

Gunn, Alfred H., "Preparing Teachers of English to Teach All the Children of All the People," College English XVI (March, 1955) pp. 352-356.


-Papers, Quoted
TABLE OF CONTENTS

I. The Preparation of This Report .......................... 1

II. Suggested Methods of Research .......................... 5

   Rating Compositions ........................................... 6
   The writer variable ........................................... 6
   The assignment variable: the topic—the mode of discourse —the time afforded for writing—the examination situation ........................................... 7
   The rater variable: personal feelings—rater fatigue ............. 10
   The colleague variable: a common set of criteria—practice rating ........................................... 11

   Frequency Counts ............................................. 15
   Clarifying examples for each type of item ................. 16
   Standard classification of types of items .................... 16
   Control or sampling of compositions according to topic, mode of discourse, and writer characteristics ............. 17
   Need for analyses of rhetorical constructions ......... 18
   Need for imaginative approaches to frequency counts .... 18
   Counting types of responses by various kinds of writers to various types of situations ........................................... 19
   Reporting frequency per hundred or thousand words .... 20
   Using the cumulative-average technique of sampling .......... 20
   Focusing investigation on narrower, more clearly defined areas and exploring them more thoroughly and carefully ........................................... 21
   Seeking key situations which are indices of larger areas of concern ........................................... 21

   General Considerations ..................................... 21
   Attitude of the investigator .................................. 22
   Meaning of terms and measures: clarity of terms and measures—direct observation—validity of assumptions —reliability of criterion application ........................................... 23
   Planning of procedures: planning before initiating research —using appropriate and consistent statistical procedures ........................................... 24
   Controlling of variables: selection of teachers and students —control of "outside influences"—control of additional influences ........................................... 25
   Need for trials and checks .................................. 26
III. The State of Knowledge about Composition

Environmental Factors Influencing Composition
Primacy of the writer's experiences
Influence of socioeconomic background
Composition interests
Flow of words
Need for case studies
Need for longitudinal studies
Instructional Factors Influencing Composition
Student correction
Frequency of writing
Student revision
Nature of marking and grading
Ineffectiveness of instruction in formal grammar
Rhetorical Considerations
Distinctive tendencies of good writers
Organizational factors
Effects on readers
Objective Tests versus Actual Writing as Measures of Writing

Interlinear tests
"Self-evident" invalidity of objective tests
Unreliable grading of compositions
Reliable grading of compositions
More on invalidity of objective tests
Reliability of objective tests
Varying emphases in college instruction
Use of objective tests for rough sorting of many students
Basing diagnosis of individual needs on actual writing
Evaluating writing from several compositions
Other Considerations
Size of English classes

IV. Summaries of Selected Research

Basis for Selecting These Studies
Explanation of Statistical Terms
The Buxton Study
The Harris Study
The Kincaid Study
The Smith Study
The Becker Study

V. References for Further Research

Summaries and Bibliographies
Indices and Abstracts
Bibliography for This Study
THE PREPARATION OF THIS REPORT

Reading a report, like driving over a bridge, is an act of faith—faith that the other fellow has done his job well. The writers of this pamphlet do not ask that the reader's faith be blind. To permit him to evaluate their work, they explain in this chapter the procedures resulting in their generalizations. The explanation also provides an opportunity to acknowledge the assistance rendered by colleagues throughout the United States and in Canada and England.

The impetus to prepare this report came from the Executive Committee of the National Council of Teachers of English. Concerned over the nature of public pronouncements about how writing should be taught—the sound and the wild seem to share space equally in the press—the Executive Committee appointed an ad hoc Committee on the State of Knowledge about Composition "to review what is known and what is not known about the teaching and learning of composition and the conditions under which it is taught, for the purpose of preparing for publication a special scientifically based report on what is known in this area." The membership of the ad hoc committee is named on the title page.

In April, 1961, the committee met in Washington to clarify the purposes of its task and to plan procedures. It agreed, among other things, to limit its task to written composition and, more particularly, to studies in which some actual writing was involved (not studies entirely restricted to objective testing and questionnaires). The committee further decided to use only research employing "scientific methods," like controlled experimentation and textual analysis. At the suggestion of the Executive Committee, the ad hoc committee set as its goal the identification of the dozen or so most soundly based studies of the foregoing type. (Actually, the committee finally identified five such studies, each of which is summarized in detail in Chapter IV.)

First instructed to complete the manuscript in six to eight months, the ad hoc committee soon realized that a review of "all" the research on composition was a prodigious undertaking which would necessitate a
much longer period of preparation. Consequently, as it began its task, the chairman of the committee applied to the Office of Education, U. S. Department of Health, Education, and Welfare, for a Cooperative Research Program grant. A grant was awarded in the amount of $15,345, supplemented by an allocation of $4,907 from the University of Iowa.

Before the grant was approved, the ad hoc committee had surveyed some 20 summaries and bibliographies (Dissertation Abstracts, Psychological Abstracts, Review of Educational Research, etc.) for titles of studies which seemed pertinent. From more than 1,000 bibliographic citations discovered by the committee, enough apparently tangential references were eliminated to reduce the number to 485 items, which were typed in a dittoed list late in the summer of 1961. The problem then was to screen the studies to determine which should be read carefully.

Because about half of the 485 studies were unpublished, the assistance of colleagues on other campuses was requested. Whenever three or more dissertations from a single campus were on the list, the services of a colleague on that campus were solicited to read the studies and advise the committee on whether or not to study them more carefully. The following people helped in this fashion:

Richard S. Beal, Boston University
Margaret D. Bickle, The Ohio State University
Francis Christensen, University of Southern California
Robert W. DeLancey, Syracuse University
Wallace W. Douglas, Northwestern University
David Dykstra, University of Kansas
Margaret Early, Syracuse University (then visiting Teachers College, Columbia University)
William H. Evans, University of Illinois
Donald J. Gray, Indiana University
Catherine Ham, University of Chicago
Arnold Lazarus, Purdue University (then University of Texas)
V. E. Leitch, Michigan State University
William McColly, University of Wisconsin
John C. McLaughlin, University of Iowa
George E. Murphy, The Pennsylvania State University
Leo P. Ruth, University of California, Berkeley
George S. Wykoff, Purdue University

THE PREPARATION OF THIS REPORT

The large majority of the 485 studies remained, of course, and these were apportioned among the members of the ad hoc committee to screen. To encourage careful screening, each person was requested to fill out a three-page questionnaire for each study he recommended.

Between the number of manuscripts recommended and the number so far inaccessible because of location on other campuses (some of them mimeographed reports not in libraries) several hundred items were still to be read. It was at this point, in the spring of 1962, that funds from the Office of Education and University of Iowa became available, providing the time and money needed to order unpublished material through interlibrary loan and to purchase microfilms, to draw together the findings and to write the pamphlet. Under the provisions of the Office of Education grant, the main responsibility for the project had to be focused in one university. Consequently, a director and two associate directors on the University of Iowa faculty were released from some of their ordinary responsibilities to accomplish these tasks—Richard Braddock, associate professor of English and Rhetoric; Richard Lloyd-Jones, associate professor of English; and Lowell Schoer, assistant professor of Educational Psychology. The grant made it possible to obtain the services of two special consultants—Alvina Tretz Burrows, consultant in Elementary Education; and Porter C. Perrin, consultant in Rhetoric, who died before his invaluable experience could be utilized.

By the end of the summer, 1962, it was possible to construct a list of studies which so far had passed the screening procedures. The directors had not had time to rescreen all recommended studies, and some items were added to the list which no one had yet examined. This list of some 100 studies was submitted to research specialists with a request for additional titles which might have been overlooked or perhaps too hastily screened. The following specialists suggested over fifty new titles to consider as well as some mimeographed bibliographies which the directors did not systematically screen:

Paul B. Diederich, Educational Testing Service
Carl J. Freundreich, New York State Education Department
Robert M. Gorrell, University of Nevada
S. I. Hayakawa, Editor, Etc.
Ernest Horst, University of Iowa
Arno Jewett, U. S. Office of Education
Walter V. Kaulfers, University of Illinois
Albert R. Kitzhaber, University of Oregon
II.

SUGGESTED METHODS OF RESEARCH

Hearing about the project of which this report is the result, a colleague wrote, "What is the sense of attempting an elaborate empirical study if there is no chance of controlling the major elements in it? I think . . . that the further we get away from the particularities of the sentence, the less stable our 'research' becomes. I do not for that reason think there should be no study and speculation about the conditions for teaching composition and about articulation, grading, and the like, but I do think that it is something close to a mockery to organize these structures as though we were conducting a controlled experiment."

Certainly there is much truth in that statement, especially if one takes it as a comment on the bulk of the research which has been conducted thus far on the teaching of written composition. But research in this area, complex though it may be (especially when it deals with the "larger elements" of composition, not merely with grammar and mechanics), has not frequently been conducted with the knowledge and care that one associates with the physical sciences. Today's research in composition, taken as a whole, may be compared to chemical research as it emerged from the period of alchemy: some terms are being defined usefully, a number of procedures are being refined, but the field as a whole is laced with dreams, prejudices, and makeshift operations. Not enough investigators are really informing themselves about the procedures and results of previous research before embarking on their own. Too few of them conduct pilot experiments and validate their measuring instruments before undertaking an investigation. Too many seem to be bent more on obtaining an advanced degree or another publication than on making a genuine contribution to knowledge, and a fair measure of the blame goes to the faculty advisor or journal editor who permits or publishes such irresponsible work. And far too few of those who have conducted an initial piece of research follow it with further exploration or replicate the investigations of others.

Composition research, then, is not highly developed. If researchers wish to give it strength and depth, they must reexamine critically the
his writing performance. Some investigators have maintained that variations in the day-to-day writing performance of individual students "cancel each other out" when the mean rating of a large group of students is considered. But this assumption is false if Kincaid's finding is true that the performance of good writers varies more than the performance of poor writers; the mean rating of the single papers from each of the good writers would not reflect their typically good writing as closely as the mean rating of single papers from poor writers would reflect their typically good writing. The importance of this realization is emphasized by the fact that annual increments in the level of writing performance have usually been reported as small—as approximately one point on a rating scale reaching from 1 to 30, or as 5 percent. Especially then, if an investigator wishes to measure individual students' improvement in writing, he should provide for at least two writing occasions as a pretest, at least two as a post-test, and count the rating only of the better composition on each occasion. If three writing occasions are used for each test, it may be wisest to average the ratings of the two best papers, but more research needs to be done on this possibility.

The Assignment Variable
A second variable—one which can be controlled but often is not—is the assignment variable, with its four aspects: the topic, the mode of discourse, the time afforded for writing, and the examination situation.

Significant variations in the writing performance of eleven-year-olds who wrote on different topics four months apart have been discovered, for example, by Wiseman and Wrigley. The children had a choice from the same set of five topics each time, but the second time had to select a topic different from the one they wrote on the first time. Evidently the investigators assumed that variations in quality of writings were associated with variations in topics not because of the topics themselves but because of the writers' abilities or the raters' idiosyncrasies. Although Wiseman and Wrigley attributed "the bulk of differences in title means [average rating for all papers written on the same topic, or title] ... to the ability of the children rather than to the idiosyncrasies of the mark-

---

2Paul Diederich wrote in 1946 that about one-fourth of a group of University of Chicago students with high scores on their SATs were in a group that wrote a second essay copy but that less than five percent changed their marks as a result of writing a second essay copy. See his "The Measurement of Skill in Writing," School Review, LIV (December, 1946), 535-547. However, in a recent comment on the draft of this report, Diederich stated that two themes are "totally independent."

3One of these considerations has been drawn from Joseph W. Miller's "An Analysis of Freshman Writing at the Beginning and End of a Year's Work in Composition" (Unpublished Ph.D. dissertation, University of Minnesota, 1953).


ers," only four raters were involved and it cannot be determined how representative they were of raters in general. Until more conclusive research has been conducted, it seems safest to select topics with care when rating compositions for purposes of research. Wiseman and Wrigley concluded that examinees might as well be given a choice of topics; the practice of the College Entrance Examination Board suggests that a single topic should be used, controlling the effects of the topic on the quality of the writing. But, whichever practice is correct, it seems very advisable when using compositions as pretests and post-tests to consider carefully the abstractness of the topics and their familiarity to the entire group of examinees. In planning composition examinations for students from a wide range of backgrounds, it seems especially necessary to consider the students' variations in intellectual maturity, knowledge, and socioeconomic background. The national examiner is not adequately controlling the topic who blithely assigns the single subject "My Vacation" or "Civil Defense," forgetting that many students may have been too poor to have had a vacation or too engrossed in farm or school activities to have learned anything about civil defense. Finally, investigators should be mindful of a possible motivational factor in the topic assigned. How many students will write their best when asked to deal with hackneyed topics like "My Vacation" or "My Autobiography"? Some investigators have even instructed students to "Write on anything you wish. It does not matter what you write, but write until you have produced 350 words." Surely there must be some stimulating factor in a topic and, if possible, in the writing situation, too, if the writing they trigger is to have any significance for research.

Another aspect of the assignment variable is the mode of discourse—narration, description, exposition, argument, or criticism. Largely ignored by people doing research in composition, variations in mode of discourse may have more effect than variations in topic on the quality of writing. Although Kincaid concluded that the writing performances of poor writers varied significantly according to the topic assigned, the fact was that his three writing assignments were very similar as topics but called for different modes of discourse.6 His conclusion may well be reinterpreted, then, to suggest that variation of the assignment from expository to argumentative mode of discourse did not seem to affect the average quality of the writing of a group of freshmen who were better writers as much as it did a group who were worse writers. At least until such time as

---

6Kincaid, op. cit.

... is the time afforded for writing. A number of studies purport to evaluate, among other things, the organization of writing when the examinees were afforded but twenty or thirty minutes to produce an essay. Although such a brief time may be sufficient for a third grader writing a short narrative on a familiar topic, it seems ridiculously brief for a high school or college student to write anything thoughtful. Even if the investigator is primarily interested in nothing but grammar and mechanics, he should afford time for the writers to plan their central ideas, organization, and supporting details; otherwise their sentence structure and mechanics will be produced under artificial circumstances. Furthermore, the writers ordinarily should have time to edit and proofread their work after they have come to the end of their papers. It would be highly desirable to discover, through research, the optimum amounts of time needed by students at various levels of maturity to write thoughtful papers. Until such research has been conducted, investigators should consider permitting primary grade children to take as much as 30 to 30 minutes, intermediate graders as much as 35 to 50 minutes, junior high school students 50 to 70 minutes, high school students 70 to 90 minutes, and college students two hours. These somewhat arbitrary allocations of time doubtless should be adjusted according to the upper limits of the range in intellectual maturity of the students and to the topic and mode of discourse of the writing assignment.

A fourth and final aspect of the assignment variable is the examination situation. The situation becomes uncontrolled if the students in the experimental group all write their papers on Wednesday morning and the students in the control group write theirs right after lunch on Wednesday (when many feel logy), or the first thing on Monday (when they are still emerging from the spell of the weekend), or on Saturday morning (when they resent having to forfeit some of their weekend, even for the glory of experimentation). The time, conditions of lighting and heating, and perhaps even the popularity of the teachers proctoring the examination should be equivalent for experimental and control groups or, if improvement is being evaluated, for pretests and post-tests. Obviously the instructions given to the students should be the same, too—preferably written beforehand and read aloud to the students to prevent
the inadvertent intrusion into the instructions for one group of a remark which may stimulate them more or less than the other group.

The Rater Variable

A third major variable in rating compositions is the rater variable—the tendency of a rater to vary in his own standards of evaluation. Any teacher recognizes how variable his own rating can be if he has dug some old papers out of a file, covered the grades, and regraded them without unusual care. Some of the variation may be the result of having forgotten the nature of the old assignment or the emphasis he had been making with the students back then. Although those sources of variability do not function when rating compositions for purposes of research, other familiar sources may operate and should be controlled. They may be characterized as personal feelings and rater fatigue.

Certainly the anonymity of the writer should be preserved to prevent the personal feelings of the rater from coloring his evaluation. That is, in a controlled experiment it should not be possible for the rater to determine from the paper in front of him whether it was written by a student in an experimental or control group. Even though the rater may not recognize the bias himself, he may be hoping that better results are obtained for one group than the other. If the rater may associate with a given group the name of the writer or of the school, the number of the class or section, or even when it was administered, such identifying features should be removed before the papers are turned over to the rater. One way to insure anonymity is to have the students write such identifying information on a 3 x 5 card numbered with the same number as the theme paper but separated from it before the themes are submitted to the raters. Even then the numbers of the material used in the experimental groups should be so mixed with the numbers used in the control groups that the raters do not associate a continuous series of numbers with any group.

In an experiment using pretest and post-test compositions, it may be desirable not to reveal which test is which. If such an experiment is intended at all to measure improvement, concealing the identity of pretest and post-test papers is essential. Not only are the procedures mentioned above essential, but additional steps must be taken to disguise the time at which the papers were written. Students should be requested not to reveal the present year or season in what they write, and papers which do refer to "the falling leaves," "the superintendent's recent speech to the graduating class," or any other such revealing incident should be removed from the compositions to be evaluated. All the paper for both tests should be purchased and prepared at the same time to insure that differences in paper stock and printing will not be apparent. Pretests after they are written and post-tests before they are written should be wrapped lightly in brown paper and stored in the dark to prevent yellowing. The numbering of pretests and post-tests should be mixed. If the pretests become wrinkled, yellowed or musty, the post-tests should be conditioned in the same manner before being submitted to the raters. To overlook some simple identifying feature which permits the personal feelings of raters to operate may render useless all the other efforts which have gone into an experiment.

The rater variable should be controlled further by allowing for rater fatigue. Fatigue may lead raters to become severe, lenient, or erratic in their evaluations, or to emphasize grammatical and mechanical features but overlook the subtler aspects of reasoning and organization. Consequently, raters should not be permitted to rate late at night or for lengthy periods during the day, and they should have regular rest periods to help them maintain their efficiency. Even so, the papers should be placed in a planned sequence which does not permit more of the compositions of one group than another to be rated during a period of probable vigor or fatigue. If pretest and post-test compositions are being rated for experimental and control groups, the four types of papers must be mixed and staggered throughout the entire rating period on each day. When several readers rate the same paper (not individual dittoed or photocopied versions), no rater should place any marks on a paper; they might influence a subsequent rater. Because there are many elements which need control in the sequence of papers, it seems highly desirable to have all of the raters working in the same or adjoining offices, where the investigator can be present and, without entering into the rating himself, insure that everything runs smoothly.

