	35	28	63
Low 40	12	11	23
Total group	47	39	86

Total group. The frequency of significant variations totaled 86 for both sources of variability. If the two sources were comparable, the expected frequencies would be 43 for each source. Actually, there were 47 significant variations between days on the same topics and 39 between topics on the same days. Application of the chi-square test indicated, at the 5 per cent level of confidence, that the observed frequencies for the two sources were not significantly different from the expected frequencies. Therefore, it seems reasonable to conclude that the two sources were responsible for comparable variations in the quality of student writing.

Strong and weak students. Likewise, the expected frequencies for the high 40 and the low 40 would be 43 for each group. The observed frequencies were 63 and 23, respectively. Application of the chi-square test indicated that the difference between the observed and the expected frequencies was significant at the 1 per cent level of confidence. In a like manner, application of the chi-square test to the frequencies of high and low ranking students for each of the sources separately, indicated at the 1 per cent level that the difference between the observed and expected frequencies was significant in each case. Therefore, the assumption that variations in the quality of writing by high and low students are comparable can be rejected.

Comparability of Similar and Dissimilar Topics

Although the evidence presented above indicated that content factors and variations in efficiency are equally responsible for variations in the quality of student writing, the content factors involved both similar and dissimilar topics. Since groups A and C wrote on similar topics, while groups B and D wrote on dissimilar topics, it was possible to note the frequency of significant variations in the quality of writing due to each set of topics.

Thus, if it is assumed that similar and dissimilar topics have comparable effects on the variability in the quality of student writing, the expected frequency of significant variations for each set of topics would be about the same. Likewise, if it is assumed that high and low ranking students have comparable variations in the quality of their writing on similar and dissimilar topics, respectively, the expected frequencies of significant variations for high and low ranking students would be about the same.

Therefore, as in the preceding section, to test each of these assumptions it was necessary to determine whether the difference between the observed and the expected frequencies, in each case, may have been due to chance or due to a real difference in the factors involved.

By tabulating the observed frequencies of significant variations in the quality of writing on similar topics and on dissimilar topics, for the high and low ranking students separately, again it was possible to apply the chi-square test to determine the comparability of similar and dissimilar topics in relation to student writing ability. The

tabulation of those observed frequencies are shown below:

	Dissimilar Topics	Similar Topics	Combined Totals
High 20	9	10	19
Low 20	1	0	1
High 40	15	13	28
Low 40	10	1	11
Total group	25	14	39

Total group. The frequency of significant variations totaled 39 for both similar and dissimilar topics. If these two sets of topics produced comparable variations in the quality of writing, the expected frequencies would be 19 1/2 for each set. The observed frequencies were 25 for dissimilar topics and 14 for similar topics. Application of the chi-square test indicated that the difference between the observed and expected frequencies was not significant at the 5 per cent level of confidence. Therefore, the assumption that similar and dissimilar topics have comparable effects on the variability in the quality of student writing cannot be rejected.

Strong and weak students. However, the trend of frequencies from high to low ranking students suggests that although there was no significant difference in the frequencies of 15 and 13 on dissimilar and similar topics, respectively, for the high 40, there was a difference in the respective frequencies of 10 and 1 for the low 40. Also, for the dissimilar topics, the frequencies for the high 40 and the low 40 were 15 and 10, respectively; while on the similar topics the frequencies were 13 and 1, respectively--

all of which suggests that dissimilar topics did account for a considerable number of significant variations for the weaker students, while the similar topics did not.

This interpretation was supported by the application of the chi-square test to the frequencies on dissimilar topics (15 and 10), and to the combined frequencies (28 and 11) for the high 40 and the low 40, respectively. For the frequencies on dissimilar topics, the chi-square found was not significant even at the 30 per cent level of confidence. Yet, for the combined frequencies, the chi-square found was significant at the 1 per cent level.

Therefore, it seems reasonable to conclude that similar and dissimilar topics were about equally responsible for significant variations in the quality of writing by the better students, while only the dissimilar topics were responsible for significant variations in the quality of writing by the weaker students.

Effect of the Pressure Factor

As previously stated, the test for the comparability of content factors and variations in efficiency included the psychological pressure as an element in the efficiency factor.

The third assumption states that the psychological pressure introduced by the examination situation has no adverse effect on the quality of a student's writing. In order to test that assumption, it was necessary to compare the quality of writing by individual students who wrote without pressure on one day and with pressure on another day and then to compare those results with control students who wrote on both days without pressure.