The Colleague Variable

A fourth and last major variable to be considered here is the colleague variable—the tendency of several raters to vary from each other in their evaluations. The existence of this inter-rater variation has been substantiated very frequently by research. As is explained in "Objective Tests versus Actual Writing" in Chapter III, ratings of the same compositions by different raters have been found to correlate from as low as .31 to as high as .96. Consciously or unconsciously, raters tend to place different values on the various aspects of a composition. Unless
they develop a common set of criteria about writing and unless they practice together applying those criteria consistently, raters may be expected to persist in obtaining low agreement.

A common set of criteria seems essential in coping with the colleague variable; if raters are not evaluating for the same qualities, they cannot be expected to rate with validity or reliability. Three principal means of achieving this commonality are composition scales, a general impression method of rating, and an analytic method.

Some forty years ago, composition scales were in wide use to standardize rating. A scale was a carefully selected set of compositions, ranging in quality from, for instance, 1 to 10. A rater would compare the paper before him to the ten sample compositions in the scale, assigning the rating of the sample composition closest in general quality to the paper in question. (The Smith study summarized in Chapter IV made use of two such scales.) The common difficulty with composition scales, however, is that the paper before the rater is seldom closely like any one of the sample compositions or that the rater notices certain similarities in which he is especially interested and overlooks or minimizes dissimilarities in other aspects of the writing. Furthermore, different scales were needed for different modes of discourse and different levels of maturity. It is easy to see why infrequent use is made of composition scales in research today. There has been a resurgence of interest in scales lately published by universities and state boards, but these graded compositions seem to be designed more to help secondary school teachers develop some commonality of practice in ordinary classroom grading or to stimulate them to approximate college standards, not to help investigators rate themes for research purposes.

The two principal means of seeking valid and reliable ratings despite the colleague variable are the "general impression" method of rating compositions and the "analytic method." In the general impression method, a number of raters, working independently, quickly read and rate each composition, the mean of their ratings being used as the final rating of each paper. According to Wiseman's procedure, four raters independently rate each paper, each rater "keeping to a rate of about 50 per

hour" to insure that he makes up his mind quickly. Wiseman has frequently reported reliabilities in the lower .90's for raters using the general impression method for the English 11+ examinations. But the topics he reports seem to call generally for narrative writing, and the purpose of the rater is "to assess the ability of the candidate to profit by a secondary education." The general impression method may not be as effective a means of reducing the colleague variable when argumentative papers, written by older students, are being rated for research purposes.

In the analytic method, two or three raters independently assign a number of points to each of several aspects of a composition and total the points to obtain an overall rating, which is then averaged in with the overall ratings of the other raters. More time-consuming than the general impression method and hence more expensive if two or more raters are used, the analytic method does have the advantage of making clear the criteria by which the rating is done.

In a comprehensive research into four different methods of rating compositions, Cast found the general impression and analytic methods more reliable than the other two and the analytic method slightly superior to the general impression method. Acknowledging that, when used by a trained and experienced rater, the general impression method may correct the errors to which "a crude, mechanical, quantitative dissection might inevitably lead," she concluded that the analytic method, "though laborious and unpopular, appears almost uniformly the best" and that the unreliability of rating "can evidently be greatly reduced by standardized instructions and by the training of examiners."

A caution must be made about the analytic method, however. The criteria used in an analytic method must be clearly defined. In one scheme, the general effect is that half of the total rating is ill-defined:

<table>
<thead>
<tr>
<th>Quantity, Quality, and Control of Ideas</th>
<th>50 marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>15</td>
</tr>
<tr>
<td>Grammar and Punctuation</td>
<td>15</td>
</tr>
<tr>
<td>Structure of Sentences</td>
<td>10</td>
</tr>
<tr>
<td>Spelling</td>
<td>5</td>
</tr>
<tr>
<td>Handwriting</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>100 marks</td>
</tr>
</tbody>
</table>

6 Stephen Wiseman disagrees with this view in "The Marking of English Composition in Grammar School Selection," British Journal of Educational Psychology, XIX (November, 1949), 206: "Indeed, it is arguable that, provided some minimum of similarity exists among the compositions in terms of diversity of viewpoint, the use of many different sets of compositions is not an objection to interrater reliability, since it offsets any such diversity of viewpoint by means from different angles, and the total mark gives a truer 'all round' picture." But this argument seems to contain a difficulty; one would not be sure that lack of high intercorrelation was the product of diversity of viewpoint or the product of erratic marking.

7 Ibid., p. 206.

8 B. M. D. Cast, "The Efficiency of Different Methods of Marking English Composition," British Journal of Educational Psychology, IX (November, 1939), 257-269, and X (February, 1940), 296-304.

To turn that analytic scheme into a meaningful system, one would have to divide or define in more detail the first category in the list. Although different in emphasis because designed for the writing of college freshmen, the theme examination criteria used at the University of Iowa seem to offer a better balance of considerations, especially when they are seen in the light of the three-page set of instructions defining each category:

- Central Idea and Analysis .................. 1.5 points
- Supporting Material .......................... 1.5
- Organization .................................. 1.5
- Expression (diction and sentence style) ... 1.5
- Literacy (grammar and mechanics) ........ 1.3

Total Possible ................................... 5-23 points\(^{11}\)

There is a danger in any analytic system that a beginning rater will first establish the total number of points according to his general impression of a composition's merit and then apportion the total points among the various categories so that they add up to the total. Such a practice, of course, undermines the basis of the analytic method and shows the need for what Cast called "the training of examiners."

Some substantiation of the importance of practice rating was provided by Stalmaker, who had an undisclosed number of college English instructors carefully reread a composition examination after a period of training. He found that rater reliability on the first reading was as low as .30 and never as high as .75 but that, after training, the reliabilities on the second reading ranged from a low of .73 to a high of .98 with an average of .88.\(^{12}\) Although the unusual nature of the examination (it included the construction of an outline and the revision of sentences, among other things) prevents Stalmaker's study from constituting conclusive proof of the efficacy of rater training for the grading of compositions, his findings are reinforced by the frequency with which rater training is reported in studies achieving high reliabilities. A caution must be offered, however. Even though raters are requested to consider in their evaluations such attributes as content and organization, they may permit their impressions of the grammar and mechanics of the compositions to create a halo effect which suffuses their general ratings. (A converse emphasis, of course, can just as easily create the halo.) Evidence of such

---

11The 5 represents "A," 4 "B," and so on to 1 "F." If a student receives an "F," in any one of the five categories, his paper fails.
12John M. Stalmaker, "The Construction and Results of a Twelve-Hour Test in English Composition," School and Society, XXXIX (February 17, 1934), 218-224.
15"Guided Practice Upon Students' Skill in Written Expression" (Unpublished Ph.D. dissertation, Stanford University, 1934).
the most important types, as exemplified by the Harris study. The importance of the frequency count (in contrast to rating procedures) lies in its potential for describing a composition in fairly objective terms which can mean the same things to most teachers and investigators and which are subject to more statistical analyses than are ratings. The frustration comes from confusion over the purpose of such studies and from failure to use methods meaningful to other investigators. A review of some of the methods used may clarify the point. Suggestions for improving the value of such studies are placed in italics.

Many investigators have counted and reported the total numbers of errors of various types which they have found in a collection of compositions. Usually, the errors they have sought have been errors in grammar, usage, and mechanics. If an investigator is seeking examples of pronoun disagreement, for instance, he replies not to every instance he sees an infraction of the rule he has in mind. One difficulty with many such error counts is that the reader does not know what “rule” the investigator has in mind. Is he counting as an error “Everybody went back to the classroom and got their books”? Or does he accept that construction as a nonerror? Does he count “It’s me” as a nonerror, an error in pronoun agreement, a problem in the predicate nominative, a failure in case agreement, or simply an example of “poor diction” or even “unidiomatic usage”?

It is essential for the investigator to give clarifying examples for each type of item he is counting. But even then the reader may feel some hesitation about the results; it is very difficult in a few examples to reveal clearly the many decisions which must be made in classifying instances of disputed and changing usage.

The more thorough the investigator, the more he may subdivide types of errors into lesser categories. Some error counts distinguish among more than 400 types, while others may divide the same problems into but 50 types. Such variation makes it impossible to compare one study, to another or to synthesize their results. If frequency count studies are to be useful to other investigators, then, they should be based on a standard classification of types of items. There is no generally accepted standard classification at this time.

Thirty years ago, one writer constructed a composite list of “the most common grammatical errors,” drawing from 33 previous error

---

---

SUGGESTED METHODS OF RESEARCH

counts. The absurdity of the list is apparent today not only because the categories of the 33 studies had been different but because the counts were made from compositions on various topics, in differing modes of discourse, and by children and adults of widely varying maturity and ability who came from various dialect regions and socioeconomic backgrounds. The composite list even lumped together error counts of oral and written language. It is appropriate to ask what the purpose of such a list would be. If it were to help English teachers and curriculum-makers determine which features of grammar need to be taught, then the frequency count should be conducted from the writing of the pupils to be taught, or from pupils similar to them at the same grade levels. If it were to help determine which types of grammatical items should go into a college English placement test, then probably the frequency count should be based on the writing of a cross section of freshmen or upperclassmen at the types of colleges in which the test will be used. If the count were to be used to establish national norms in the development of written exposition from grade 1 through grade 12, then the compositions would have to be selected from among expository papers written by “average writers” at each grade level and sampled from groups of various socioeconomic backgrounds, amounts of writing practice, and geographical areas.

The studies by Kincaid (summarized in Chapter IV) and by Wise- man and Wrighle demonstrated that the topic a person writes on affects the caliber of his writing. Seger has indicated that a person’s sentence structure is affected by the mode of discourse he is using—argumentation, exposition, narration or description. It also seems apparent that a pupil’s rhetoric, syntax, and usage vary to some degree with his general ability, experience in writing, maturity, socioeconomic background, and native geographical area. Consequently, before conducting a frequency count or using the results of one, a person should determine what his purpose is and then ascertain that the compositions used are appropriately controlled or sampled according to topic, mode of discourse, and characteristics of the writers.

A fundamental difficulty with most frequency counts is that they are simply counts of grammatical and mechanical “errors,” omitting attention to purpose and main idea, supporting material, organization, and style. Even though, in the “summary, conclusions, and implications”

---

10Wise and Wrighle, op. cit.
SUGGESTED METHODS OF RESEARCH

the basis for selecting his sample of readers and he accepted their simple statements about which articles they had read and found satisfaction from, Haskins’ article does give one more confidence in the Gillie formula. A study by Anderson attempted to validate several frequency count instruments. Although Anderson points out that his own use of 150-word samples of writing was a weakness in his study, he does show, for instance, that the widely known LaBrant subordination index does not work well if not applied under carefully prescribed conditions.

One way to break from the grip of error counting is to count the frequency of certain types of situations and the ways in which writers of various kinds respond to those situations. For instance, instead of merely counting what he happens to consider errors in the “these kind of things,” “these kinds of things,” “this kind of thing” expression, the investigator would do well to tabulate the frequency of each of the ways in which writers meet this situation (as Thornblad did[4]) and to seek correlations of the type of response and the type of writer (age, amount of experience in writing, general writing ability, socioeconomic background, and geographical area). Not only would such data help determine what usage label could be attached to each type of response, but, unlike counts of errors, the data would be meaningful even when usage is disputed or when notions of “correctness” have changed since the study was conducted. Such descriptions of actual usage would be more soundly based than the questionnaire approach used by many others who merely asked people which of several expressions they used.[5]

The reporting of frequency counts has often been meaningless or confusing because of the way in which the data have been expressed. The earliest counts seem merely to have reported the total number of errors found in the writing examined. All things being equal, if a person tabulated apostrophe situations in 200,000 words, he would find twice as many situations as if he had examined 100,000 words. To overcome this difficulty, some investigators reported their results by listing the errors in rank order of frequency. But this procedure had two shortcomings. It hid the actual frequency behind the rank; a reader could not tell whether an error of the first rank was much more prevalent or barely more prevalent than an error of the second rank, etc. It also hid the actual frequency in cases where many errors increased or decreased

by chapter of his thesis, the investigator expresses regret at the impossibility of counting rhetorical elements, the impact is often unfortunate; the study has distracted the investigator, his major professor, and readers of the report from the “larger elements of composition.” It is obvious that soundly based counts are needed of the frequency of various grammatical, word, and mechanical usages; but even more urgently needed are similar analyses of rhetorical constructions.

Imaginative approaches to frequency counts are needed. The tendency in any frequency count is to find what one is looking for. More investigators need to initiate frequency studies with fresh questions in mind, not merely attempting to find new frequencies of old “errors.” Some psychologists have been trying new approaches. Kimoto, for instance, explored the relations between dominance-submissiveness characteristics and grammatical constructions. She asked a number of subjects how they would respond in each of several situations in which their own tendencies to dominate or submit would be tested. After recording their oral responses, she counted the frequency of such grammatical features as the passive voice and discovered a number of interesting things. Although her study is not very germane here, it does exemplify an approach which may open up new dimensions in the teaching and learning of composition. Investigations have also been made, using frequency counts, into the degree of abstractness of writing, the correlative of egocentricity, simply abstract index to linguistic maturity, and the increased use of subordination with maturation. These studies have all tended to be exploratory in nature, attempting to develop new instruments for the analysis of language. The worth of such instruments becomes better known, of course, when other investigators attempt to validate the instruments. For instance, Haskins validated the Gillie abstraction index by measuring the degree of abstraction of the articles in an issue of the Saturday Evening Post and then comparing the reactions of a “nationwide sample of readers (N = 340)” to the abstraction of the articles. Although he did not explain


SUGGESTED METHODS OF RESEARCH

reach a relatively stable point for most of the frequency situations being considered.

The suggestions made thus far for the improvement of frequency counts all seem to add up to one thing—more tedious work for the investigator. This is not necessary. Instead of conducting fishing expeditions in a morass of 400 types of items, more investigators employing frequency counts should focus their studies on narrower, more clearly defined areas and explore them more thoroughly and carefully. The same amount of effort should be employed in more intensive analyses of more limited problems.

Another means of efficiency should be explored. Instead of counting many different types of items to study a larger area of concern, investigators should seek to discover certain key situations which are indices of larger areas of concern. For example, would not the usage of irregular English verbs provide an index to the general level of usage of primary school children? Does the subordination index truly provide an index to a broader aspect of linguistic development in writing, as the Strickland study shows in speech? Over the years and through the cumulative efforts of many investigators, if a number of key indices can be developed, frequency counts may become a very efficient means of studying written composition.

General Considerations

The consideration of "Rating Compositions" and "Frequency Counts" dealt in detail with two concerns unique to research in written composition. Here some more general suggestions will be offered on designing and reporting research in composition, drawn up as the writers of this report, especially the specialist in educational research, noted the strengths and weaknesses of the studies being reviewed. Helpful in writing this section were articles by Carroll,13 Dolch,14 Rivlin,15 and Singleton.16

The following discussion is not addressed to investigators as a substitute for formal study of research design and statistics; rather the dis-

[References are cited and relevant passages are numbered accordingly.]
cussion is intended to introduce the reader of research to some basic considerations to have in mind as he interprets and evaluates reports.

**Attitude of the Investigator**

One day a doctoral candidate consulted the director of a large freshman English program. The graduate student had heard that the professor had "a lot of data" on the two thousand freshmen who went through the program each year, and he wondered if the professor would permit him to use it. "I'd like to do something on reading," the student explained, "on the effects of reading on composition." He elaborated: "If I could get the grades the students receive in your courses and compare them to the amount of reading they report having done in high school, I'd be able to tell how much effect high school reading has on composition and be able to recommend that the high schools stimulate more reading."

The student had made a number of assumptions which needed— and received—questioning. Here are some of them:

1. That course grades in freshman English reveal the quality of a student's writing. (But a student's course grade may have been lowered for failure to do assignments, for poor work in the reading aspect of the course, for failure to take part in enough class discussion, for being personally obnoxious, for being in the class of a severe grader, etc. The grade could have been raised for doing unassigned work, for having other students write his papers for him, or for the converse of the other matters mentioned above.)

2. That these college freshmen are generally typical of high school students. (But many high school students have ability, finances, or motivation too low to bring them to college. Furthermore, high school graduates attending this particular university doubtless are not representative of all high school graduates who are college freshmen.)

3. That the freshmen can remember and report accurately the amount of reading they had done in high school. (But many freshmen might, consciously or unconsciously, report more reading than they actually did.)

4. That "amount of reading" is a clear concept. (But it is not clear whether or not "amount of reading" refers to number of books regardless of size, number of pages, books well understood or

books vaguely understood, magazines and newspapers as well as books, books checked out of the high school library whether or not read, etc.)

5. That "amount of reading" is equivalent to "reading." (But the nature of the things read, of the attitude the students held toward their reading, or of their comprehension of the reading may be more important than the amount of reading they did.)

6. That a correlation between amount of reading and quality of writing would reveal a causal relationship. (But these two matters could be caused by some third, unknown factor, or several factors, such as the intellectual atmosphere of the students' homes or school requirements that certain amounts of reading and writing be done.

7. That he knew what the results would be before he had made the investigation. (But such a bias could color his entire investigation, interpretation of results, and recommendations based on the results.)

It is clear that research must be carefully designed if it is to be effective. Basic to a good design is the honest desire to discover or test some generalization about which the investigator does not believe he is fully informed, to discover or test some answer to a sincere question. Coupled with that honest search for knowledge should be a rather antithetical unwillingness to believe anything without being shown, moderated a little by the realization that some things cannot be shown as conclusively as others.

**Meaning of Terms and Measures**

Basic to honesty is clarity. Terms and criteria may mean nothing in the abstract. It should be clear what they represent. If a composition is being rated in part for "fluency," for example, the meaning of that term should be made clear. It could refer to the number of words a student writes, the speed with which he writes, writing without correcting or adding elements, or even writing so the reader proceeds smoothly from one idea to the next. Terms and criteria should be defined carefully, preferably in an operational manner, permitting others to use the terms and criteria with the same results. It is not enough, for example, to refer to the amount of predication a student uses or whether "glanced covertly and winked suggestively at him" represents one predicate or two.
RESEARCH IN WRITTEN COMPOSITION

Behavior should be studied by direct observation and measurement, not only by indirect methods. It is not enough to ask a student to identify a logical fallacy in the writing of others if the problem is whether or not the student uses good logic in his own writing; he should be given an assignment which directs him into writing where his own valid or fallacious logic will be displayed. Similarly, it is not enough to ask a student whether or not he constructs a written outline before he writes; it would be much preferable to give him scratch paper with his theme paper and examine both afterward to see whether the student has written any notes which can be termed an outline. Investigation should not depend upon such indirect measures as multiple-choice or true-false tests, questionnaires, self-inventories, and the like.