If the pressure factor had no adverse effect, then there should be no more losses and no fewer gains in scores, from the first to the second day, for the pressure students than for the non-pressure students. In other words, if it is assumed that the pressure factor has no adverse effect on the quality of writing, then the expected frequencies of significant gains and losses in scores from the first to the second day should be about the same for the pressure groups, C and D, as for the non-pressure groups, A and B. Likewise, if it is assumed that the high and low ranking students have comparable variations in the quality of their writing with and without the pressure factor, the expected frequencies of significant gains and losses in the quality of writing by the high and low ranking students should be about the same.

Thus, again, to test these two assumptions it was necessary to determine whether the difference between the observed and the expected frequencies, in each case, may have been due to chance or due to real differences in the factors involved.

By tabulating the observed frequencies of gains and losses for non-pressure groups A and B and the pressure groups C and D, divided between the high and low ranking students, it was possible to apply the chi-square test to determine whether the difference between observed and expected frequencies were due to chance or due to the factors involved. The tabulation of those observed frequencies is shown below:

Non-pressure Groups A and B	Gains	Losses	Combined Gains and Losses
High 40	6	11	17
Low 40	3	2	5
Total group:	9	13	22
Pressure Groups C and D			
High 40	10	7	17
Low 40	5	3	8
Total group	15	10	25
Grand Total	24	23	47

Total group. Examination of the total frequencies supports the assumption that the frequencies of gains should be no fewer and the frequency of losses no more for the pressure groups than for the non-pressure groups. Of the grand total of 24 gains, 15 were by the pressure group and 9 by the non-pressure group. Of the grand total of 23 losses, only ten were by the pressure groups, while 13 were by the non-pressure groups. Although the data indicate a trend of more losses than gains by the non-pressure groups on the second day, and more gains than losses by the pressure groups on the second day, application of the chi-square test indicated, at the 5 per cent level, that these observed differences were no greater than would be expected by chance. The data merely suggest that a larger sample might substantiate the trends indicated. Therefore, it seems reasonable to conclude that the psychological pressure of the examination is not likely to have an adverse effect on the quality of writing by a number of students.

Strong and weak students. The frequencies of significant gains and losses for the low ranking students in both the pressure and non-pressure groups

were too small for application of the chi-square test. However, the trends for the high and low ranking students in both groups are quite similar to those for the total group. Therefore, it seems reasonable to conclude that the pressure of the examination had little or no effect on either high or low ranking students.

CHAPTER VI

SUMMARY, CONCLUSIONS, AND IMPLICATIONS

Summary. The purpose of this thesis was to determine whether a single paper written by a student on a given topic at a particular time can be considered as a representative sample of his writing ability--and thus provide a valid basis for evaluating ability at any time in a writing course.

For that purpose, it was necessary to test the following four basic assumptions involved in the use of a single paper for evaluating a student's writing ability--the assumptions that:

1. any given topic provides the same stimulus as any other topic,
2. any given topic elicits constant responses at different times,
3. the psychological pressure introduced by an examination situation has no adverse affect on the quality of writing, and
4. the quality of writing is stable from topic to topic and from time to time for any student regardless of his writing ability.

To test the first assumption, it was necessary to determine whether the assignment of different topics during the same period has any effect on the quality of writing.

To test the second assumption, it was necessary to determine whether the assignment of the same topic on different days has any effect on writing quality.

To test the third assumption, it was necessary to determine whether the assignment of the same topics, with and without the pressure of an examination, has any effect on the quality of writing.

To test the fourth assumption it was necessary to rank the students according to their writing ability and to determine whether the frequency and/or the degree of variations in the quality of writing differs from strong to weak students.

The data for testing those assumptions were obtained by having four groups of students (twenty in each group), enrolled in the Written and Spoken English Department at Michigan State College, write two papers on different topics on each of two different days. Two groups, A and B, wrote under the same conditions on both days--with the understanding that the results would not affect their grades for the course. The other two groups, C and D, wrote their two papers on the first day under the same conditions as groups A and B. On the second day, groups C and D were assigned to write papers on the same two topics, but with the understanding that the grades assigned those papers would count as a part of their final examination for the course. Thus groups A and B provided a basis for determining variations due to topics on each day, a basis for determining variations in writing due to efficiency variations from the first to the second day, provided a control group for determining variations due to the examination pressure on groups C and D the second day.

Steps were taken to assure that the evaluation score assigned each of the four papers written by each of the 80 students was as reliable and valid as could be obtained with current rating methods. These scores were

analyzed to determine whether each group of students was representative of the total group and whether different topics, efficiency variations, or the pressure of an examination had any effect on the average quality of writing by the appropriate groups.

After it had been determined that the four groups were comparable, and that none of the three factors had any significant effect on the average quality of writing, it was then possible to analyze those factors in relation to their effects on different individuals with different writing abilities.

In planning this study, an attempt was made to provide the controls necessary to preclude the intervention of extraneous factors. It seems advisable, however, to restate the following qualifications before presenting the conclusions:

1. The 80 students involved in this study were only a small sample (about three per cent) of the entire student population enrolled in the course. However, there is little reason to believe that these students were not representative of that population. Therefore, conclusions relative to the 80 students can be applied with reasonable confidence to the entire student population.

2. The use of the highest score of each student to rank the students according to their writing ability could provide some distortion. The conclusions presented, however, are based on the analyses of the data on the high 40 and the low 40--and it has been shown that using the highest score, for each student, rather than the average of all four scores, resulted in changing the position of only four students between these two

groups. It would seem, therefore, that the conclusions concerning variations in the quality of writing for strong and weak students should be valid.

3. There was no assurance that the rating method provided scores consisting of equal units. Therefore, the results of the statistical tests may be open to question. In other words, the conclusions based upon a statistical analysis of such scores can not be stated with as much certainty as would be justified if the scores were known to consist of equal units.

Conclusions: From the analyses of the writing performance by the groups of students involved, it seems reasonable to conclude that:

1. Under normal conditions, variations in the efficiency of individual students do not significantly affect the average quality of writing by a group of twenty or more students from day to day on the same topics.

2. When sufficient care is taken to provide topics which are appropriate to the students' background of experiences, the accidental content factors resulting from the assignment of any one of those topics do not affect the average quality of writing by a group of twenty or more students.

3. When students are assigned the task of writing a theme as a part of a final examination, the average quality of the writing under such conditions will be about the same as though the papers had been written as an assignment which would not affect their grades for the course.

From the analyses of the writing performance by individual students in each group, it seems reasonable to conclude that:

1. Content factors due to different topics and variations in efficiency from day to day result in significant variations in the quality of writing for a considerable number of the students involved, and these variations are likely to occur with significantly greater frequency for strong students than for weak students.

2. Content factors and variations in efficiency are about equally responsible for such variations in the quality of students' writing as occur.

3. For the strong students, dissimilar topics are responsible for no more significant variations than similar topics. For the weak students, however, dissimilar topics appear to result in more significant variations in the quality of writing than similar topics.

4. Psychological pressure introduced by having papers written as a part of a final examination, does not result in a significantly greater frequency of gains and/or losses for individual students than are likely to occur when such papers are written as an assignment which will not affect the students' grades. Thus, it appears that such pressure has no adverse effect on a significant number of students.

5. Although significant variations in the quality of writing, either from topic to topic or from day to day, were noted for 47 of the 80 students involved in this study, no such variations were noted for the other 33 students. This does not mean, however, that there were no variations in the quality of writing by any of those students. It can be said, only, that if such variations did occur, they did not exceed the error of the ratings.

On the other hand, it seems reasonable to assume that for some of those students, the two assigned topics provided about equal stimuli--resulting in no significant difference in the quality of writing on those two topics. Likewise, some of those students may have been writing at about the same efficiency level on both days--resulting in no significant difference in the quality of their writing on the same topics on different days.

Implications. In Chapter I, two major problems were raised concerning the use of a single sample of a student's writing as indicative of his writing ability--first, the problem of evaluating his achievement (assigning a grade) at the conclusion of a writing course; and, second, the problem of evaluating individual student improvement from the beginning to the end of a course.

In relation to the first problem, the findings from this study cast considerable doubt upon the justification of the customary practice of using five letter-grades to designate achievement in a writing course when a single paper provides the basis for that designation.

For example, the findings from this study indicate that variations in the quality of a student's writing from topic to topic, or from day to day on the same topic, plus the error in assigning a score for each paper, are likely to result in variations of two or more letter-grades in the evaluation of his writing ability--and that such variations are likely to occur more frequently for strong than for weak students.

Thus, a student might have his writing ability rated as "A" from one paper and as "C" from another paper. However, if only one of those papers

had been written as an indication of his writing ability, he would have received only one of those ratings. It is entirely possible that the "A" rating would not have been the best estimate of that student's writing ability. From the evidence presented in this study, a grade of "B" might have been a better estimate--since error in rating could account for such a difference, even when the most reliable rating methods currently available are used.

On the other hand, if that student's single paper had received a rating of "C", there would be no assurance that a rating of "A" would not have been a better estimate of his writing ability--since a combination of unsuitable topic, or a low efficiency level at the time the paper was written, and error in rating could account for such a difference.