Statistical analyses in composition research are based upon criterion measures about which certain assumptions must be made. The nature of these assumptions should be made clear, and there should be fairly adequate evidence that the assumptions are valid and that the criterion measures can be applied reliably. For instance, if an investigator wishes to measure originality of diction by determining how many words in a composition are not on a list of 10,000 most frequently used words, he must be prepared to demonstrate that "infrequently used words" usually yield diction which recognized authorities would characterize as "original." Furthermore, he must define what is meant by "not on the list" so that other investigators can analyze the same compositions with the same results; he must explain, for instance, how to count a word which has a different suffix from a word on the list.

Planning of Procedures

As the anecdote about the graduate student showed, good research generally does not consist of an after-the-fact description or mere correlation study of the results of behavior over which the investigator had no real control. (Correlation studies are necessary, of course, but the investigator should be wary of accepting a correlation as a cause.) He should plan his groups and treatments before he gathers his data, specifying before beginning his study what statistical procedures he will use. Then he should be reasonably consistent in his use of those procedures. He should not, for example, switch criterion measures in different parts of an experiment when there seems to be little basis for the change, as did the investigator who used the subordination index as a measure of language development from grade four to grade eight but then depended upon sentence length as his measure through grades nine to twelve. The

SUGGESTED METHODS OF RESEARCH

statistics need not be complicated, but they should be appropriate. For instance, the progress of a range of students should not be examined only by mean scores, when average gains may be achieved merely by speeding up one end of the range; distributions of scores should be examined.

Controlling of Variables

If the investigation entails the comparison of one method of instruction to another, all variables other than the method should be controlled—the personality, knowledge, and experience of the teacher; the mental ages, writing proficiency, and socioeconomic and intellectual home backgrounds of the students. The attitudes of teachers and students should be controlled (that is, selected) before the experiment and measured after it. 23 It is frequently better to control such variables than to choose teachers and students at random and hope for the best. If students at the extremes of a range of proficiency on one measure are used, it is wise to allow for a regression effect in other areas; that is, just because a group of generally good writers is being contrasted to a group of generally bad writers does not mean that the two groups are far apart in spelling ability or in attitudes toward reading. In fact, the likelihood is that they will be closer on all other measures than they are on the one used to differentiate them. It is better to select subjects than to use volunteers, because only certain kinds of people are willing to volunteer. Moreover, the students should be chosen in such a way that they represent some meaningfully defined student population; otherwise, the results of the experiment cannot validly be generalized beyond those involved in the experiment. Enough students and teachers should be involved that unobserved, uncontrolled factors may cancel out each other.

Other "outside influences" should be controlled or otherwise accounted for—the time of day classes meet, motivation by classroom guests or rewards, size of classes, and demands upon time and initiative. A startling example of outside influences was noted in an experiment in which 23 fourth and fourth graders were being taught to typewrite. The effect of this instruction on written composition was being compared to the effect of no such instruction on the written composition of a control group. The experimental boys and girls came to a university campus during the summer, were instructed in a room newly redecorated for

23A constructive and understanding attitude toward an experiment may be generated and pitfalls may be avoided if the investigator explains the experiment beforehand to the teachers and students involved. Mr. King's suggestions and utilized them whenever possible. Of course the same treatment here must be afforded experimental and control groups.
the purpose and furnished with new desks adjusted to each child's height. A new portable electric typewriter was provided for the use of each child, the typewriters of different pastel colors being switched from child to child each two weeks. Four assistants were in the room to help the professor who taught the class, among other things demonstrating typewriting finger and wrist movement at the piano keyboard. Newspaper publicity covered the experiment throughout that area of the state. Meanwhile, what was happening to the 24 third and fourth graders in the control group? The published report did not say. This experiment illustrates not only the influence of "outside" factors but of the "Hawthorne Effect," the added stimulation received by an experimental group when a new method is being compared to an old method (or, more likely in this case, no method).

Some additional influences should be watched for. If students are doing two or more things in a sequence, the sequence effects must be controlled. For example, if oral and written samples of the same narrative are being collected, some students should first speak and then write, and other students should first write and then speak. If a procedure or instrument is being used which would not be employed in a regular teaching situation (such as a kymograph, recording on a moving drum the starts and stops of a student's writing), steps should be taken to insure that the atypical element did not affect the outcome of the experiment. And, finally, the influence of time and disuse should not be ignored, as it usually is in composition research. Often a follow-up measure should be taken, months or even a year after a new method has been tried, to see how learning stands up, for experimental and control groups, when instruction and practice lie in the past. And then the investigator should take steps to determine if one group has advanced through further study or practice, while the other group retrogressed through disuse, and to determine whether or not these differing post-experiment behaviors were generated by attitudes developed during the experiment.

**Need for Trials and Checks**

This cursory review of elements in the design of composition research should make one point clear: there are many elements to measure and control, and unexpected influences which can spoil a carefully designed experiment. For that reason, it is very prudent to have a trial run before actually beginning an experiment. Different teachers and students must be used for the trial, of course, to prevent the trial itself from affecting the subjects in the main experiment. The complexity of composi-
Finally, if the report is a thesis or dissertation which will be available on microfilm, the investigator should make certain allowances for the medium used. He can assist the reader by including an abstract at the beginning of the manuscript and making the table of contents detailed enough to permit ready location of subsections of chapters. Furthermore, he should avoid the unnecessary use of pages too wide to read in the microfilm reader and the use of color in graphs which become meaningless in the black and white medium.

III.

THE STATE OF KNOWLEDGE ABOUT COMPOSITION

Some months ago, one of the writers of this report mentioned to a colleague doing research in internal medicine that it was disappointing to see how little was really known about the teaching and learning of written composition, how inconclusive most of the research has been. The colleague replied that 95 percent of the research in his area was inconclusive or trivial. "Keep at it," he said. "As you learn more, you'll slowly learn to define your problems in a useful manner and to refine your techniques of analysis. Then you'll be in a position to learn something substantial." His emphasis on the importance of research goals and procedures is reflected in this report; much of what is known about composition teaching is actually known about the procedures of research—and has been considered in Chapter II.

The purpose of Chapter III, however, is to review what is known about the teaching and learning of written composition, as distinct from research procedures. Although the emphasis of this chapter is on the highly selected research summarized in Chapter IV, reference is also made to numerous other studies which merit attention. The findings of these secondary studies should not be taken as conclusive, however. Sometimes the studies make claims which go far beyond the statements extracted here. Other times the investigators base their conclusions entirely on objective testing rather than on actual writing, or they themselves do the teaching in their experiments or rate the themes on which their results are based. Such cautions are not intended to disparage the secondary studies so much as to remind the reader that there is more to be discovered on the subject, as the authors of these studies often state in their own reports.

Environmental Factors Influencing Composition

The primacy of the writer's own experiences has been explored in two interesting, similar studies—by Anderson and Edmund—of the


relationship between prior experiences and short story writing. In each case the investigator discovered, by means of a questionnaire, which short stories (or the settings and characters employed, in the Anderson study) were based on derived experience. Anderson's study used 42 short stories written out of class by freshmen—many of them war veterans—in the investigator’s college composition class, and Edmund’s study was drawn from the in-class writing of 90 seventh graders in three suburban New York State schools. Although both investigators were involved in evaluating the stories (Anderson did all the evaluating; Edmund evidently was one of the three judges) and certain other problems cast doubt on the significance of the results, still the general problem and some of the procedures of the studies merit further investigation. With these cautions, it is interesting to note that Anderson’s college freshmen veterans seemed to write better narration when they based it on direct experiences, Edmund’s seventh graders when they drew their ideas from such derived experiences as television programs and motion pictures.

McClellan also conducted a study of “creative writing,” but with a different purpose and at a different level. His data were 200 papers randomly selected from those of 1,005 children in grades 3 through 6. Of particular interest is the investigator’s study of the socioeconomic level of the children. They were attending three schools representative of the upper, middle, and lower economic classes, and he found that, with almost every factor studied, the higher the socioeconomic level the better the performance.

An indirect key to environmental factors influencing composition is the composition interests of school pupils. If we know what they want to write about, we may be able to approach them on their own ground more readily. There is much research on the reading interests of school pupils, but little has been done on their writing interests. Although now too old to provide reliable clues to writing interests, Coleman’s study affords some suggestive conclusions to those who may wish to attempt an investigation of current interests. The weaknesses of Coleman’s study—and his may be the best of those few done—reflect the difficulty of this type of research: (1) The teachers were not requested scrupulously to avoid influencing their pupils’ choices of topics for writing, (2) The papers were classified as to interest evidently by title only, not

by the actual content of the papers. (3) Most basic of all, distinct and mutually exclusive classifications were not developed, suggesting that it was often arbitrary in just which classification of interest a composition was placed. For instance, “war” was not one of the categories used. Of those used, under which would a war story be placed—“Adventure,” “Contemporary Famous People,” “Historic Events, Sites, or Characters,” “Outdoor Activities,” “People,” “Sympathy,” or “Social Problems”? (4) The pupils’ preferences for forms of discourse were sought by questionnaire only, not by examination of the papers they wrote, and the nature of some items on the questionnaires may have tended to encourage some types of responses and discourage others. (5) The investigator did not attempt in any systematic way to discover the intensity of the pupils’ interests or the relation of their interests to such factors as IQ and socioeconomic level. In short, the area of pupils’ writing interests is wide open for more research, but the problems involved are prodigious. Probably an entirely fresh approach is needed if progress is to be made in this area—an approach which goes deeper than the superficial concerns of the student.

One study which seems to probe into the psychological realm underlying or accompanying the act of composition is Van Bruggen’s investigation into the flow of words. In an experiment using a kymograph, Van Bruggen measured the “rate of flow” of words while junior high school pupils were writing compositions. He sought “to determine how it [rate of flow] is affected by various compositional, academic, personal, and environmental factors” and “to discover how the composing structure—that is, the number, length, and location of pauses between words—differs in compositions of superior and inferior quality and in compositions written with rapid and slow flow of words.” Although the investigator took careful precautions to see that use of the kymograph did not obtrude on the students’ writing, some of his procedures for analyzing the compositions seem very questionable, and some of his conclusions seem to leap beyond the reasonable distance for a step from data to generalization. But this study of the number and length of pauses between words may provide leads and procedures for valuable exploration of the psychological dimension of the composition act.

The Van Bruggen study, among others, suggests to the writers of this report that the psychological dimension of writing needs to be investigated by case study procedures. Individual differences may
"cancel out" in studies using the mean as the measure of a group. Case studies have done much to help remedial reading specialists understand and assist their "clients," and the similar complexities of writing suggest that much may be gained by developing case study procedures, against a background of experimental group research, to investigate the factors affecting the learning of composition and the procedures which will accelerate and maintain learning. But before composition teachers can conduct case study investigations, they must learn how to do so.

Another promising type of investigation is the longitudinal study—the type of study which follows the same individuals through a protracted period of time. The longitudinal study is especially appropriate for written composition, in which change usually seems to take place slowly. Note the effectiveness of the Harris study (summarized in Chapter IV) because it extended for two years instead of one. But the Harris study would not ordinarily be termed a longitudinal study. A clear example is the Loban project, which began in September, 1952, to study the development of language ability of 338 children in eleven kindergarten classes from socioeconomically diverse districts in Oakland, California. Working with a team of assistants and financed by a grant from the U. S. Office of Education, Loban has analyzed oral language samples obtained yearly from his subjects. He has also obtained written samples, beginning with the third grade (a written response to a colored picture), but, so far at least, has subjected the writing only to rough evaluation for the light it may throw on the development of oral language. Here is one finding from his analysis of oral language: "Those subjects who proved to have the greatest power over language . . . were the subjects who most frequently used language to express tentativeness. Supposition, hypothesis, and conditional statements occur much less frequently in the language of those subjects lacking skill in language." Numerous investigations using the above clauses of concession, though small, becomes more apparent in a somewhat consistent pattern as one analyzes the writing of older children. If one were to teach students, over a period of years, to express "tentativeness" (among other things) in their writing, would he help them develop their writing ability more than they otherwise would? Such questions

---

Instructions Factors Influencing Composition

As the discussion of the Loban study indicated, it is impossible to construct mutually exclusive categories of research studies. That study has dealt with environmental factors, but not to the exclusion of instructional factors. The next study considered here does, however, focus on instructional factors. But one should remember that teachers are part of the student’s environment, as is (some say) his own physical makeup—his visual and auditory acuity, his muscular coordination, etc. The student is inextricably intertwined with his environment. Regardless of one’s philosophical orientation on the nature of Self, it is convenient at times to categorize things, as is being done here, so long as one is not

---

"At a certain university there is a "room full of mud" that no one quite knows what to do with. A geology major years collecting tens hortage of the soil of the state. After accumulating this cornel, the professor who has run out of space in public print is a number of extra hortage, including two things a tower and the mud. Finally the professor who wants to use the mud is instructed to complete his work, but no one wishes to discard the data. At a time when space is needed badly on that campus, the roomful of mud sits undisturbed.

---


RESEARCH IN WRITTEN COMPOSITION

blinded to the reality that the categories are more a matter of convenience than of fact.

An instructional approach designed for retarded college freshmen but, with crowded classrooms, increasingly interesting to high school and college English teachers was examined experimentally by Maize.9 His experimental group consisted of 75 remedial college freshmen who wrote 26 themes some 250 words in length in addition to a 1,500-5,000 word autobiography, a 3,000-word research paper, and several business and social letters. These papers were read and corrected in class by the students with the guidance of the instructor, who used no time outside of class working with themes. A control group of 74 remedial freshmen wrote fourteen 250-word themes which were marked outside the classroom by the instructor. The control group was taught language usage according to traditional workbook-drill methods. The original study lists the matters covered in the control group. Unfortunately, the pretest and post-test theme examinations are vaguely described, though much attention is devoted in the study to the many objective tests used as well as two scales designed to measure attitudes toward the subject and the instructor. The statistically significant superiority of the experimental over the control group at the end of the study suggests that some combination of frequent writing and student correction is effective. The possibility also is strong, however, that superiority of the experimental group resulted primarily from the inferiority of the control group, which seems to have been taught a course which many people would now consider outmoded.

As it was one variable in the Maize study, so was frequency of writing a variable in the Buxton study (summarized in Chapter IV), the control group having received no regular instruction and practice in theme writing.10 Both of these studies and others indicate that college freshmen who receive a composition more complete and well written than those who do less writing.11 A problem which needs investigation, however, is just how much writing is the optimum amount. To carry the problem further than research seems to have done so far, one might well ask what kinds of writing following what kinds of instruction for what kinds of students. It does not seem reasonable that doubling the number of ains less writing assignments which are then marked in a perfunctory manner would necessarily stimulate students to improve their writing more than developing a somewhat less numerous but carefully planned sequence of assignments which are marked in an instructive manner related to the sequential pattern of the assignments. Another factor which may rival frequency of writing in importance is extensiveness of reading, as the recent Hays experimentation with eleventh graders indicated.12 But most such studies do not reveal the nature of the writing assignments as clearly as the Buxton study did, and it is thus difficult to compare and contrast one experiment with another. Although most of the research in this area supports the familiar contention that students learn to write by writing, with the help of a qualified teacher, it is not clear yet with what kinds of writing it is true, how much is the optimum amount, and with what kinds of students such generalizations are effective.

Another instructional factor influencing composition learning is the procedure of revision. Is it valuable for a student to revision his papers after his teacher has marked them? Again, the Buxton study affords a tentative "yes" for college freshmen revising their papers under supervision—tentative because revision was one of several variables in that study. It is not clear, however, which aspects of composition are learned better through revision, if one is concerned with that question. Fellows was, in an experiment he conducted with ninth graders.13 Although the experiment extended only 12 weeks and the gains were very small, the Fellows study does indicate that a combination of theme correcting by the teacher and revision by the pupil does have a more favorable effect on the grammar and punctuation (but not the spelling or capitalization) of above average ninth graders than does the practice of writing without correction and revision. In another experiment involving 1,059 Chicago area children in grades 6 through 9, Lyman demonstrated that even in a period as short as seven weeks pupils can reduce their grammatical and mechanical errors—including spelling and capitalization—more than by half by learning how to correct errors before submitting their papers.14 Remembering that it has yet to be shown what


10W. H. Buxton, "An Experiment to Test the Effects of Writing Frequency and Guided Practice upon "Outgoing" Skill in Written Expression" (Unpublished Ph.D. dissertation, Stanford University, 1948). (Note: for one, a study evaluated entirely by objective tests—Virgil L. Lokke and George S. Wysong, "Double Writing in Freshman Composition: Experiment," School and Society, LXXVIII (December, 1948), 417-430.


effect revision has on such elements as organization and supporting details, it does seem clear that what little research has been done in this high school pupils, and college freshmen.

It is somewhat surprising that little research has been done in that aspect of a composition teacher's work in which he spends so much of his time—the marking and grading of papers for instructional purposes. (The rating of compositions for purposes of research has been discussed as is available on the relative effectiveness of various grading procedures. One of Buxton's conclusions was that college freshmen whose writing is graded and thoroughly marked improve their writing more than those no grades or intensive marking. It is not clear, however, to what extent the intensive marking was responsible for that superiority and to what extent revision of the papers merits the credit. Certainly it has not been proved that intensive marking, with or without revision, is the best procedure to use with upper elementary grade or junior high school students.

One study which might have clarified the problem of grading if it had extended beyond its mere 11 weeks is Wormsbecker's thesis. Work-elementary schools, Wormsbecker investigated the effect on writing improvement of three different grading methods: (1) an overall impression, (2) a split grade, representing content and form as a teacher graded each set of themes for a different single consideration would be used. The fact that there were no significant differences among of the experimental period. A similar investigation was conducted by symbols and abbreviations to indicating merely the general location of errors by placing check marks in the margins of the themes. Halvorson's results also were inconclusive. In short, research has so far given no clear indication of the most efficacious way to mark papers except for the qualified endorsement by Buxton of thorough marking, a sentence

---

21Nelvin O. Halvorson, "Two Methods of Indicating Errors in Themes," College English, II (December, 1940), 272-278.
22Halvorson's results also were inconclusive. In short, research has so far given no clear indication of the most efficacious way to mark papers except for the qualified endorsement by Buxton of thorough marking, a sentence

---

THE STATE OF KNOWLEDGE ABOUT COMPOSITION

or two of constructive comment at the end, and a grade, and that endorsement does not necessarily hold good for younger people. (A recent review of the studies of various marking procedures may be found in Chapter I of an investigation by McMeekan.)

One of the most heavily investigated problems in the teaching of writing concerns the merits of formal grammar as an instructional aid. Study after study based on objective testing rather than actual writing confirms that instruction in formal grammar has little or no effect on the quality of student composition. Many of the summaries of research listed in Volume V provide title after title of studies confirming this generalization. For example, Butterfield contrasted the effectiveness of a grammatical approach and a "thought" approach in teaching 9th, 10th, seventh, and eighth grade pupils to punctuate. He concluded that, although the "students who were taught grammar per se learned significantly more grammar than students who were not taught grammar . . . . significantly superior results in punctuation are obtained by direct methods rather than by methods which are based upon a knowledge of so-called related grammatical elements."

Uncommon, however, is carefully conducted research which studies the effect of formal grammar on actual composition over an extended period of time. The Harris study (summarized in Chapter IV) compared the effect of instruction in formal grammar and functional grammar over a period of 2 years on the writing of 228 London pupils aged 12 to 14. Using a short-answer test on the terms and application of formal grammar as one measure and frequency counts of various types of sentence constructions and common grammatical errors in the students' compositions as a second measure, he concluded that, although the formal grammar classes made significantly higher gains on the grammar test, "the study of English grammatical terminology had a negligible or even a relatively harmful effect upon the correctness of children's writing in the early part of the five Secondary Schools."

In view of the widespread agreement of research studies based upon many types of students and teachers, the conclusion can be stated in strong and unqualified terms: the teaching of formal grammar has a negligible or, because it usually displaces some instruction and
RESEARCH IN WRITTEN COMPOSITION

practice in actual composition, even a harmful effect on the improvement of writing.

It may seem incomplete to conclude a discussion of instruction in composition without mentioning some of the more descriptive works on composition—such as the reports of McCall, Mears, Hargis, and Burrows. To mention several works of creative writing, or the recent analysis by Kitzhaber of college writing investigations, have a different approach to truth than the scientific studies of Kitzhaber. The book uses a kind of "informed intuition" with its "contrasts" rooted in the author's accumulated experience.

Rhetorical Considerations

If little has been proved about the instructional factors influencing composition, it is fair to say that almost nothing has been proved in a "rhetorical" sense about the rhetorical aspects of written composition. By "rhetorical" is meant here those aspects of writing which (to simplify somewhat) are larger than the unit of the sentence—in expository writing, for instance, the main idea and its analysis; the support of the topic; details; examples, statistics, and reasons; and the organization of the previous elements into an orderly and meaningful whole. It is a challenge to investigate these aspects of writing in a scientific way.

A rather sweeping attempt to identify some rhetorical considerations as well as other concerns in writing was made by Culpepper-Hagen. After she and two other raters identified the best 30 and the worst 30 writers in a group of 120 freshmen at the University of Denver (purposely excluding the mechanics of expression from their criteria of evaluation), and supplementing the evaluation of three papers from each student by a battery of objective tests and by the student's grade in the Basic Communication course, she administered to the 60 students a Writing

---

Methods Inventory. Included among her many conclusions were such statements as these about the distinctive tendencies of the best writers:

1. They follow a previously made plan or outline to guide them rather than think out and organize their topic as they write.
2. They have no unusual difficulty concentrating on the job of writing.
3. They feel no handicap because of lack of skill in punctuation, spelling, and grammar; therefore they do not worry about such details.
4. They leave their work largely as they write that, finding little use for elaborate revision and correction by themselves or others.

Although the Culpepper-Hagen investigation offers many leads for further investigation, it suffers, among other things, from the same weakness which plagues most questionnaire studies: one cannot be sure whether a response indicated the student's actual behavior or the behavior which he thought the investigator wanted.

A more focused and behaviorally oriented study on rhetorical considerations is one by Parker. Although not enough details are afforded in this report to enable the reader to evaluate it, the study appears to provide evidence that topic sentences and concluding summaries (in an expository pamphlet 2,000 words in length) increased comprehension for college freshmen. Cross headings and beginning summaries had little or no effect. Although the use of topic sentences increased comprehension when the subjects were tested immediately, no increase was noted when one week had elapsed between reading the pamphlet and taking a test on it.

Some people may say that Parker studied reading, not writing. What he did, however, is what some other investigators should do, too: study the effect of various rhetorical considerations on the reader. This may be a proof-of-the-pudding type of research which has deeper and more far-reaching possibilities than most investigations of writing, which, after all, tend strongly to be investigations of the effectiveness of writing for composition teachers. Doubtless many such studies undertaken by specialists in reading may have implications for effective writing, just as similar research by speech scholars (like Goyer) may suggest topics for research in the effectiveness of written composition.

---

"John P. Parker, ""Some Organizational Variables and Their Effect upon Comprehension," Journal of Communication, XII (March, 1932), 223.

"Robert S. Goyer, ""A Study of Individual Differences in Ability and Achievement of College Freshmen in the Organization of Ideas"" (Unpublished Ph.D. dissertation, The Ohio State University, 1935)."
Objective Tests versus Actual Writing as Measures of Writing

Sometimes treated with more passion than wisdom and often dealt with by people who have a psychological or financial stake in the outcome, the controversy over the relative merits of objective testing and actual writing as measures of writing will not be settled in the comments which follow. This statement may be useful, however, by touching upon some illustrative studies of the problem, bringing into focus some of the major factors involved, and offering some tentative suggestions to help those who have practical decisions to make before more conclusive research has been conducted.

"Objective tests" is usually taken to mean the familiar multiple-choice tests, seldom true-false or completion tests, in measuring the general merit of composition. A variation was the interlinear test developed some ten years ago by the Educational Testing Service. This test attempted to approximate the editing phase of composition by presenting the examinee with a passage which he was to edit in thirty minutes by crossing out incorrect or infelicitous expressions and inserting his attempts at correction and improvement between the lines of the passage. Trained readers, provided with lists of preevaluated possible responses for each of the problem situations consciously introduced into the passage, assigned each examinee's response a number of points, depending upon which of the listed possibilities the examinee's response most closely approximated. Although a test reliability of .51 and a reader reliability of .95 to .96 were reported for the interlinear test, this test has been criticized for its use of subjective raters. One can conjecture why; examinees' responses would not often enough match the lists of possible responses, and the readers, even though trained, would disagree on their evaluations of responses not anticipated in the tests. Fretting over the examinees' human departures from the anticipated responses, the examinees would find it difficult to resist using the anticipated responses as the possible answers in a multiple-choice form of the test. In addition, the interlinear test doubtless is characterized by the problems common to the more usual multiple-choice tests while much more expensive to score.

In general, multiple-choice tests of composition suffer from a number of shortcomings. The most common objection heard among composition teachers is that they do not require the examinee to select his own

---

words and to compose—to formulate and organize his own ideas into paragraphs and sentences; such tests are therefore, say these critics, inevitably not valid measuring instruments.9 The defenders of multiple-choice tests are quick to respond; perhaps even acknowledging what may seem to be "self-evident invalidity," they point nevertheless to the helplessness of these objective measures as predictors of success in future composition classes10 and to the notorious unreliability of the grading of actual compositions.

Certainly it is true that disagreement among theme graders is common. Perhaps the clearest testimony of this unreliability was developed in a study by Diederich and two colleagues.12 They analyzed the way ten English teachers rated 300 two-hour compositions by college freshmen in comparison to 45 other raters: social scientists, natural scientists, writers, and editors, lawyers, and business executives. The raters were given no standards or criteria for judging the papers, merely asked to sort the themes into nine piles in order of general merit, with not less than 4 percent of the papers in any pile. It was "disturbing to find that 94 percent of the papers received either seven, eight, or nine of the nine possible grades, that no paper received less than five different grades, and that the median correlation between readers was .51. Readers in each field, however, agreed slightly better with the English teachers than with one another." Not enough better was the reliability reported by Noyes for the one-hour English Composition test included among the Achievement Tests used by the College Entrance Examination Board during 1943 and 1944: "A lot of people worked very hard at setting and scoring the themes, in six successive tests during 1943 and 1944. The highest reliability of reading we ever attained was about .57..." Many similar instances of unreliable grading have been reported by other investigators.

But the defenders of objective tests of writing ability seem to overlook or regard as suspicious the high reliabilities obtained by some investigators. Noyes himself refers to reliability indices of .57, .88, and .87 attained with the one-hour theme which was part of the CEEB...
Comprehensive Examination in English before World War II. In studies summarized at length in Chapter IV, Buxton reported reader reliability coefficients of .91 and .88, and Kincaid obtained reliabilities ranging from .77 to .91. In one of the most thorough explorations of this problem, Finlayson found reliabilities ranging from .786 to .961 in the rating of one-hour papers written by 850 sixth graders in 21 Edinburgh primary schools. It seems clear, then, that high reader reliabilities are possible. Suggestions were made in Chapter II for obtaining such reliable grading.

The most serious charge against multiple-choice tests of writing is their lack of validity. Not only do they not require the examinee to perform the actual behavior being measured—she does no actual writing; but these tests also make little or no attempt to measure the "larger elements" of composition, even indirectly. In her study of a multiple-choice test and the rating of actual writing, Huddleston found that the objective test contained 27 punctuation items, 33 problems in idiom, 47 in grammar, and 70 in sentence structure—evidently no items on the clarity, focus, and analysis of the main idea; the logic and clarity of the organization; or the soundness and specificity of the supporting content. It is little wonder, then, that the written part of her test consisted of three 20-minute essays; she did not have to be concerned with affording the examinees time to analyze and formulate their ideas. Although his analysis, too, was made of a test which (though widely used at the time) has been largely supplanted now, Vordenberg found invalidity even within the "smaller" elements of composition, which the objective test attempted to cover. An analysis of the errors in freshmen themes in comparison to the types of errors included in the test led him to conclude that capitalization errors were emphasized more on the test than their occurrence in writing warranted, that the seriousness of sentence fragments in writing was not reflected in the weight accorded them in the test, and that other types of errors were insufficiently represented, if at all, in the test: wordiness and repetition, omissions, superfluous commas, italics, various hyphen problems, certain types of spelling errors, etc. And, again, the examination did not test for such matters as unity,

...in the case of marked differences between the essay and objective ratings of a student in English composition the objective rating is probably the more valid estimate. Students who receive high essay scores and low objective scores tend to be an inferior group as judged on the basis of their intelligence-test scores or on the basis of their class performance in English composition. Certainly one has reservations about this study. The number of cases used for this conclusion is small, and the shortness of the writing samples probably precludes opportunity for the freshmen to demonstrate well their ability to handle the larger elements of composition. Furthermore, one may ask questions about the reliability of the essay grading and the emphasis of the composition course. In how many schools has objective testing been a good predictor of success in composition classes precisely because those classes have emphasized grammatical and mechanical matters with little or no emphasis on central idea, analysis, organization, and content?

Doubtless one reason that the College Entrance Examination Board has given up grading the composition it includes in its examination is that colleges differ in their emphases in composition instruction. How can the CEEB rate papers in a manner valid for all colleges if some colleges emphasize narration, others exposition, still others argument or literary criticism? And within those emphases, one college may
stress content and organization, another style, still another "correctness." The CEEB has found a partial solution by requiring a paper on an assigned topic and forwarding the paper to the candidate's preferred institution for evaluation. But even that is not completely satisfactory, for doubtless a number of colleges would prefer different topics, eliciting the kind of writing they attempt to teach. Until that dull, gray utopian day when all freshman composition courses have the same emphasis, the CEEB composition examination will have varying validities for the institutions receiving its papers. Meanwhile, the CEEB continues to include an essay examination in its battery of tests, partly to counteract the critics who fear a "backwash effect" on the secondary schools if no actual writing is demanded. The CEEB probably discounts this argument, however, since French \[...\]

In view of all these conflicting beliefs, what is the English teacher, department head, or chairman of freshman composition to do about objective tests and actual writing as measures of writing ability? The answer must depend in part upon what he wishes to accomplish. If a school department head or college chairman of freshman composition wishes to section large numbers of students according to their predicted success in composition, he probably will find a relatively sound objective test the most efficient instrument. Although some missectioning doubtless will occur, usually the individual students can be reclassified during the first week or two of the year on the recommendation of their teachers. (This would work, of course, with the 11-point examinations used in England. There a consideration of actual writing seems in order.) The expense and effort involved in reliable rating of actual compositions are not often justified where only a rough sorting of large numbers is needed and where occasional errors in classification can be corrected with little pain. Meanwhile, those who construct such objective tests should continually strive to improve and update them. Especially they should bend their best efforts to include items on the larger elements of composition.

---

**The State of Knowledge about Composition**

If a teacher is interested in diagnosing the strengths and weaknesses in various aspects of the composition of individual students, he will do best to base his diagnosis on an examination of their actual writing, realizing that he must gear the assignment to the matters he wishes to diagnose. The objective testmakers seldom claim that their tests are useful for this purpose, and those who do have yet to substantiate their claims. When a beginning teacher does not feel confident of his ability to diagnose his students' writing, he would be wise to exchange a set of themes for a set from a more experienced colleague and then study the colleague's comments.

If a teacher or chairman wishes to evaluate the writing performance of students after instruction, he should plan carefully to use several papers written by the students. Objective tests will not accomplish this purpose; if the teacher has been emphasizing rhetorical matters as well as grammar and mechanics, objective tests simply are not valid—they make little or no attempt to measure the larger elements of composition. But, as Kincaid demonstrated (see Chapter IV), one composition is not enough, especially for more capable students. If a large number of students is to be evaluated on a pass or fail basis, those students who do not pass a first composition may be given a second opportunity; there is little point in examining a second time the student who passes the first. But if students are to be evaluated for the purpose of assigning grades, probably they should be given as many as three compositions, with only the best two of the three grades being averaged for the final grade. The assignments and the basis of evaluation should be similar to or the same as those used in instruction, and the papers should be graded by at least two raters using a mutually understood and accepted rating system with which they have previously practiced. Other considerations were discussed in Chapter II. Additional suggestions may be found in articles by Dieterich,41 Gerber,42 and Wiseman.43

**Other Considerations**

**Size of English Classes**

Although there has been much research on size of class in general, little has appeared on the size of English classes and less on its influence

---

on written composition. The Smith study, published in 1961 and summarized in Chapter IV, investigated the effect of class size on various aspects of ninth grade English. It seems possible that investigation that teaching composition to superior ninth graders may be generally as effective in large classes as in small classes if the teacher has the requisite ability and is given adequate time to prepare and adequate assistance in handling the routine aspects of class management. Now that the overhead projector has been tried out with reported success in a number of large lecture situations, it seems time for a new experiment in size of class, using lay readers and the overhead projector as aids to composition instruction.

**Lay Readers**

One way in which more than routine assistance can be given the English teacher is to provide the services of a "lay reader," usually a person not certified to teach in the public schools but qualified by other experience to assist in the marking and grading of compositions. No convincing study of the effectiveness of a lay reader program was discovered by those preparing this report, and that is not surprising. It is a natural thing—as in the Sauer study—to have a lay reader experiment in which the experimental and control classes are taught by the same teacher, automatically alleviating the total instructional load of the teacher, who may tend, then, to devote more time to his usual control class because he is being assisted by the lay reader in his experimental class. It seems clear that a carefully designed experiment with lay readers necessitates, in part, a time study of the teachers involved. Another unique necessity, if the findings are to have meaning to school systems outside the experiment, is to include some objective descriptions of the qualifications of the lay readers and teachers involved; otherwise, a conclusion valid for a high school in, for example, a university community where there are many competent faculty wives who would like to be lay readers has little meaning for a less fortunate community where there is no abundant supply, if any, of people competent to mark compositions. If one has read the Dusel report and is cognizant of the enrollment pressures on public schools and the shortage of first-rate

---

**THE STATE OF KNOWLEDGE ABOUT COMPOSITION**

English teachers, he knows that the use of lay readers deserves careful investigation.64

**Teaching by Television**

Despite the fact that much money has been granted by foundations for experiments in the use of television as an instructional aid, little of the research, as published, seems convincing, at least where instruction in written composition is concerned. Often the nature of the course of instruction is not made clear, or the evaluation of the experiment is based solely on objective testing, or the attitudes of the students have not been carefully examined. In one experiment, the cost of the equipment and studio technicians seemed to exceed by far the salaries which were saved by reducing the number of instructors. Sometimes glowing published reports seem to overlook rumors of inattentive students and disgruntled teachers whose normal classroom procedures have been interrupted rather than supplemented by the television set in the front of the darkened room. Until such time as more conclusive research has been published, one is tempted to reserve the television teaching of composition for correspondence students who cannot be present on a college campus.

The above criticisms do not seem very applicable to the Becker study (summarized in Chapter IV), in which kinescopes (films made with television cameras) were used in the teaching of college composition.65 Kinescopes, like teaching films, can be used with more flexibility of timing than television, even though they were not so used in the Becker experiment. Although the kinescopes were not always technically smooth and the results of the experiment would have been more convincing if a trial period had permitted the instructors to adapt their teaching better to the medium being used, the study does indicate that kinescopes may be a useful instructional aid, especially where instructors are teaching composition for the first time. The study makes much the same claim for releasing freshmen from one class meeting each week (of four usual meetings) to supplement their classwork by library study guided by a bibliography of readings on, among other things, rhetoric and written composition. After reading the Becker study, one wonders if a series of professionally prepared and mature teaching films on composi—

64Dore V. Smith, *Class Size in High School English: Methods and Results* (Minneapolis: University of Minnesota Press, 1931).

65William J. Dusel, "Determining an Efficient Teaching Load in English," *Illinois English Bulletin*, vol. XLI, no. 9 (October, 1958), 1-19. (This issue of the Illinois bulletin is distributed by the National Council of Teachers of English.)
tion would not prove even more effective with freshmen than kinescopes or television. The problem is to find teaching films which are well done and which cover the particular matters emphasized in the individual school or college.

Writing Vocabulary

A fruitful product of research on writing vocabulary has been the word count—a count of the frequency with which words appear in various kinds of writing. Most apparently pertinent to this pamphlet are the counts based on the writing of students and adult nonprofessional writers. Not designed for spelling lists, they are guides to persons wondering which words may be used commonly enough at a given grade level to warrant inclusion in a spelling list for that grade. Word lists based on children’s writing certainly are not designed to be used for reading vocabulary study; reading vocabulary begins to exceed writing vocabulary soon after a child starts to read and quickly becomes far more extensive than his writing vocabulary. Graded writing vocabulary word lists can also be useful guides in judging one aspect of the maturity of a pupil’s writing. Among the more carefully constructed of these word lists are the following, all of which antedate any impact of television on vocabulary:


The sources of this list are (1) lists of words which children were asked to construct in grades 2 through 5 in Illinois, Manhattan, and Brooklyn and in grade 2 in New England, and (2) ten other word lists of various kinds, including those based on children’s writing, speech, and literature. The sources and recommended grade levels are indicated after 7,000 words of the 19,000.