Even if four samples of a student's writing were obtained (on two different topics on each of two different days), perhaps not more than a three-level grading system could be justified--in view of the amount of error in present rating methods, plus possible variations in the quality of writing from topic to topic and from time to time. Since the findings from this study indicate little variation in the quality of writing by the weaker students, an unsatisfactory grade might be assigned from an evaluation of four papers with considerable confidence that no injustice is being inflicted upon the student. Since considerable variations in the quality of writing were noted for the stronger students, perhaps no more than two levels of passing grades could be assigned with confidence in their accuracy.

The second problem raised in Chapter I concerned the evaluation of improvement in the quality of student writing during the period of a writing course. If an evaluation of over-all or average improvement is all that is desired, it can be obtained from a single sample of each student's writing for a pre-test and a post-test, as indicated in Chapter I. Such an evaluation, however, provides little useful information for evaluating the effectiveness of different teaching methods with students who vary in writing ability--since variations in quality of writing, plus error in ratings, will tend to nullify or exaggerate the actual improvement for a considerable number of the students involved.

From the results of this study, it would seem that in order to develop a program for evaluating individual student improvement in writing (for strong as well as for weak students), it would be advisable to obtain several samples of writing by each student--samples of writing on different topics on the same day and on the same topics on different days. And such samples should be obtained for both the pre-test and the post-test. If the topics assigned for the post-test are not to be the same as for the pre-test, then they should be similar in nature--especially for the weak students.

Such a program of pre-testing and post-testing would provide a basis for evaluating the improvement in writing proficiency for both strong and weak students--reducing the possibility that content factors and/or variations in efficiency would affect the quality of writing in such a way as to nullify or exaggerate actual improvement.

Furthermore, it would seem, from the results of this study, that the pressure of the examination situation does not result in more adverse

variations in the quality of writing than when such pressure is absent. This does not mean, however, that such variations will not occur, occasionally. On the other hand, it seems reasonable to assume that, if proper precautions are taken, a post-test may be utilized as a final examination without seriously affecting the results.

Such a program of evaluation, as outlined above, would entail a considerable amount of time and effort on the part of the staff members involved. It seems to this writer, however, that unless such a program is utilized for evaluating individual improvement in a writing course, there is little possibility of improving courses, programs, or teaching methods concerned with that objective. And unless improvement does take place in the effectiveness of courses, programs, or teaching methods concerned with the improvement of student writing, teachers in that area can continue to expect the same criticisms they have received in the past.

BIBLIOGRAPHY

Adkins, Dorothy C., Construction and Analysis of Achievement Tests, U. S. Government Printing Office, Washington, D. C., 1947.

Anderson, H. A., and Traxler, A. E., "The Reliability of the Reading of an English Essay Test", School Review, vol. 48, 1940, pp. 521-30.

Coward, Ann F., "A Comparison of Two Methods of Grading English Compositions", Journal of Educational Research, vol. 46 (October 1952), pp. 81-93.

Cronbach, Lee J., Essentials of Psychological Testing, Harper and Brothers, New York, 1949.

Darsie, Marvin L., "The Reliability of Judgments Based on the Willing Composition Scale," Journal of Educational Research, vol. 5 (January 1922), pp. 89-90.

Diederich, Paul B., "The Measurement of Skill in Writing", School Review, vol. 55 (April 1948), pp. 375-84.

_____, "Readers' Methods Scrutinized in English Essay Scoring," Educational Testing Service Developments, No. 1 (October 1951).

_____, "The Use of Essays to Measure Improvement," College English, vol. 10 (April 1949), pp. 395-99.

Dolch, Edward W., Jr., "More Accurate Use of Composition Scales," English Journal, Vol. 11 (November 1922), pp. 536-44.

Ebel, Robert L., "Estimation of the Reliability of Ratings," Psychometrika, vol. 16, No. 4 (December 1951), pp. 407-24.

Edmiston, R. W., and Gingerich, C. N., "The Relation of Factors of English Usage to Composition," Journal of Educational Research, vol. 36, 1942, pp. 269-71.

Gates, Arthur I., Psychology for Students of Education, The Macmillan Company, New York, 1931.

Gerber, John C., "Testing and Evaluation in the Skills of Communication," College English, vol. 9 (April 1948), pp. 375-84.

Greene, "English--Language, Grammar, and Composition," Encyclopedia of Educational Research, Macmillan, New York, 1950.

Hartung, M. L., and others, "Aspects of Thinking," Appraising and Recording Student Progress, Harper and Brothers, New York, 1942.

Hawkes, H. E., and others, The Construction and Use of Achievement Examinations, Houghton Mifflin Company, New York, 1936.

Hillegas, Milo B., "Scale for the Measurement of Quality in English Composition by Young People," Teachers College Record, vol. 13 (September 1912) pp. 331-84.

Huddelson, Earl, "The Effect of Objective Standards Upon Composition Teachers' Judgments," Journal of Educational Research, vol. 12 (December 1925), pp. 329-40.

Huddleston, Edith M., Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Techniques (Ph. D. Thesis), New York University, 1952, 104 numb. leaves; published as Research Bulletin RB-52-7, Educational Testing Service, Princeton, N. J., 1952.

Lange, Phil C., "A Sampling of Composition Errors of College Freshmen in a Course Other Than English," Journal of Educational Research, vol. 42 (November 1948), pp. 191-200.

Leonard, S. A., "Building a Scale of Purely Composition Quality", English Journal, vol. 14 (December 1925) pp. 760-75.

Lindquist, E. F., Statistical Analysis of Educational Research, Houghton Mifflin Company, New York, 1940.

Lockwood, H. R., Correlation of the Mental Maturity of One Hundred College Freshmen and Their Ability to Write English Composition, (Master's Thesis), University of Chicago, 1925.

Loevinger, Jane, "A Systematic Approach to the Construction and Evaluation of Tests of Ability," Psychological Monographs, vol. 61, No. 4, 1947.

Lyman, R. L., "Investigations in the Field of Written Composition," Summary of Investigations Relating to Grammar, Language, and Composition, Supplementary Educational Monographs No. 36, The University of Chicago, 1929.

McNemar, Quinn, Psychological Statistics, John Wiley and Sons, New York, 1949.

Merrill, Maud A., "The Significance of IQ's on the Revised Stanford-Binet Scales," Journal of Educational Psychology, vol. 29 (December 1938), pp. 641-51.

Mowrer, O. H., "Learning Theory," Review of Educational Research, vol. 22, No. 5 (December 1952), pp. 475-95.

President's Commission on Higher Education, "Establishing the Goals" (vol. 1), Higher Education for American Democracy, U. S. Government Printing Office, Washington, D. C., 1947.

Sackett, Leroy W., "Comparable Measures of Composition," School and Society, vol. 5 (February 24, 1917), pp. 233-39.

Sims, Verner M., "The Essay Examination as a Projective Technique," Educational and Psychological Measurement, vol. 8, 1948, pp. 19-20.

_____, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," Journal of Educational Research, vol. 24 (October 1931), pp. 216-23.

_____, "Reducing the Variability of Essay Examination Marks Through Eliminating Variations in Standard of Grading," Journal of Educational Research, vol. 26 (May 1933), pp. 637-47.

Stalnaker, John N. "The Essay Type Examination," Educational Measurement, American Council on Education, Washington, D. C., 1951.

_____, "Question IV, The Essay," English Journal (College Edition), vol. 26 (February 1937), pp. 133-40.

Starring, Robert W., "A Study of Ratings of Comprehensive Examination Themes When Certain Elements Are Weakened," (unpublished Ed. D. Thesis), Michigan State College, 1952, 109 numb. leaves.

Stroud, James B., Psychology in Education, Longmans, Greene, and Company, New York, 1946.

Thorndike, E. L., The Elements of Psychology, The Mason-Henry Press, Syracuse, New York, 1907.

_____, Mental Work and Fatigue and Individual Differences, Teachers College, Columbia University, New York, 1923.

Thorndike, R. L., "Reliability," Educational Measurement, American Council on Education, Washington, D. C., 1951.

Trabue, M. R., "Supplementing the Hillegas Scale," Teachers College Record, vol. 18, (January 1917) pp. 51-84.

Traxler, A. E., and Anderson, A. A., "The Reliability of an Essay Test in English," School Review, vol. 43 (September 1935), pp. 534-39.

Tyler, Ralph W., "Purposes and Procedures of the Evaluation Staff," Appraising and Recording Student Progress, Harper and Brothers, New York, 1942.

Virtue, John B., "The Proficiency Examination in English Composition at the University of Kansas," College English, vol. 9 (January 1948), pp. 199-203.

APPENDICES

APPENDIX A

THEME RATING SCALE AND DEFINITIONS OF SEPARATE CATEGORIES

	Superior 10	9	8	7	6	5	4	3	2	Unsatisfactory 1
Content										
Conventions of Grammar										
Sentence Structure										
Diction										
Organization										

Content

Content refers to the quality and adequacy of the substantiating material (examples, statistics, arguments) employed in support of ideas expressed in the paper. A theme of college caliber should concern itself with matter worthy of adult consideration and express a reasonably mature point of view.