This list of 2,650 words was selected from a much larger list compiled through study of seven other lists drawn from children’s writing, both in and out of school, and from adult writing.

Horn, Ernest. A Basic Writing Vocabulary: 10,000 Words Most Commonly Used in Writing. University of Iowa Monographs in Education. Iowa City: College of Education, University of Iowa, 1926. 225p.

THE STATE OF KNOWLEDGE ABOUT COMPOSITION 49

The sources of this list are business correspondence, personal letters, and a variety of other letters and types of writing, written by college students and adults. Although this list was published in 1928, many people still consult it because of the diversity and level of its sources.


This list is constructed from the writing of children in grades 1 through 8, submitted by the officials of 708 schools throughout the United States. The graded list includes 14,571 words.


Based on the Thorndike general count of 1931, the Lorge magazine count, the Thorndike count of 120 juvenile books, and the Thorndike-Lorge semantic count, this list gives the frequency of occurrence of each word according to each of the four sources and a frequency which represents a summary from all four counts. Although based on professional writing, this list is still useful in developing criteria to judge the maturity and style of student writing or to guide the selection of words by the professional writer who does not wish his vocabulary to be too difficult for his prospective readers.

Other types of research in writing vocabulary would be valuable, too—studies in the range, freshness, and precision of diction at various levels of maturity and intelligence, and studies of the variation in the connotations which words have for readers of different types. A promising procedure for getting at the range of a student’s writing vocabulary was used in a study by Newman.4

Spelling

Research in spelling is barely touched upon in this report for the simple reason that, however important accurate spelling may be in the

clarity and social acceptability of composition, many of the factors of good spelling do not seem to be closely involved with the factors of good composition. A number of investigators have shown, for example, that it is more effective to base the selection of words for spelling instruction on carefully prepared spelling lists than to base selection on spelling words chosen from the compositions of the pupils. A recent confirming study of this type was conducted by Deacon, working with 18 teachers and 414 second and third graders. 49

One study of spelling which should be on every composition teacher’s reference shelf, although it is not a study of spelling in student writing, is the Emery report. 50 Emery compared the variant spellings of the same words in five standard college-level dictionaries and found many discrepancies, suggesting that teachers should exercise caution (by consulting the 28-page “Reference List of Variant Spellings” in the Emery pamphlet) before marking some words as misspelled.

One final word regarding spelling research is in order. Although it evidently makes sense to teach children in grades 1 through 5 to spell a basic core of words from high-frequency lists rather than from words they misspell in their own compositions, that is no reason why studies of misspelling should not be based on actual spellings in student compositions rather than on multiple-choice tests or dictated lists. When one reads published reports on misspelling, the printing of the report obscure a great truth—that many words in student writing have been formed so carelessly that one cannot be sure: whether or not they are misspelled, or, if so, how. More use of photographs of handwritten words in published research would help to convey some sense of the problem the spelling investigator faces in illegibility.

Handwriting

Like spelling, handwriting appears to be a field of research which more often than not has less to do with composition than one might suppose. As with spelling, the writers of this report made no attempt to review the thousands of published reports on the subject, although they did read those few studies recommended to them and listed in the bibliography in Chapter V. Two observations do seem pertinent. An alarming proportion of the small sample of studies read seems to be based on the handwriting used when pupils copy some sentences from the

---


THE STATE OF KNOWLEDGE ABOUT COMPOSITION

blackboard or, more commonly, write from memory some familiar passage like the opening lines of the Gettysburg Address. Surely such artificial circumstances cannot yield a valid sample of pupils’ typical handwriting, done during the heat of composition. The other observation is more in the nature of a request—that the titles of reports on handwriting use the word “handwriting,” not the ambiguous “writing.”

Typewriting

The effect of typewriting on progress in written composition evidently has been the subject of very little investigation, most of that with elementary school children. Moreover, because most, if not all, of the studies have been subsidized by grants from typewriter manufacturers, one cannot help but feel some reservations about the findings. One such investigation, conducted in the late twenties by Wood and Freeman, and utilizing 6,125 pupils in the typewriter group, 8,824 in the control group, concluded that typewriting stimulated more and longer original compositions, both in typewriting and in handwriting, especially at the first and second grade levels. 51 The most favorable influence of typewriting seemed to be on spelling, as measured by that portion of the Stanford Achievement Test. Generally speaking, Wood and Freeman found the typewriter equally helpful to bright, normal, and dull pupils. A more recent study by Erickson and Clow, based on 24 matched pairs of fifth graders, showed mixed results. 52 Certainly the effect of the typewriter on composition deserves careful investigation, financed by sources other than manufacturers. The ease with which kindergarten and primary school children seem to be able to operate portable electric typewriters argues especially for a careful investigation in that area, perhaps using Van Bruggen’s “rate of flow” index, described early in this chapter, to determine whether or not the typewriter can increase the fluency of writing in general.

Relationships of Oral and Written Composition

With the realization revivified by structural linguists that written language evolves from oral language, the need is more apparent than ever for investigations into the rhetorical and syntactic relationships between the two types of language. The Harrell study did investigate

---

some features of the relationships in the writing and speaking of children in grades 4, 6, 8, and 10.\textsuperscript{63} His language samples seem artificial, however, inasmuch as the compositions were merely narrative reproductions of a story and the oral narratives were delivered into a visible microphone when the speaker was alone with the investigator. On the other hand, Harrell's idea of using ten-minute films to trigger oral and written composition appears to be a good means for achieving standard stimuli. If the content of films—especially silent films, to forestall copying—can be tied to pupils' desires to write, perhaps more original, spontaneous language samples will result. The relationships between oral and written composition are still wide open for research, as is the important matter of stimulating a desire to write well.

Unexplored Territory

Some questions which seem fundamental in the teaching and learning of written composition apparently have gone almost untouched by careful research. This chapter concludes with a list of questions, not considered previously in this chapter, which indicate areas in which future investigators may wish to direct their efforts:

1. What kinds of situations and assignments at various levels of schooling stimulate a desire to write well?
2. What do different kinds of students prefer to write about when relieved of the expectations and requirements of teachers and others?
3. What are the sources of fear and resentment of writing?
4. How do the kinds of writing which adults compose vary with their occupations and other factors?
5. What is the effect on writing of having the student compose his paper for different kinds of readers?
6. At which levels of maturation does it seem appropriate to introduce the various modes of discourse—narration, poetry, drama, exposition, argument, and criticism?
7. What is the relative effectiveness of writing shorter and longer papers at various levels of maturity and proficiency?
8. At which levels of maturation does it seem appropriate to introduce the various rhetorical elements of writing?

9. What are the effects of various kinds and amounts of reading on the quality and kinds of writing a person does?
10. What are the direct and indirect effects of particular sensory experiences and guided observation upon writing?
11. At what stages of maturity do students spontaneously seek specific help in improving particular aspects of writing, such as specificity of details, transitions, parallel structure, and metaphor?
12. At which levels of maturation can particular aspects of writing most efficiently be learned?
13. Does the oral reading of rough drafts help the elementary school child strengthen "sentence sense"? How does it?
14. What techniques of composition most effectively help build self-discipline and pride in clarity, originality, and good form?
15. What procedures of teaching and learning composition are most effective for pupils of low socioeconomic patterns?
16. What procedures of teaching and learning composition are most effective for pupils learning to write English as a second language?
17. Can study of the newer types of linguistics help writers?
18. Can formal study of rhetorical theory or of logic help writers?
19. How is writing affected by extensive study and imitation or parody of models?
20. What forms of discourse have the greatest effect on other types of writing? For example, does writing poetry help a writer of reports?
21. What is involved in the act of writing?
22. How does a person go about starting a paper? What questions must he answer for himself?
23. How does a writer generate sentences?
24. Of what does skill in writing really consist?

IV.
SUMMARIES OF SELECTED RESEARCH

Basis for Selecting These Studies

One strong impression which should emanate from Chapters II and III is that there are many variables to control in research with written composition. There are so many that it is an unusual study which does not leave several important variables uncontrolled or undescribed. Consequently, the writers of this report were tempted more than once not to select any “most soundly based studies” to summarize at length, as the Executive Committee of the National Council had requested. (See Chapter I.) Recognizing, however, that “most soundly based” does not mean “perfect in all respects” but “most soundly based of all those studies available,” the writers have selected five distinctly superior investigations to summarize here in Chapter IV. Because three of these studies have never been published and the other two are out of print, the summaries are detailed. The fullness of the summaries serves to inform the reader who wishes to determine for himself whether or not to be convinced by the conclusions of the studies, the investigator who wishes to replicate or extend the experiments described, and the student of research who wishes to discover how some superior investigations have been designed, executed, and reported.

Although the methods of research suggested in Chapter II are, in effect, criteria which can be used in selecting “most soundly based studies,” the details of those suggested methods (or criteria) tended to become apparent to the writers of this report as they screened the studies recommended to them. From the beginning of this screening, however, certain broad criteria were employed. Generally, the selected studies, the writers agreed, should do these things:

1. Base the investigation at least in part on the direct observation of actual writing, not entirely or mostly on such indirect measures as objective tests or questionnaires.

2. Study the writing either of a generous number of students (never actually specified, “generous” seemed to mean at least...
seventy or eighty) or of as few as twenty students who were very carefully selected or very carefully matched with another twenty.

3. Describe the procedures in a controlled experiment or the features of writing in a textual analysis in enough detail that it is clear what was being studied.

4. Use procedures of statistical analysis which, though not necessarily complicated, were appropriate and consistent and did not obscure the raw data being analyzed.

5. Maintain as objective an investigation as possible by controlling or reporting the salient variables; that is, by keeping the investigator as "removed" from the study as possible, by preserving the anonymity of the students when evaluating or analyzing their writing, by describing the abilities of the pupils used, etc.

As the writers of this report screened the studies to a more and more limited number, they naturally became increasingly detailed in their expectations, drawing more heavily from the considerations covered in Chapter II. They sometimes had to weigh an outstanding merit against a striking flaw or to consider the desirability of representing a particular type of study in this report. Moreover, they anticipated some criticism because one of the five studies had been conducted by a member of the ad hoc committee (the Smith study) and another had been accomplished at the institution with which the three writers of this report are associated (the Becker study). But the writers finally concluded, with the support of the ad hoc committee, that those two investigations should be included because of their intrinsic merits and interest, despite the charges of favoritism which might be made.

**Explanation of Statistical Terms**

For the benefit of the reader who is not familiar with statistical terminology, the writers include here explanations of the terms used in a number of the reports.

**Level of confidence**

Because there are always a number of uncontrolled factors operating in an experiment, the effects that are observed cannot all be attributed to the particular treatment under consideration. There are available a number of statistical procedures to determine the probability that an observed effect is due to these "chance" factors rather than to the experi-

**SUMMARIES OF SELECTED RESEARCH**

mental variable. This probability is referred to as the level of confidence or significance; for example, if an effect is significant at the .01 level of significance, it means that there is only one chance in one hundred that differences as large as those observed could have occurred by chance.

**Variance**

The spread or scatter among the scores in a distribution is frequently a very useful bit of information to have. One such measure of scatter is the variance, the square root of which is referred to as the standard deviation. In addition to their descriptive usefulness as measures of variability, the variance and standard deviation are used in a number of the statistical tests that are employed to determine significance of an effect.

**Analysis of variance**

This is a statistical procedure frequently used to test the significance of an effect, particularly when there are more than two experimental groups. Basically it involves comparing different estimates of the variance of a distribution. These comparisons yield an F-value which when used with appropriate tables indicates the level of the significance of an effect.

**Analysis of covariance**

It is frequently desirable to match subjects in an experiment, but this matching becomes very difficult when dealing with certain types of subjects, e.g., intact classrooms. Analysis of covariance is a statistical procedure which adjusts the scores obtained in such a way as to match the groups statistically when actual matching is not possible.

**Chi-square**

Chi-square ($\chi^2$) is used when there is a need to determine whether or not a distribution of scores is like some hypothetical distribution the experimenter has set up. A value based on the discrepancies between the actual and hypothetical scores is referred to a table, and significance of the discrepancies can then be determined.

**Critical ratio**

Most tests of significance are based on a ratio of some sort. The ratio varies with the size of the differences between or among the means. The critical ratio is that ratio which is of such a magnitude as to be considered statistically significant.
F-ratio

The F-ratio is a statistic that is used under certain conditions to test whether or not the results of an experiment may be considered to have arisen by chance.

t-test

The t-test, like the F-ratio, is a statistic used to determine the significance of an effect.

At the beginning of each summary, a bibliographic reference to the study is given. When the University Microfilms number is included, the dissertation cannot be borrowed through interlibrary loan but must be purchased, in microfilm, directly from University Microfilms, 313 North First Street, Ann Arbor, Michigan.

The Buxton Study


Purpose

The purposes of this experiment, conducted at the University of Alberta during the academic year 1956-1957, were to determine (1) whether or not “regular practice in writing over a period of seven months (a University of Alberta ‘year’) [would] result in a significant improvement in writing skill” and (2) which of the following methods was superior in improving writing skill: (a) a “freedom from restraint,” or “Writing,” method in which no interlinear or marginal marks were employed on the student papers, no grades were noted on the papers, the only remark was a paragraph of generous comment at the end of the paper, and the students were not advised to correct or revise their papers, and (b) a “revision, writing, and revision,” or “Revision,” method, in which the papers were thoroughly marked, graded, commented upon (adversely, when necessary) in a paragraph at the end of the paper, and discussed by the raters and revised by the students during 35 to 50 minutes of class time when each paper was returned. Thus the experiment at-
Table 1: Courses and Instructors in the Junior Elementary Program at the University of Alberta

<table>
<thead>
<tr>
<th>Course Number</th>
<th>Name</th>
<th>Class Hours</th>
<th>Control Group</th>
<th>Writing Group</th>
<th>Revision Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ed.101</td>
<td>Physical Education</td>
<td>48</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Ed.105</td>
<td>Health</td>
<td>48</td>
<td>12</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Ed.106</td>
<td>Social Studies</td>
<td>72</td>
<td></td>
<td></td>
<td>Social Studies Staff</td>
</tr>
<tr>
<td>Ed.121</td>
<td>Primary Work</td>
<td>72</td>
<td>14</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Ed.127</td>
<td>Mathematics</td>
<td>30</td>
<td>10</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Ed.129</td>
<td>Science</td>
<td>48</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Ed.130</td>
<td>English</td>
<td>72</td>
<td>20</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Ed.132</td>
<td>Art</td>
<td>48</td>
<td>21</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>Ed.133</td>
<td>Music</td>
<td>48</td>
<td>21</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>Ed.161</td>
<td>Administration</td>
<td>36</td>
<td>19</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Ed.176</td>
<td>Psychology</td>
<td>72</td>
<td>11</td>
<td>15</td>
<td>9</td>
</tr>
</tbody>
</table>

Notes on Table:
1. Numbers in columns under Group C, Group W, and Group R refer to instructors; thus, Affleck is number 1, Buxton number 2, Caldwell number 3, and Worth number 23.
2. Education 106 (social studies) was taught by instructors 4, 16, and 23. Each instructor taught one unit to all classes; therefore, each class had the same allotment of time with each instructor.
3. The division of classes among instructors in Education 121, Education 127, and Education 161 was not so desirable as it might have been for this investigation. However, since the assignments of these classes largely involved collecting materials for primary work, surveying primary texts, working with numbers and number materials, observing classes, and planning timetables, it is felt that any differences in students' experience in essay writing were negligible.

The treatment of assignments and papers in the Writing group was characterized by "freedom from restraint." If a student did not wish to write on the topic assigned, he was encouraged to choose any topic he desired, even an anecdote, imaginative experience, or personal happening. No intrainstructor or marginal marks or comments were written on their papers, but a three- or four-sentence comment was written at the end, giving a general impression of the essay by praising it as much as possible but also by suggesting one or two ways in which the next paper could be made clearer or more interesting. No grades of any kind were placed on the Writing group essays; they were returned to the students without comment, and no suggestion was made that the papers be revised or corrected.

The assignments and papers in the Revision group were treated with considerably more direction. Although the students could develop the assigned topic "in their own way," they were expected to write on the same topic and to include some critical thinking, a central idea, and material that was organized and developed. They were encouraged to organize preliminary ideas into an outline before beginning the theme itself, to choose their words and illustrations carefully, and to develop good unity within paragraphs and transitions between them. They were also warned against using unqualified and fallacious statements. These qualities—as well as errors in spelling, punctuation, and sentence structure—were marked on the papers, and a few sentences of general evaluation were written at the end, including mention of commendable qualities as well as suggestions for improvement. Each paper was given two grades, one for content and organization, another for general correctness and accuracy.

The papers of the students in the Revision group were returned at the beginning of a class period. The general strengths and weaknesses of the essays were pointed out at that time, and excerpts exemplifying certain good features were read to the class to elicit comments on how the effectiveness was achieved. Then the students were required to correct the errors indicated on their papers while the reader went from student to student, giving assistance where it was needed. No more than one 30-minute class period was devoted to these procedures, the average time having been 30 minutes for each assignment—a total of some nine hours during the year.

The investigator reported that all of the essays of the Writing and Revision groups "were read by a group of qualified readers appointed by the investigator and paid on an hourly basis out of funds allotted to this investigation by the Alberta Advisory Committee on Educational Research." More specifically, the readers were graduate students, each specializing in the field of the instructor with whom he was working. The papers in art were thus read by a graduate student in art, those in music by a graduate student in music. The investigator noted that "most" of the readers "had had teaching experience," but he did not indicate how much. Presumably, few of them had been English teachers. The readers also conducted the follow-up work (described in the previous paragraph) undertaken when papers were returned to the Revision students.
But the regular instructors assigned the topics, discussed the assignment with the students, and collected the papers when they were due. The instructors also returned the papers to the Writing students.