Conventions of Grammar

Conventions of grammar refers to such matters as reasonable spelling, correct punctuation at major junctures, the usual grammatical agreements

(subject-verb, pronoun-antecedent), and the correct use of possessives. It refers also to the avoidance of sentence fragments, comma faults, period faults, and dangling modifiers.

Sentence Structure

Effective sentence construction means the strategic use of such things as the periodic sentence, subordination, and parallelism. It means that by a variety in sentence length, in sentence structure, and in sentence order, monotony and childishness of expression may be avoided and variety and maturity of expression achieved. It means sentences which are free from awkwardness and obscurity. It means that successful attention has been given to the requirements of sentence euphony and rhythm.

Diction

Good diction means the use of words well chosen to express the writer's meaning. It means the avoidance of expressions which are crude or trite, of wordiness, of pompousness. It means the use of accepted idioms, of expressions which are vigorous and alive, of the specific and concrete in preference to the general and abstract.

Organization

The size of the topic should fit the length of the paper. The theme as a whole should have a single, controlling idea or purpose, expressed or clearly implied, to which each part of the theme contributes. Each paragraph should be recognizable as a unit (i.e. developing a single topic or sub-topic) in the development of the theme. The ideas presented should

be smoothly and logically linked together. Such linking is achieved by a recognizable pattern of development and by the use of such transitional devices as the connective, parallelism, pronoun reference, and repetition. By the use of such things as position, proportion, and repetition, that which is of most importance in the theme should be made to seem so to its reader.*

---

* Written and Spoken English Syllabus, Michigan State College Press, East Lansing, 1949, pp. 17-18.

APPENDIX B

ESTIMATION OF THE RELIABILITY FOR INCOMPLETE SETS OF RATINGS*

Whether or not it is desirable to remove "between-raters" variance in estimating the reliability of ratings depends upon the way in which the ratings are ultimately used in grading, classification, or selection. In any case where differences from rater to rater in general level of rating do not lead to corresponding differences in the ultimate grades, classifications, or selections, the "between-raters" variance should be removed from the error term. Specifically, the "between-raters" variance should be removed where the final ratings on which decisions are based consist of averages of complete sets of ratings from all observers, or ratings which have been equated from rater to rater such as ranks, Z-scores, etc. Likewise, if comparisons are never made practically, but only experimentally, between ratings of pupils by different raters, the "between-raters" variance should be removed. But if decisions are made in practice by comparing single "raw" scores assigned to different pupils by different raters, or by comparing averages which comes from different groups of raters, then the "between-raters" variance should be included as part of the error terms.

---

* Robert L. Ebel, "Estimation of the Reliability of Ratings," Psychometrika, vol. 16, No. 4 (December 1951), pp. 411-412.

APPENDIX B - Continued

TABLE 2

ANALYSIS OF RATINGS FOR PROBLEM 5--INCOMPLETE SETS

	Ratings		k	Sum
Pupil 1	8 6 4 4 3		5	25
Pupil 2	6 9 9 4 9 6 5 10 8		9	66
Pupil 3			3	23
	Sums		17	114
Sum of squared ratings				858
Sum of products (pupil sum times pupil mean)				785.3333
Product of sum and mean				764.4706
Sum of squares				
For total	858 -- 764.4706	=		93.5294
For pupils	785.3333 -- 764.4706	=		20.8627
For error	93.5294 -- 20.8627	=		72.6667
Mean square				
For pupils	20.8627 + 2	=		10.4314
For error	72.6667 + 14	=		5.1905
Average value of k				
Reliability	$\frac{10.4314 - 5.1905}{10.4314 + (4.1176)(5.1905)}$	=		.1648

Table 2 illustrates application of this formula to a simple problem in which the table of ratings is incomplete and the sources of ratings are not identified. In this case only two components of the variance, attributable to pupils and error, are separated. Thus any difference in general level of ratings between the various raters is automatically included in the error term.

APPENDIX C

DIFFERENCES BETWEEN PAIRS OF PAPERS

(Significant differences are underlined. Notations equal: T1 - Topic 1, T2 - Topic 2; D1 - First day, D2 - Second day; W/O - Without pressure; W - With pressure. Significant gains and losses from first day to second day are indicated by plus and minus marks.)

Group A

STUDENT NUMBER	T1 D1 W/O - D2 W/O	T2 D1 W/O - D2 W/O	D1 W/O T1 - T2	D2 W/O T1 - T2
2	12.0 +	14.1 -	4.1	22.0
3	6.5	8.5	4.5	2.5
4	14.0 -	5.0	3.5	15.5
5	3.0	3.0	1.0	5.0
6	2.0	2.0	0.0	0.0
8	1.0	5.5	0.0	4.5
9	6.5	8.0	18.0	3.5
10	5.5	5.0	1.0	0.5
11	1.0	3.0	2.0	4.0
12	6.0	0.0	0.5	5.5
15	1.5	11.0 -	5.5	4.0
16	6.0	10.7 +	2.7	14.0
17	2.7	10.5 -	8.0	0.2
18	4.5	5.0	7.0	7.5
19	11.5 -	1.5	4.0	6.0
21	4.0	0.0	1.5	2.5
22	13.0 +	1.5	2.0	16.5
23	3.0	1.5	5.5	1.0
24	13.0 -	22.5 -	9.5	0.0
25	6.5	5.0	6.0	5.5

APPENDIX C - Continued

	Group C			
	T1	T2	D1 W/O	D2 W
	D1 W/O - D2 W	D1 W/O - D2 W	T1 T2	T1 T2
26	1.0	2.5	5.0	3.5
27	8.0	2.0	2.5	3.5
28	5.0	1.5	2.0	1.5
29	6.0	2.0	7.5	3.5
30	2.0	7.0	8.0	1.0
31	13.5 +	2.5	4.5	4.5
32	20.5 +	9.0	7.5	22.0
33	0.2	3.5	8.3	5.0
35	1.0	2.5	7.5	4.0
36	8.0	2.5	5.5	0.0
37	2.5	6.0	9.0	12.5
38	3.0	3.0	0.0	6.0
39	2.5	1.5	2.0	6.0
40	11.5 -	7.0	0.5	3.0
41	4.5	7.5	12.0	0.0
42	6.5	0.5	3.0	9.0
43	15.0 +	5.0	1.0	11.0
44	16.4 -	17.5 +	28.0	5.9
45	2.0	1.0	0.5	3.0
46	11.5 +	2.5	6.0	8.0

APPENDIX C - Continued

	Group B			
	T1	T3	D1 W/O	D2 W/O
	D1 W/O - D2 W	D1 W/O - D2 W	T1 - T2	T1 - T2
48	3.0	0.5	13.0	9.5
49	2.5	7.0	6.5	11.0
51	0.0	4.0	5.5	1.5
52	4.5	8.5 -	5.5	9.5
54	6.8	5.0	4.5	2.7
55	19.5 -	4.0	21.5	2.0
56	0.5	5.0	5.5	1.0
57	2.5	0.0	8.0	5.5
58	8.5	4.5	11.5	7.5
59	2.5	11.5 +	12.5	3.5
61	10.2 -	3.5	14.0	7.3
62	5.5	1.5	1.5	4.5
63	2.5	7.5 +	7.0	2.0
64	0.0	0.7	3.7	3.0
65	10.0 +	1.0	13.0	4.0
67	4.0	9.5 +	6.0	7.5
68	9.0 -	2.5	8.5	2.0
69	13.0 +	9.5 -	6.0	16.5
70	1.0	0.0	6.5	5.5
71	7.0	10.5 -	0.5	4.0

APPENDIX C - Continued

	Group D			
	T1	T3	D1 W/O	D2 W
	D1 W/O - D2 W	D1 W/O - D2 W	T1 - T3	T1 - T3
73	4.0	4.0	1.0	1.0
74	1.0	6.0	9.0	4.0
75	10.5	0.0	6.0	16.5
76	10.0	8.5 -	0.5	18.0
77	9.0	0.5	4.5	4.0
79	9.5 -	0.5	11.5	2.5
80	2.0	1.5	1.5	2.0
81	4.0	13.5	11.5	2.0
83	13.0	8.0 -	6.0	15.0
84	3.0	0.0	2.0	1.0
86	0.5	15.0 -	0.5	15.0
87	11.0	10.5	2.0	1.5
88	10.5 -	8.0 -	8.0	10.5
89	2.5	0.5	11.5	14.5
90	4.5	15.5	19.0	1.0
92	6.5	3.5	2.5	7.5
93	9.5 -	2.0	1.5	9.0
94	8.0	7.5 -	5.0	10.5
95	5.0	15.5	11.0	0.5
96	0.0	13.5	4.0	9.5

APPENDIX D

FREQUENCY OF SIGNIFICANT VARIATIONS FOR STRONG AND WEAK STUDENTS

Note: T1 - Topic 1, T2 - Topic 2, T3 - Topic 3;
D1 - 1st Day, D2 - 2nd Day;

D1-D2 - Difference between papers on topic indicated for different days

T1-T2,3 - Difference between papers on different topics for days indicated

Top 40

Top 20, in descending order, presented on this page.