In order to maintain consistency from one class to another, a general meeting of all the instructors was held in September, before the program was begun, to discuss the plan of the experiment. The investigator reported that the instructors "were interested in the investigation and eager to cooperate" when it was assured that they would not have the paper grading added to their usual duties. Sheets outlining the procedures for each group were distributed to all of the instructors and readers. (Copies are included in the final report as Appendices A, B, and C.) Furthermore, the instructors frequently conferred with the investigator about the topics they planned to assign, representative examples of which follow:

"What Music Means to Me"
"My Observation of Science Teaching in the Elementary School"
"My Visit to the Alberta Legislature" (and my conclusions)
"My Opinion of the Film 'Near Home'"
"Handwriting Can Be Improved"

After considering whether or not the theme grades should count as part of their regular subject grades, the investigator visited each of the six classes before the first assignment was given, "described the experiment, stated the responsibility of each group, and asked for their cooperation." He advised the students in the Writing and Revision groups that the practice would probably improve their writing, but he explained that essay grades would not affect their grades in the subject areas. "This approach seemed to prove quite satisfactory," the investigator stated, "because the students throughout the year made every effort to complete all assignments on time in spite of a rather heavy subject-matter program.

To determine whether or not the three groups were equivalent for the purposes of the experiment, all of the students were given three tests prior to the beginning of classes in September, 1956. (1) the Cooperative English Test A, Mechanics of Expression, Form 2, (2) the Cooperative English Test Bb, Effectiveness of Expression, Form Z, and (3) an original essay written during a 50-minute period. Parallel forms of these three examinations were used again in March, 1957, to measure the performance of the students after instruction. In March, Form X of each of the Cooperative tests was used. These multiple-choice tests were administered on each occasion by the University of Alberta Student Advisory Services, and the answer sheets were scored by IBM machine. The scaled scores were used in all computations.

The instructions for the essay pretest and post-test were given on mimeographed sheets (included in the final report as Appendices E and F). Because of the widely varying backgrounds of the students, a broad topic was assigned and suggestions were made on the instruction sheet of different ways in which the topic could be narrowed. The topic of the first of these 50-minute themes was "High Schools," and the topic for the post-test was "My Opinion," the latter permitting practically any kind of stand on any subject. On the instruction sheet for the post-test, the students were advised "Since your name will be replaced by a number before your essay is read, all opinions expressed are entirely anonymous so far as your readers are concerned."

Reliability in rating the essays was sought through a number of precautions. The anonymity of the student writer was preserved by having a colleague mix the 257 examination booklets so thoroughly that "both class and alphabetical arrangement were eliminated." Each paper was then numbered (from 1-257) on the cover and the first inside page, and the cover (with the student's name and number) was removed.

The essay tests were graded by two raters. One was the investigator; the other was the Chairman of the Grade Twelve Essay Marking Committee of the Alberta Department of Education. Both raters had had more than twelve years' experience in teaching English in high school. A score sheet was prepared to guide the essay test raters. The sheet provided for rating each theme according to general areas of consideration, under each of which its constituent factors were listed and the maximum number of points to be awarded each factor was given. A feature which may have contributed strongly to the validity of the experiment is that the score sheet went through several preliminary stages as the raters practiced with it on trial themes. (The fact that the raters helped develop the score sheet as they practiced with it suggests that they not only understood it but believed in it and followed its instructions.) The final score sheet (reproduced as Figure 1) contained fifteen categories, each of which could receive a maximum number of points.

In addition, each paper received a "basic mark" of 75 points, and deductions amounting to as much altogether as 100 points could be made for spelling, punctuation, usage, grammar, sentence, and form. The maximum total theoretically possible was 300 points.
Practice rating was provided to the two raters. Themes were obtained for practice by having fifty freshmen not in the experiment write on the same topic and under the same conditions as the students in the experiment. The two raters marked and discussed each of these papers. They then began rating the experimental papers. (The report does not make clear whether or not the raters were aware which papers were pretests and which post-tests. Since the papers were numbered from 1 to 257, we may infer that the pretests and post-tests were rated separately and that the raters knew which was which, though obviously they did not know which papers were written by the Control group, which by the Writing, which by the Revision. Consequently, the results of the rating may be considered useful for comparing the relative composition quality of the three groups but not for measuring progress from pretest to post-test.)

Each rater read each paper at least three times:

During the first reading, the errors in spelling, usage, punctuation, and sentence structure were noted, and an attempt was made to evaluate the content, introduction, and conclusion. During the second reading, examples of effective diction, concreteness, figures of speech, and critical thinking were noted on the back of the scoring sheet and marks given for these. During the third reading, an attempt was made to evaluate the logical sequence of paragraphs, transition between paragraphs, unity, and the general coherence of the whole. Frequently a fourth reading was necessary to assess the accuracy of scores given.

The two raters independently marked and graded each paper, leaving to "a committee" the computation and recording of final scores. The raters jointly reviewed every twenty-fifth paper, as they went along, to insure continuous consistency but not to change the grade. Interpretations which had to be made of the score sheet as the rating progressed were recorded (on pages 39-41). A check on accuracy, each rater marking mechanical and grammatical errors in the margins, but "the number of errors was not considered in the awarding of marks for positive qualities." In order to adjust for the number of opportunities for error in longer and shorter themes, the number of errors was divided by the number of words written (and then multiplied by 1,000) to find an "error rate." The final score for each rater was computed by adding the 75-point "basic mark" and the points awarded for material, organization, sentences, and diction; then, from that possible maximum of 300 points, was subtracted the "error rate." The final scores of the two raters were averaged and then divided by 3 to yield a scaled score—the number of points based on 100, which facilitated comparisons of the essay and objective
test scores and made other statistical computations more manageable. The raw scores, subsequent adjustments, and resulting scaled scores for each of the 257 papers are given in Appendix J, though the scores for two post-test themes were estimated by the missing plots technique.

Rater reliability coefficients estimated by the Pearson product-moment formula were .01 for the pretest themes and .88 for the post-test themes, indicating a high degree of consistency between the two raters.

Results

The mean scores for each group on each of the three tests are given in Table 2. (The rest of this discussion will concern only the essay ex-

| Table 2 |
| Mean Scores of Pretests and Post-tests on All Tests for All Three Groups |
| Control Group (N=86) | Writing Group (N=86) | Revision Group (N=85) |
| Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| Mechanics of Expression | 53.29 | 56.99 | 53.78 | 59.88 | 52.89 | 58.18 |
| Effectiveness of Expression | 57.05 | 59.05 | 57.16 | 59.09 | 43.35 | 57.28 |
| Essay | 42.65 | 49.28 | 42.10 | 48.97 | 40.67 | 51.79 |

amination, not the two objective tests, other than to note here that the differences among the three groups on the two objective pretests were found not to be significant at the .05 level of confidence.

Before analyses of variance and covariance could be conducted with the essay test scores, it was necessary to determine the degree of homogeneity of the variances among the three groups—Control, Writing, and Revision. Then the significance of the differences between the post-test means was calculated by an analysis of covariance to adjust the post-test means in terms of differences in the pretest means. Significance was determined by applying the t-test to the differences between the adjusted post-test means.

When the investigator computed F-ratios as tests for the homogeneity of the within-groups variances on the essay pretest, the investigator found one ratio significant at the .05 level—the ratio of the Control group to the Revision group. Consequently, he applied Bartlett’s test of the significance of the differences among the within-groups variances. The value of the chi-square resulting from the calculation was 2.818, when it would have had to be 5.991 or greater to be significant for two degrees of freedom. This indicated that an analysis of variance and covariance was justified.

Analysis of variance on the essay pretest yielded an F-ratio of .8688, where 3.035 is significant at the .05 level of confidence with 2/253 degrees of freedom (and 4.695 is significant at the .01 level), indicating that there was no significant difference between any two groups at the beginning of the experiment. Analysis of covariance of the essay post-test scores, adjusted for differences in the pretest scores, yielded an F-ratio of 4.97 for the among means mean squares divided by the within-groups mean squares where 4.959 is significant at the .01 level of confidence with 2/251 degrees of freedom. Inasmuch as the three adjusted post-test means were thus seen to differ significantly, the investigator applied the t-test to the differences between pairs of adjusted post-test means. The difference between the means of the Writing and Control groups was significant at the .01 level, and the difference between the means of the Revision and Writing groups was significant at the .05 level of confidence. In other words, the Revision group improved very significantly over the Control group and significantly over the Writing group, though the Writing group did not improve significantly over the Control group.

The investigator explored further to determine whether or not the differences in achievement on the post-tests were in fact the result of the different experimental methods of instruction. He obtained the mean scores achieved by all the students on their final examinations in all their subjects:

| Control Group | 63.3 |
| Writing Group | 64.4 |
| Revision Group | 63.2 |

After determining by Garrett’s method that the within-groups variances were homogeneous, the investigator computed the significance of the differences between means. Where t = 1.99 at the .05 level of confidence for 25 degrees of freedom, the critical ratio between the Writing and Control groups was 0.18, between the Revision and Control groups was 0.13, and between the Revision and Writing groups was 0.18. In other words, it is not apparent that instructional factors other than those in the experiment affected the results.
The investigator conducted one last analysis—an analysis of the gain in mean raw score from pretest to post-test essay made by each group on each of the fifteen categories on the score sheet. The gain was computed for each student in each of the fifteen categories. The mean gains and standard deviations are given in Table 3.

### Table 3
**Mean and Standard Deviation of Gain-Scores by Groups and Categories**

<table>
<thead>
<tr>
<th>Category</th>
<th>Control Group</th>
<th>Writing Group</th>
<th>Revision Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean S.D.</td>
<td>Mean S.D.</td>
<td>Mean S.D.</td>
</tr>
<tr>
<td>1. Significance</td>
<td>5.07 8.22</td>
<td>3.14 8.13</td>
<td>6.68 8.70</td>
</tr>
<tr>
<td>2. Critical Thinking</td>
<td>1.53 6.64</td>
<td>1.58 6.09</td>
<td>1.04 8.64</td>
</tr>
<tr>
<td>3. Originality</td>
<td>1.18 6.66</td>
<td>1.23 6.57</td>
<td>2.24 8.54</td>
</tr>
<tr>
<td>4. Title</td>
<td>1.60 4.20</td>
<td>2.80 4.02</td>
<td>5.20 4.30</td>
</tr>
<tr>
<td>5. Introduction</td>
<td>.46 3.92</td>
<td>.61 3.60</td>
<td>2.81 4.16</td>
</tr>
<tr>
<td>6. Logical Sequence</td>
<td>1.20 5.80</td>
<td>1.76 3.26</td>
<td>2.26 4.54</td>
</tr>
<tr>
<td>7. Unity</td>
<td>1.12 3.52</td>
<td>1.24 2.96</td>
<td>1.06 3.76</td>
</tr>
<tr>
<td>8. Transition</td>
<td>1.80 6.22</td>
<td>1.76 6.60</td>
<td>1.66 6.54</td>
</tr>
<tr>
<td>9. Coherence</td>
<td>.99 2.58</td>
<td>.46 2.88</td>
<td>1.26 3.24</td>
</tr>
<tr>
<td>10. Conclusion</td>
<td>1.08 5.12</td>
<td>1.94 4.92</td>
<td>2.12 5.76</td>
</tr>
<tr>
<td>11. Variety in Sentence Structure</td>
<td>1.97 7.11</td>
<td>2.99 7.08</td>
<td>4.28 6.30</td>
</tr>
<tr>
<td>12. General Fluency</td>
<td>.64 2.58</td>
<td>1.29 2.98</td>
<td>1.46 2.26</td>
</tr>
<tr>
<td>13. Diction</td>
<td>2.03 10.68</td>
<td>4.34 10.35</td>
<td>5.93 10.56</td>
</tr>
<tr>
<td>14. Figures of Speech</td>
<td>4.16 9.15</td>
<td>5.93 8.31</td>
<td>8.46 10.86</td>
</tr>
</tbody>
</table>

From the figures in Table 3, the investigator estimated the significance of the difference between each mean difference and the smallest mean difference (.48), using the usual critical ratio method. Using the 5 percent level of confidence, the investigator found those critical ratios significant which are shown in Table 4.

TheRevision group gained significantly over the Writing group in three categories and over the Control group in six of the fifteen categories. (As has been shown on page 65, because the raters evidently knew which papers were pretests and which were post-tests but not which were written by students of the three respective groups, these "gains" should be viewed as evidence of the relative effectiveness of the three methods but probably not as clear evidence of absolute gain in a particular category.)

### Table 4
**Significant Critical Ratios for Groups and Categories**

<table>
<thead>
<tr>
<th>Category</th>
<th>Control</th>
<th>Writing</th>
<th>Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Significance</td>
<td>4.95</td>
<td>2.88</td>
<td>6.39**</td>
</tr>
<tr>
<td>2. Critical Thinking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Originality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Title</td>
<td>2.07</td>
<td>4.41</td>
<td>8.78***</td>
</tr>
<tr>
<td>5. Introduction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Logical Sequence of Paragraphs</td>
<td></td>
<td>4.35***</td>
<td></td>
</tr>
<tr>
<td>7. Unity</td>
<td>3.27</td>
<td></td>
<td>5.12</td>
</tr>
<tr>
<td>8. Transition</td>
<td></td>
<td></td>
<td>2.39</td>
</tr>
<tr>
<td>9. Coherence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Conclusion</td>
<td></td>
<td></td>
<td>2.43</td>
</tr>
<tr>
<td>11. Variety</td>
<td></td>
<td>3.09</td>
<td>5.09*</td>
</tr>
<tr>
<td>12. Fluency</td>
<td></td>
<td></td>
<td>2.56*</td>
</tr>
<tr>
<td>13. Diction</td>
<td></td>
<td>3.34</td>
<td>4.67*</td>
</tr>
<tr>
<td>14. Figures of Speech</td>
<td>3.59</td>
<td>5.88</td>
<td>6.56*</td>
</tr>
<tr>
<td>15. Error Rate</td>
<td>3.09</td>
<td>5.27</td>
<td>5.76</td>
</tr>
</tbody>
</table>

Total: 4 5 11

*Gain of Revision over Control significant
**Gain of Revision over Writing significant
***Gain of Revision over both Control and Writing significant

Among the conclusions which may be clearly drawn from this experiment are these:

1. Raters who use a scoring system which they understand and endorse, who have practiced rating papers by this system, and who make periodic checks on their rating may achieve a high degree of consistency in rating essays, or themes.
2. Since the students in the three groups did not differ significantly in their gain on the two objective post-tests but did differ significantly in their gain on the post-test essay examination, it may be seen that essay examinations can be used to measure changes which may not be measured by objective tests.
3. College freshmen whose writing is graded and thoroughly marked and criticized and who revise their papers in the light of
these matters can improve their writing more than college freshmen whose writing receives a few general suggestions but no grades or intensive marking and who do not revise their papers. (It is not clear, however, what the relative influence is of each of these three factors.) Further experimentation would be necessary, of course, before one could feel confident in making such a generalization about, say, junior high school pupils.

The Harris Study


Problem

This study investigated the relative teaching usefulness of what might be loosely referred to in the United States as "formal grammar" and a "direct method" of instruction. Although previous investigators have focused on this problem, most of them have either restricted their data to objective test results or have conducted their studies over too short a period of time. Harris' research was based on frequency counts of actual writing done before and after a two-year period of instruction. The reader's attention is also directed to the manner in which Harris tried out his procedures in a three-month pilot experiment and validated his criteria of measurement before he undertook his regular experiment.

The investigation compared two forms (or classes) of pupils in each of five London schools; "... one form had as nearly as possible the same general English work as the other, but with the one major difference that while one form studied and applied the terms of formal grammar in its composition work, the other used none of these terms and devoted the time saved to direct practice in writing." Although the investigator refers to the two groups of pupils as the "Grammar" and the "Non-Grammar" groups, actually grammar and composition were taught in each group. To prevent misunderstanding by American readers, the terms "Formal Grammar" and "Direct Method" will be used here to distinguish between the two groups. The essential differences were, as the investigator explained on pages 131-139, that the Formal Grammar group followed a logically organized program of traditional grammar

SUMMARIES OF SELECTED RESEARCH

instruction "through the parts of speech, with stress on the function of words" and employed the traditional grammatical terminology in classroom teaching and in correcting compositions, and the Direct Method group evidently used no textbook or grammatical terminology but considered the elements of "sentence building and structure" which came to the teachers' attention as they read the children's writing, treating common errors in the classroom and in compositions "by means of example and imitation, instead of by the abstraction and generalization of the approach through formal grammar—which did not itself, of course, exclude the use of examples." The time saved by not teaching formal grammar in the Direct Method group was devoted to additional composition work.

To insure that the distinction between the two groups, and hence the problem of the experiment, is clear, an extended quotation from the final report is included here:

Such mistakes as "Jim and me was going into the cave"... would be tackled in the grammar forms by direction to agreement of subject and verb and to the proper use of the nominative forms of pronouns when the pronouns are used as the subject. First, of course, subject, verb, agreement, pronoun, nominative or subject case, were terms that had to have been absorbed—at least, this was the ideal supposition. Probably the example which followed ("Jim and I were going...") actually drove home the point, and one hoped that the rule would spread the influence of the example to other parallels. In the non-grammar forms, there was time for more examples. The sentence would be looked at in this way:

"Jim and me was going into the cave."

Teacher: "Would you say 'We was going into the cave'?"

General dissent—one or two brave individuals aver that they would or might, and why not sit? But the vast majority of the class bring home to them the weight of convention. They admit to being wrong, which was more than was asked of them.

Teacher: "What would you say then?"

Class: "We were going into the cave."

Teacher: "How many is 'we'?"

Class: "Depends upon how many there are. More than one, anyway."

Teacher: "Well, 'Jim and me' means more than one. So they must be followed by 'were.' Let's try some more examples... . . . And the class, and the teacher, provide twenty or thirty examples in five minutes orally. With a group that needed this sort of practice, it was wise to try to correct the pronoun error in the same lesson. One has to decide on priorities.

*The textbook used in four of the schools was Humphreys and Roberts' Active English Course, Books 1 and 3 (London: University of London Press, 1939); and the text used in the remaining school was Allen and Mason's An English Grammar of Function, Book 1 (London: Edward Arnold and Company, 1939).
Procedures

To refine his procedures and measuring instruments, the investigator conducted a three-months' pilot experiment with a pair of classes, first-year forms at a Technical School. The classes contained 22 and 24 boys. As in the main experiment, the plan for each class was the same except that the one lesson each week devoted to formal grammar in the one class was spent on writing and functional grammar in the other. The work done by both groups included "practice in summarising, in comprehension, in reading and composing verse, in composition, in silent reading, reading aloud and discussion." The Formal Grammar group studied sentence-types, subjects and predicate, the parts of speech, and the idea of function, using Active English Course, Book 1, the same text used by four of the five schools in the main experiment. Instead of studying those matters, the Direct Method group wrote a story about the exploits of a boy separated from his parents during the war. Evidently the class worked out a common plan for the story but each pupil wrote his own version, doing a chapter each week. The teacher afforded instruction in planning the story, achieving inherent plausibility (even if not factual accuracy), maintaining suspense, and using details to suggest character and establish atmosphere. No class or group instruction was given the Direct Method group in the imitation of sentence structures; but individuals were aided in rephrasing their sentences, and the results were "often brought to the attention of the class." When the pupils finally completed their stories, they exchanged them and evaluated them in committees of three, for suspense, characterization, and vividness of scenes. The investigator noted that the procedures used with the Direct Method group took as much but no more time than those used with the Formal Grammar group and that they elicited "not only enthusiasm but also a certain self-criticism and purposive modification of habits of writing."