Second 20 of top 40 continued on next page.

Low 40

Low 20, in ascending order, presented on this page.

Second 20 of low 40 continued on next page

Student Number	ABOVE MED. High 20 T1 D1-D2	T2,3 D1-D2	D1 T1-T2,3	D2 T1-T2,3	Student Number	BELOW MED. Low 20 T1 D1-D2	T2,3 D1-D2	D1 T1-T2,3	D2 T1-T2,3
56	-	-	-	-	73	-	-	-	-
83	13.0	8.0		15.0	25	-	-	-	-
22	13.0	-	-	16.5	63	-	7.5	-	-
55	19.5	-	21.5	-	80	-	-	-	-
64	-	-	-	-	57	-	-	-	-
4	14.0	-	-	15.5	62	-	-	-	-
88	10.5	8.0	-	10.5	30	-	-	-	-
44	16.4	17.5	28.0	-	26	-	-	-	-
42	-	-	-	9.0	21	-	-	-	-
48	-	-	13.0	9.5	74	-	-	9.0	-
32	20.5	-	-	22.0	11	-	-	-	-
43	15.0	-	-	11.0	84	-	-	-	-
58	-	-	11.5	-	86	-	15.0	-	15.0
61	10.2	-	14.0	-	52	-	8.5	-	9.5
75	10.5	-	-	16.5	92	-	-	-	-
9	-	-	18.0	-	3	-	-	-	-
37	-	-	9.0	12.5	5	-	-	-	-
69	13.0	9.5	-	16.5	68	9.0	-	-	-
2	12.0	14.1	-	22.0	6	-	-	-	-
10	-	-	-	-	49	-	-	-	11.0

APPENDIX D - Continued

Student Number	Top 40 (Cont'd.) Second 20 of top 40 in descending order. T1 D1-D2	T2,3 D1-D2	D1 T1-T2,3	D2 T1-T2,3	Student Number	Low 40 (Cont'd.) Second 20 of low 40 in ascending order. T1 D1-D2	T2,3 D1-D2	D1 T1-T2,3	D2 T1-T2,3
16	-	10.7	-	14.0	51	-	-	--	-
17	-	10.5	-	-	81	-	13.5	11.5	-
96	-	13.5	-	9.5	22	13.0	-	-	16.5
24	13.0	22.5	9.5	-	28	-	-	-	-
27	-	-	-	-	70	-	-	-	-
71	-	10.5	-	-	77	9.0	-	-	-
38	-	-	-	-	95	-	15.5	11.0	-
87	11.0	10.5	-	-	12	-	-	-	-
15	-	11.0	-	-	35	-	-	-	-
65	10.0	-	13.0	-	54	-	-	-	-
76	10.0	8.5	-	18.0	18	-	-	-	-
93	9.5	-	-	9.0	79	9.5	-	11.5	-
19	11.5	-	-	-	36	-	-	-	-
40	11.5	-	-	-	33	-	-	-	-
29	-	-	-	-	46	11.5	-	-	-
45	-	-	-	-	90	-	15.5	19.0	-
59	-	11.5	12.5	-	8	-	-	-	-
94	-	7.5	-	10.5	39	-	-	-	-
41	-	-	12.0	-	67	-	9.5	-	-
31	13.5	-	-	-	89	-	-	11.5	14.5

Note: Students numbered from 2 to 25 and from 48 to 71 wrote on both days without the examination pressure. Students numbered from 26 to 46 and from 73 to 96 wrote on second day with that pressure added. Students numbered from 2 to 46 wrote on similar topics; while students numbered from 48 to 96 wrote on dissimilar topics.

APPENDIX E

LETTER-GRADE VARIATIONS

The mean and standard deviation of the 320 scores were used to form a distribution (approximately normal) of letter grades from A to F for all of the papers. Then a check was made to determine how many letter-grade variations may have been due to error in ratings and how many were due to significant variations in the quality of writing.

Most of the variations of one letter grade may have been due to error in ratings, while most of the variations of two or more letter grades probably were due to significant variations in the quality of writing--as shown below:

Variations of one letter grade, which may have been due only to error in the ratings	105
Variations of one letter grade, which probably were not due only to error in ratings but which occurred as a result of significant variations in the quality of writing	33
Variations of two letter grades which may have been due to error in ratings	5
Variations of two or more letter grades which probably were not due only to error in ratings but which occurred as a result of significant variations in the quality of writing	50