Two measures were administered to each group before and after the three-months' instruction. One was a short-answer formal grammar test, also used in the main experiment and included in the report as Appendix 3, which required the identifying in sentences and naming from recall of the parts of speech, subject and direct object, phrase and clause, tense and mood, and the like. The examinees were also required to explain "in grammatical terms" what was wrong with a number of sentences. The Formal Grammar group showed much better progress than the

SUMMARIES OF SELECTED RESEARCH

Direct Method group on this measure: the difference between means was 7.345, the standard error was 1.349, and t was 5.45.

The other measure was a composition. Both classes wrote a descriptive paper on "A Day in the Country" at the beginning and on "A Day at the Seaside" at the end of the three-month period of instruction, being afforded two 40-minute class periods for the writing each time. Instead of evaluating the papers by solely subjective judgment, the investigator conducted a number of frequency counts, using from each paper the first 500 words to the nearest end of a sentence. Two types of counts were made: (1) the number of words per error, as listed in Table 1, and (2) the number of complex or noncomplex sentences and the number of correct or incorrect sentences. Illustrations were not given of which types of usages and sentences were considered correct and which incorrect.

Table 1

(Appendix 1 in the Harris Report)
A List of Common Errors in Children's Writing

1. Certain errors in punctuation, notably—
   a. omission of question mark;
   b. omission of full stop;
   c. omission of commas between items in a list, words in apposition, main and non-defining clauses.
   d. omission of apostrophe in 's for possession.

2. Failure to use capital letters—
   a. at beginning of sentence.
   b. in proper nouns and adjectives.

3. Minus of various parts of speech—
   a. adjective or preposition as adverb.
   b. wrong comparatives and superlatives.
   c. faulty positioning of adverbs.
   d. failure to give pronouns a clear antecedent.
   e. wrong or ambiguous use of prepositions.
   f. mismanagement of conjunctions.
   g. minus of object forms of pronouns.
   h. failure to use relative pronoun or adverb.

4. Failure to give a finite verb to each clause, error in or omission of any important verbal word.

5. Lack of agreement between verb and subject.

6. Faulty sequence of tenses.

7. Unrelated or false participle, e.g., "laying" for "lying."

NOTE: In the error count referred to in the pilot experiment, and in the final reckoning of the third measuring instrument, errors 1a, 1d, 2a, and 2b were omitted from the calculations as being too remote from grammar as such and too simply mere matters of convention which training in formal grammar should not be expected to help in eradicating.
RESEARCH IN WRITTEN COMPOSITION

but a familiarity with examples elsewhere in the Harris report (especially in Tests A and B in Appendix 3) suggests that a rather strict sense of formality was observed. (For instance, a distinction in meaning is drawn on page 154 between "He only found three ducats" and "He found only three ducats").

The differences between the means of the Formal Grammar and Direct Method groups on the frequency count measures were all small, insignificant, and favoring the Direct Method group. Encouraged by those results, considering the small number of children examined and the shortness of the pilot period of instruction, the investigator decided to use the frequency count approach in the main experiment but to refine the procedures involved. But he did not reach that decision until he had tried to gauge the "differences in correctness" between boys of 12 and 14 years of age by using two tests similar to those common in the United States, tests in which the examiner chooses which of two alternate forms is the better. Finding no regularity of increase in scores with age when he administered the tests to boys aged 11, 12, 13, and 14 at one school, the investigator abandoned that type of test. The formal grammar test (Test C), on the other hand, not only achieved a coefficient of reliability of .923 ± .014 by the split-half method, but it yielded regular increases in scores with age, as Table 2 shows.

Table 2: Validation Data for Test C

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Average Score</th>
<th>Range of Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 years</td>
<td>8.1</td>
<td>0-16</td>
</tr>
<tr>
<td>12</td>
<td>13.0</td>
<td>4-23</td>
</tr>
<tr>
<td>13</td>
<td>22.8</td>
<td>12-34</td>
</tr>
<tr>
<td>14</td>
<td>33.1</td>
<td>19-48</td>
</tr>
</tbody>
</table>

In order to find some frequency count measures which would reflect "the rate of growth of a child's maturing style" as well as Test C reflected maturity in knowledge of formal grammar, the investigator made a comparative analysis of some compositions written by 63 ten-year-old children and by 25 fifteen-year-olds. After a number of initial assessments which did not prove to differentiate satisfactorily between the two groups, the investigator developed a set of 11 criteria "which occurred sufficiently often to give a clear measurement, and which did not over-lap too much." The 11 criteria are given in Table 3. For each criterion, the investigator had computed the difference between the mean scores

at age 10 and age 15, the standard error, and the critical ratio (t). The three statistics are given in parentheses at the end of each item in Table 3, demonstrating the apparent significance of each of the items finally selected.

To discover how reliable were the differences established by the criteria, the investigator applied them to two descriptive compositions written, a week apart, by 27 thirteen-year-old boys who were allowed forty minutes to write on each of these topics: "A House on Fire" and "A Foggy Night." Table 4 shows the product moment correlation of their scores (r), the standard error, and the reliability quotient (t), with the level of significance. It is evident that all criteria except a and g reached a good or high level of reliability.

Other investigators interested in conducting frequency counts using Harris' 11 criteria may also wish to use the simplified process he developed for tabulating his counts and scores, a sample of which is given in Figure 1.

SUMMARIES OF SELECTED RESEARCH

Table 3: Criteria of Maturing Style

| a. Average length of the correct simple sentence; i.e., the simple sentence containing none of the errors listed in Table 1. (D 4.86, SE .832, t 5.84) |
| b. Instances of the omission of the full stop at the end of a sentence and of inadequate separation between clauses, scored by subtracting the number of omissions from the number of sentences in which the stop was correctly inserted. (D 11.44, SE 1.870, t 6.79) |
| c. Number of words per common error. (See Table 1, including the note at the foot.) (D 141.0, SE 22.90, t 6.16) |
| d. Number of different sentence patterns used by each writer, counting only correct sentences but accounting, for example, different types of adversarial clauses or inversion of the order of similar clauses as sufficient to provide a new pattern. (D 6.31, SE 1.13, t 5.00) |
| e. Number of non-simple sentences, whether correct or incorrect, from which the number of simple sentences was subtracted. (D 7.00, SE 1.21, t 5.78) |
| f. Number of subordinate clauses used—the actual number (not in proportion to the number of main clauses). (D 11.44, SE 1.71, t 6.07) |
| g. Total number of words written in a limited time. (D 235.4, SE 23.00, t 10.24) |
| h. Number of correct simple sentences, minus those containing an error (even though minus scores were frequent at 10 and 11 years of age). (D 9.513, SE 0.916, t 10.38) |
| i. Number of correct simple sentences containing two or more qualifying phrases, each phrase consisting of three or more words. (D 1.063, SE 0.85, t 1.24) |
| j. Total number of correct sentences minus the number of incorrect. (D 13.66, SE 1.573, t 9.96) |
| k. Number of adjectival clauses and phrases, each phrase consisting of three or more words. (D 12.97, SE 1.341, t 9.67) |
Table 4

<table>
<thead>
<tr>
<th>Reliability of the Criteria</th>
<th>t</th>
<th>SE</th>
<th>level of signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. simple sentence length</td>
<td>-0.039</td>
<td>0.181</td>
<td>worse than 0.1</td>
</tr>
<tr>
<td>b. deletion of full stop</td>
<td>0.574</td>
<td>0.129</td>
<td>better than 0.01</td>
</tr>
<tr>
<td>c. words per common error</td>
<td>0.844</td>
<td>0.055</td>
<td>better than 0.001</td>
</tr>
<tr>
<td>d. number of sentence patterns</td>
<td>0.701</td>
<td>0.073</td>
<td>better than 0.001</td>
</tr>
<tr>
<td>e. non-simple sentences</td>
<td>0.362</td>
<td>0.163</td>
<td>better than 0.05</td>
</tr>
<tr>
<td>f. number of subordinate clauses</td>
<td>0.705</td>
<td>0.097</td>
<td>better than 0.001</td>
</tr>
<tr>
<td>g. total words</td>
<td>0.235</td>
<td>0.178</td>
<td>worse than 0.1</td>
</tr>
<tr>
<td>h. complex sentences</td>
<td>0.685</td>
<td>0.102</td>
<td>better than 0.001</td>
</tr>
<tr>
<td>i. simple sentences with two or more phrases</td>
<td>0.447</td>
<td>0.154</td>
<td>better than 0.02</td>
</tr>
<tr>
<td>j. total correct sentences</td>
<td>0.839</td>
<td>0.057</td>
<td>better than 0.001</td>
</tr>
<tr>
<td>k. adjectival clauses and phrases</td>
<td>0.446</td>
<td>0.154</td>
<td>better than 0.02</td>
</tr>
</tbody>
</table>

With his measuring instruments prepared, the investigator completed the design of his experiment. After some difficulty from other schools because of the length of the experiment, the investigator secured the cooperation of five London schools representing a range of socio-economic conditions: two from West London (Isleworth and Dormers Wells), one from nearer the center (Christopher Wren), and two from "rather East" (Shoreditch and Owens). See Table 5.

Because of administrative necessities at the five schools, the investigator was not able to select the pupils for each group and hence equate the groups for intelligence, attainment, or success in English. Fortunately, however, the pupils had not been grouped according to ability, and the tables of classes in each school were roughly equivalent (See Table 34, page 278, in the original study), with the Formal Grammar groups having in three of the five schools a slight superiority in IQ. Although some pupils were lost from the experiment during the second year, 119 were able to spend both years in the Formal Grammar classes, 109 in the Direct Method.

<table>
<thead>
<tr>
<th>Paper No.</th>
<th>Sentence No.</th>
<th>Error No.</th>
<th>CLASSES USED</th>
<th>No. of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1 M</td>
<td>Vi M</td>
<td>VT M</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>Vi M</td>
<td>M M</td>
<td>VT M</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>Vi M</td>
<td>M J</td>
<td>T M</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>M Noj M</td>
<td>M</td>
<td>VT M</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>M Noj M</td>
<td>M M M M</td>
<td>M</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>VT M</td>
<td>M M M</td>
<td>T M</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>1 b</td>
<td>M</td>
<td>VT M</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>1 b</td>
<td>M</td>
<td>M M</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>M M M</td>
<td>M</td>
<td>M M</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>1 b</td>
<td>M M M</td>
<td>M M M Noj M</td>
<td>40</td>
</tr>
<tr>
<td>11</td>
<td>6,3 d</td>
<td>M</td>
<td>M VT M Noj M</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>M</td>
<td>M M M M</td>
<td>24</td>
</tr>
</tbody>
</table>

Scores:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>--</td>
<td>+1</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>+1</td>
<td>38.5</td>
<td>12</td>
<td>12</td>
<td>20</td>
<td>308</td>
<td>+1</td>
<td>--</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Key:
- S = simple sentence without any qualifying phrase.
- S + 2 = simple sentence but with two or more qualifying phrases.
- M = main clause.
- N = noun clause.
- Noj = noun clause as object.
- Nappos = noun clause in apposition.
- V = adverbial clause.
- Vreass = adverbial clause of reason.
- Vconc = adverbial clause of concession.
- Vi = adverbial clause of time.
- J = adjectival clause.

Table 5

Participants in the Experiment

<table>
<thead>
<tr>
<th>Grammar Schools</th>
<th>Formal Grammar Classes</th>
<th>Direct Method Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isleworth</td>
<td>29 boys</td>
<td>26 boys</td>
</tr>
<tr>
<td>Dormers Wells</td>
<td>29 girls</td>
<td>21 girls</td>
</tr>
<tr>
<td>Comprehensive/Technical</td>
<td>21 boys</td>
<td>23 boys</td>
</tr>
<tr>
<td>Christopher Wren</td>
<td>19 boys</td>
<td>14 boys</td>
</tr>
<tr>
<td>Shoreditch</td>
<td>119</td>
<td>109</td>
</tr>
</tbody>
</table>
RESEARCH IN WRITTEN COMPOSITION

In four of the schools, one teacher taught both classes, but in the Secondary Modern school the two classes were taught by different individuals. Though the investigator acknowledged that "no certain existence that each teacher was absolutely impartial in his attitude to the teaching or abandonment of formal grammar, or indeed equally adept at the one approach as at the other," he judged that "all the teachers were willing to be convinced by evidence, desirous of finding out the truth about the matter in hand, competent and practised in the present grammar of the language." (No further attempt seems to have been made to estimate the probable effects of the teacher attitudes and skills on the outcome of the experiment.) The investigator also reported "The general method of grammar teaching among the five teachers was surprisingly uniform," grammatical form and terminology being "closely tied to function and meaning," but he did not describe any procedures used to observe or control the teaching in either group, none, that is, other than the teacher direction for the Formal Grammar groups implied by the textbooks, the formal grammar test, the List of Common Errors, and "the relevant sections of the Course of Study." The teacher instructions for the Direct Method groups "excluded any reference to formal grammar and specified attention to practice in writing and imitating conventional forms and structures." The investigator also stated that the teachers were careful not to use the terminology of formal grammar when correcting the homework of the Direct Method pupils.

Each class met for five 40-minute periods per week and divided its work during four of the periods among the usual areas of reading, drama, poetry, and composition. The difference came in the fifth period, when one group emphasized formal grammar and the other focused on direct methods of instruction. (Details are given early in this summary, under "Problem.") The investigator described the difference further: "The work attempted in the [Direct Method] lesson was basically an extension of the usual composition practice [in Formal Grammar classes], but most teachers gave cohesion and interest to the work by engaging in a variety of longer projects such as the compilation of a diary, a form newspaper, an adventure story similar in general outline to the one discussed in the pilot experiment, or a book of hobbies." Thus some of the time released by omitting study of the terminology of formal grammar could be devoted to writing activity projects and to drawing illustrative sentences, points of usage, and paragraphs from the stories to teach the improvement of writing. A piece of continuous writing was also attempted in the Formal Grammar classes, but not much time was available for the project.

SUMMARIES OF SELECTED RESEARCH

Inasmuch as these classes were carefully following the integrated grammar and composition lessons in their textbooks.

Results

The investigator reported the results of the pretest and post-test in formal grammar (Test C) as shown in Table 6. He wrote, "It is evident from these scores that good progress was made by the grammar forms [Formal Grammar classes] in learning not only the terms of formal

| Table 6 |
| Scores on Test C, the Formal Grammar Test |

<table>
<thead>
<tr>
<th>Group</th>
<th>First Average</th>
<th>Second Average</th>
<th>Gain</th>
<th>Difference</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>School A</td>
<td>Formal</td>
<td>33.03</td>
<td>49.71</td>
<td>16.68</td>
<td>16.11</td>
<td>1.955</td>
</tr>
<tr>
<td>(Grammar)</td>
<td>Direct</td>
<td>34.12</td>
<td>54.69</td>
<td>0.97</td>
<td>0.97</td>
<td>2.393</td>
</tr>
<tr>
<td>School B</td>
<td>Formal</td>
<td>36.53</td>
<td>66.62</td>
<td>30.99</td>
<td>26.97</td>
<td>3.32</td>
</tr>
<tr>
<td>(Grammar)</td>
<td>Direct</td>
<td>34.88</td>
<td>38.50</td>
<td>3.62</td>
<td>3.62</td>
<td>2.393</td>
</tr>
<tr>
<td>School C</td>
<td>Formal</td>
<td>20.10</td>
<td>34.34</td>
<td>14.24</td>
<td>12.86</td>
<td>1.825</td>
</tr>
<tr>
<td>(Sec. Mod.)</td>
<td>Direct</td>
<td>22.69</td>
<td>33.47</td>
<td>1.58</td>
<td>1.58</td>
<td>2.393</td>
</tr>
<tr>
<td>(Comp. Tech.)</td>
<td>Direct</td>
<td>21.09</td>
<td>21.83</td>
<td>0.74</td>
<td>0.74</td>
<td>2.393</td>
</tr>
<tr>
<td>School E</td>
<td>Formal</td>
<td>15.84</td>
<td>32.54</td>
<td>17.00</td>
<td>15.50</td>
<td>1.669</td>
</tr>
<tr>
<td>(Comp. Tech.)</td>
<td>Direct</td>
<td>22.71</td>
<td>24.21</td>
<td>1.50</td>
<td>1.50</td>
<td>2.393</td>
</tr>
</tbody>
</table>

grammar, but how to apply these, for Test C contained questions requiring both recognition and use of the common grammatical terminology at the appropriate level." The gains of the Formal Grammar group were, of course, significantly higher on this test than the gains of the Direct Method group.

Frequency counts for the criteria listed in Table 3 were made from essays written at the beginning and end of the two-year experiment. A similar analysis was conducted of essays written at three of the schools after nine months of instruction, "to check that things were proceeding according to plan." Although each class used the same topic for the final essay that it used for the first, other topics (but ones calling for the same mode of discourse) were employed for the intermediate composition. (The investigator did not report the topics used, but he indicated that the mode of discourse was narrative or descriptive in each case.)

It seems highly important, in the light of the results of the final composition (reported later), to note that the results of the intermediate
composition analysis were very inconclusive. The differences between means of the Formal Grammar and Direct Method groups at each school were computed for each of the eleven criteria, a total of 33 differences. Of these 33, only four differences were significant, where t = 3 or over. Three of those favored the Direct Method group, one the Formal Grammar group. Another four differences, reaching a fair level of significance (t = 2.70 or over), all favored the Direct Method group. Of the total of 33 differences, 16 favored the Formal Grammar group and 17 the Direct Method group, leading the investigator to conclude "The period of nine months or one academic year is too short to expect any stable changes to have been made, and to allow for variation of pace within a course."

Significant differences appeared much more clearly in the comparative analysis of the compositions written at the end of the two-year experiment, as is shown in Table 7. Of the 55 differences computed, 11 were significant, and all 11 favored the Direct Method group. Of the 25 differences represented by the five very reliable measures (c,d,f,h, and j), only six favored the Formal Grammar group, and none of those differences was significant.

Table 7
Final Scores on Frequency Count Criteria

<table>
<thead>
<tr>
<th>Rel. Measure and School</th>
<th>Diff. bet. Means</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Words per common error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>14.26</td>
<td>13.63</td>
<td>1.05 g</td>
</tr>
<tr>
<td>B</td>
<td>25.41</td>
<td>26.68</td>
<td>3.19 *</td>
</tr>
<tr>
<td>C</td>
<td>37.41</td>
<td>33.32</td>
<td>3.33 *</td>
</tr>
<tr>
<td>D</td>
<td>38.41</td>
<td>11.92</td>
<td>3.23 *</td>
</tr>
<tr>
<td>E</td>
<td>41.46</td>
<td>21.84</td>
<td>1.59</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Different sentence patterns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.322</td>
<td>0.685</td>
<td>0.47 g</td>
</tr>
<tr>
<td>B</td>
<td>3.331</td>
<td>1.042</td>
<td>3.19 *</td>
</tr>
<tr>
<td>C</td>
<td>2.779</td>
<td>0.767</td>
<td>3.62 *</td>
</tr>
<tr>
<td>D</td>
<td>2.349</td>
<td>0.900</td>
<td>2.59</td>
</tr>
<tr>
<td>E</td>
<td>2.824</td>
<td>1.149</td>
<td>2.46</td>
</tr>
<tr>
<td>F-U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Non-simple sentences minus simple sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2.120</td>
<td>1.140</td>
<td>1.86</td>
</tr>
<tr>
<td>B</td>
<td>3.649</td>
<td>2.056</td>
<td>1.77</td>
</tr>
<tr>
<td>C</td>
<td>0.710</td>
<td>1.688</td>
<td>0.42</td>
</tr>
<tr>
<td>D</td>
<td>1.374</td>
<td>1.848</td>
<td>0.74 g</td>
</tr>
<tr>
<td>E</td>
<td>1.305</td>
<td>1.468</td>
<td>0.89 g</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Subordinate clauses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.174</td>
<td>1.637</td>
<td>0.72</td>
</tr>
<tr>
<td>B</td>
<td>1.116</td>
<td>2.333</td>
<td>0.49</td>
</tr>
<tr>
<td>C</td>
<td>1.951</td>
<td>2.874</td>
<td>0.86</td>
</tr>
<tr>
<td>D</td>
<td>0.036</td>
<td>1.919</td>
<td>0.02</td>
</tr>
<tr>
<td>E</td>
<td>0.255</td>
<td>2.376</td>
<td>0.11 g</td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g. Total words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>70.43</td>
<td>21.54</td>
<td>3.27 *</td>
</tr>
<tr>
<td>B</td>
<td>57.99</td>
<td>36.63</td>
<td>1.31</td>
</tr>
<tr>
<td>C</td>
<td>52.16</td>
<td>31.39</td>
<td>1.12</td>
</tr>
<tr>
<td>D</td>
<td>37.02</td>
<td>19.18</td>
<td>1.93 g</td>
</tr>
<tr>
<td>E</td>
<td>30.16</td>
<td>37.33</td>
<td>0.81 g</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h. Complex sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.048</td>
<td>1.107</td>
<td>0.59 g</td>
</tr>
<tr>
<td>B</td>
<td>5.042</td>
<td>1.468</td>
<td>3.44 *</td>
</tr>
<tr>
<td>D</td>
<td>4.495</td>
<td>1.185</td>
<td>3.08 *</td>
</tr>
<tr>
<td>C</td>
<td>4.953</td>
<td>1.935</td>
<td>3.86 *</td>
</tr>
<tr>
<td>E</td>
<td>5.257</td>
<td>1.779</td>
<td>3.56 *</td>
</tr>
<tr>
<td>F-U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Simple sentences with two or more phrases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.076</td>
<td>1.899</td>
<td>0.04 g</td>
</tr>
<tr>
<td>B</td>
<td>0.436</td>
<td>2.483</td>
<td>0.17</td>
</tr>
<tr>
<td>C</td>
<td>0.899</td>
<td>0.383</td>
<td>2.26 g</td>
</tr>
<tr>
<td>D</td>
<td>0.417</td>
<td>0.473</td>
<td>0.89</td>
</tr>
<tr>
<td>E</td>
<td>0.918</td>
<td>0.793</td>
<td>1.16</td>
</tr>
</tbody>
</table>
RESEARCH IN WRITTEN COMPOSITION

(Table 7 cont'd)

<table>
<thead>
<tr>
<th>Rel.</th>
<th>Measure and School</th>
<th>Diff. bet. Means</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j.</td>
<td>Number of correct sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2.007</td>
<td>2.166</td>
<td>1.30 g</td>
<td>2.42</td>
</tr>
<tr>
<td>B</td>
<td>10.230</td>
<td>4.225</td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1.087</td>
<td>2.581</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>8.278</td>
<td>2.451</td>
<td>3.38</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>8.308</td>
<td>3.434</td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>F-U</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k.</td>
<td>Adjectival clauses and phrases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>3.454</td>
<td>1.323</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2.791</td>
<td>2.408</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>2.063</td>
<td>1.820</td>
<td>1.10 g</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>4.217</td>
<td>1.861</td>
<td>2.27 g</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2.850</td>
<td>2.767</td>
<td>1.03 g</td>
<td></td>
</tr>
</tbody>
</table>

In addition to the above computations, the investigator tabulated the number of papers containing each of the Common Errors listed in Table 1 and the total number of each of such errors in all papers in each group. Because the total number of each error has not been adjusted according to the total number of words, only the number of essays containing each error is given in Table 8.

From these results the investigator concluded, among other things, that there was "the lack of effective tie between a relatively high grammatical score and improvement in the measured items of the essay"—the highest correlation between the top third of the scores on the final grammar test with changes in score between the first and final essays was merely .229 ± .138. The investigator also expressed dissatisfaction with the low levels of achievement of the Formal Grammar group on Test C, only one class having more than 50 percent of the answers correct. That low achievement, he wrote, "may be seen as a major factor throwing doubt on the advisability of studying formal grammar in the early part of the Secondary School." Noting that his students were drawn from a range of schools and socioeconomic backgrounds, he also concluded that "The failure to profit from instruction in [traditional formal] grammar is thus not confined to any one educational environment or category of children" (of those he studied). In short, the investigator concluded, "It seems safe to infer that the study of English grammatical terminology had a negligible or even a relatively harmful effect upon the correctness of children's writing in the early part of the five Secondary Schools." (Based as it was on the use of traditional grammar, the Harris study does not necessarily prove, of course, the ineffectiveness of instruction based on structural or generative grammar.)

Table 8

<table>
<thead>
<tr>
<th>Number of Essays Containing Each Common Error</th>
<th>First Essays</th>
<th>Final Essays</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Formal</td>
<td>Direct</td>
</tr>
<tr>
<td></td>
<td>Formal</td>
<td>Direct</td>
</tr>
<tr>
<td>1. Omission of full stop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Lack of use of comma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Adj. or prep. as adverb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Wrong compar. or super.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Faulty position of adverb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. No clear antecedent for pronoun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Error of prepositions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Error of conjunctions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. No clear antecedent for pronoun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Failure to use relative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. No finite verb or omission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of an important verbal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Lack of subj.-verb agreement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Faulty sequence of tenses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Unrelated participes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Problem

Noting the widespread practice of judging a student's writing by one sample of his work, the investigator sought "to determine whether a single paper written by a student on a given topic at a particular time" constituted a valid basis for evaluation. The investigation was based on the evaluation of 320 papers written between the fall term, 1931, and the winter term, 1932, by students enrolled in the first quarter of the Written and Spoken English course (later renamed Communication Skills) at Michigan State College (now Michigan State University). The reader's attention is directed especially to the superior design of this experiment and to the care with which the theme rating was planned.
Procedure

The students whose papers were evaluated in this investigation came from the classes of two instructors, who collaborated in the selection of topics similar in nature to writing assignments used previously in the term. In order to test the hypothesis that "any given topic provides the same stimulus as any other topic," some students wrote the same day on two similar topics, some on two dissimilar topics. The similar assignments used were "It Is (Is Not) Too Far Between Classrooms for Students to Travel During the Ten-Minute Period Allowed" and "Textbooks for Freshmen at Michigan State College Are (Are Not) Too Expensive," hereinafter referred to as "Distance Between Classes" and "Cost of Textbooks." Each student writing on one of these topics was "provided a brochure of information" pertaining to that topic. A third, dissimilar, assignment was also used: "Give Directions to a Stranger at the Union Building Enabling Him to Get to Building A-6 on South Campus," henceforth referred to as "Giving Directions." Each student writing on this topic was provided with a map of the campus. (Although these topics were selected to prevent content factors from affecting the quality of the writing, the nature of these assignments may lead some readers to question whether the topics were pitched at a level representative of assignments commonly used in other colleges and whether the material provided would tend to make the writing of various students more alike than it normally would be.)

A second hypothesis was "that any given topic elicits constant responses at different times." To test this, some students wrote on the same topic on different days, seeking to determine "whether variations in efficiency from day to day have any effect on the quality of student writing."

That "the psychological pressure of the examination situation has no adverse effect on the quality of student writing" was a third hypothesis. This was tested by having some students write without such a pressure on one day and with it, on the same topic, on another day. Paralleling them, a control group wrote on the same topic without any examination pressure on either occasion.

A fourth hypothesis was "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." The investigator tested this assumption by analyzing the variations in the

quality of the writing in relation to the rank of the student's writing ability.

The design of the experiment provided for the variables and controls in the manner shown in Table 1.

<table>
<thead>
<tr>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>W/O</td>
</tr>
<tr>
<td>Topic 2</td>
<td>W/O</td>
</tr>
<tr>
<td>Group A</td>
<td></td>
</tr>
<tr>
<td>W/O</td>
<td>W/O</td>
</tr>
</tbody>
</table>

Table 1

Design of Experiment

To prevent the intrusion of extraneous influences into the experiment, the investigator instituted a number of controls to allow for fatigue, motivation, and the like. Each of the four groups of students wrote all four papers during their regular two-hour laboratory periods, when they were accustomed to write. All four groups regularly met on the same days—Tuesdays and Thursdays. The presence of a minimum of twenty students in each group could confidently be expected. To eliminate the effects of fatigue, every other student wrote on the first topic during the first hour, took a ten-minute break, and then wrote on the second topic. The other students first wrote on the second topic and then, after a ten-minute break, on the first topic.

On the first day in which the students wrote, they were all instructed that their grades for the term would not be affected by the papers they wrote that day. As a stimulus to good writing, however, they were informed that grades would be assigned to their papers by a team of raters and reported to the students for comparison to their usual grades. On the second day of writing, the students in Groups A and B were
RESEARCH IN WRITTEN COMPOSITION

given much the same instruction, again being assured that their grades would not count but in addition being informed that "The results of the grading will be posted on the bulletin board early next term." (If the students interpreted this to mean that their names would be posted with their grades, this procedure probably should be considered a substitution of social pressure for academic pressure, not an elimination of pressure.) The students in Groups C and D were informed in their instructions on the second day not only that their grades would be posted but that the grades would be considered as a part of their final examination and hence would affect their final grades for the term. It was also pointed out that, having written on the topics before, they "should be able to do a better job of writing" on them this time.

To prevent the papers written on the second day from being merely second editions of the first papers, the investigator did not wish to space them very close. On the other hand, he did not wish to space them so far apart that new learning experiences would cause more differences in quality of writing than variations in efficiency would cause. He favored the shorter time in his compromise, spacing the two days of writing one week apart.

To select reliable raters to evaluate the student themes, the investigator examined "the records of ratings by more than forty staff members of the Department of Written and Spoken English," selecting nine who "demonstrated a high degree of sensitiveseness and reliability." Of these, three were "willing and able to devote the time necessary" to rate each of the 320 papers. All three raters judged all 320 papers, rating "at the same time and under the same external conditions," employing the usual rating scale for the Written and Spoken English course at Michigan State. Each paper was rated from 1 (unsatisfactory) to 10 (superior) on each of five equally weighted categories: conventions of grammar, sentence structure, diction, organization, and content. The average of the two closest ratings became the score of each paper. But if the three ratings were equidistant (22, 24, 20) or if no two ratings came within five points of each other, the average of all three ratings was taken as the score.

To preclude influences due to variations in rater efficiency from day to day, all three graders judged all four papers of each student during the same rating period; Rater 1 judged the first paper and passed it to Rater 2, who judged it and passed it to Rater 3. (The report did not state whether or not a grader could tell, at the time of rating, what score another grader had assigned the paper, but the investigator has

SUMMARIES OF SELECTED RESEARCH

advised the writers of this pamphlet that "No rater had any indication of a score assigned by a previous rater." In addition, the papers were placed in such an order that any two themes by the same student were separated by the themes of at least two other students. (The investigator did not state whether or not the raters knew any or many of the student writers or could identify the papers of a particular group, but the investigator has advised that the students' names were removed from the themes and the papers were assigned numbers so that no grader knew what student's paper he was rating at any time.)

The raters judged papers for six hours each day, with two 90-minute periods each morning, separated by a 15- to 20-minute rest period, and a similar schedule each afternoon, for a total of three days. Some papers were eliminated from the study by random procedure as a means of equalizing the size of each group of students. These extra papers were studied for clues to the criteria by which the organization and content of the other themes on those topics could be evaluated.

As a means of seeing that the 80 students were representative of the freshman population at Michigan State, the investigator relied on the usual registration procedure for the course. Students had been admitted to registration in alphabetical order during the three-day registration period, and enrollment in each section was limited each day, automatically distributing throughout the alphabet the names of the students in each of the four groups.

It was at least as important for the four groups to be equal in their ability to write. The fact that all 80 students wrote their first theme on "Distance Between Classes" provided a basis for determining whether or not the four groups were equivalent. The mean scores of the first papers were as follows:

<table>
<thead>
<tr>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.8</td>
<td>23.4</td>
<td>25.6</td>
<td>22.4</td>
</tr>
</tbody>
</table>

An analysis of variance produced an F-ratio of 7.457 with 3/76 degrees of freedom where F = 2.73 for significance at the 5 percent level of confidence and F = 4.06 for significance at the 1 percent level of confidence. In other words, the indication was very strong that each of the four groups was representative of the total group of 80 students.

Results

To test whether or not fatigue did reduce writing efficiency, the investigator compared the scores on the Topic 1 papers written during the first hour to the scores on the Topic 1 papers written during the second
hour. Since the alternate topics were distributed systematically to every other student, it was reasonable to assume that the two groups of students were equal in writing ability. The mean score of the Topic 1 papers written during the first hour was 23.735, during the second hour was 23.925, indicating that fatigue was not a factor or that it was slightly more than offset by another, unknown factor—perhaps a "practice effect." Application of the t-test to the difference of .19 indicated that such a difference would happen by chance alone between 40 and 50 times out of a hundred cases. The investigator then concluded 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."

To determine whether or not the day-to-day efficiency or the differences in topics were affecting the caliber of the students' writing, the investigator compared the mean scores of the Group A papers and the mean scores of the Group B papers, both of these groups having written on both days without examination pressure. The means for Group A, writing on similar topics, were these:

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>24.5</td>
<td>23.9</td>
</tr>
<tr>
<td>Topic 2</td>
<td>23.7</td>
<td>21.3</td>
</tr>
</tbody>
</table>

The F-ratio obtained for 3/76 degrees of freedom was .7480, where $F=2.75$ for significance at the 5 percent level and $F=4.06$ for significance at the 1 percent level. The means for Group B, writing on dissimilar topics, were these:

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>25.4</td>
<td>23.4</td>
</tr>
<tr>
<td>Topic 2</td>
<td>21.4</td>
<td>19.2</td>
</tr>
</tbody>
</table>

Here the F-ratio obtained for 3/76 degrees of freedom was .9837—not approaching significance at the 5 percent level. In short, there being no significant difference between the means of any two sets of papers written by Group A or any two sets by Group B, the investigator concluded that, with groups of writers, there were no significant effects on the quality of writing as a result of variations in day-to-day efficiency or variations in topics assigned. (The reader should be reminded here that the three topics were all simple topics for college freshmen and that standard information was supplied to the students. Obviously, different results might well have been obtained if topics had been used which demanded more complicated analysis or more specialized knowledge. Remembering that Topics 1 and 2 were argumentative in nature, Topic 3 expository, one might reasonably conclude that this variation in mode of discourse—rather than variation in topic—had no significant effect on the average quality of the writing of a group of students.)

As a test of the effect of examination pressure on the writing of groups of students, the investigator compared the mean scores of the Group C papers as well as the mean scores of the Group D papers. The primary variables were the topics (similar for Group C, dissimilar for Group D) and the psychological pressure of the examination situation (applied to both groups the second day). Again, analyses of variance were used to seek significant differences among the mean scores of the four sets of themes for each group. The means for Group C, writing on similar topics, were these:

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>23.4</td>
<td>23.9</td>
</tr>
<tr>
<td>Topic 2</td>
<td>23.0</td>
<td>24.6</td>
</tr>
</tbody>
</table>

The F-ratio with 3/76 degrees of freedom was .2700, not approaching significance at the 5 percent level. The means for Group D, writing on dissimilar topics, were these:

<table>
<thead>
<tr>
<th></th>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>22.4</td>
<td>19.2</td>
</tr>
<tr>
<td>Topic 3</td>
<td>21.4</td>
<td>20.9</td>
</tr>
</tbody>
</table>

The F-ratio with 3/76 degrees of freedom was 1.48, not approaching the $F=2.73$ needed for significance at the 5 percent level. With no significant difference found among the means of the four sets of papers written by either Group C or Group D, the investigator concluded that neither the examination pressure, content factors (from the use of similar and dissimilar topics), nor a combination of such factors significantly affected the average quality of writing of these students.

In a further attempt to test the effect of psychological pressure, the investigator combined the scores of Groups A and B on Topic 1 on each day (writing "without pressure" either day) and combined the scores of Groups C and D on Topic 1 on each day (writing "with pressure" the second day). Reasoning that this combination of scores was justified by the lack of significant differences among the four groups on Topic 1 on the first day and by the doubling of the number of student papers
which could be used in testing the effect of pressure, the investigator obtained these means:

<table>
<thead>
<tr>
<th>First Day</th>
<th>Second Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups A and B</td>
<td>Groups C and D</td>
</tr>
<tr>
<td>24.9</td>
<td>22.9</td>
</tr>
</tbody>
</table>

An analysis of variance yielded an F-ratio of .7957 with 3/156 degrees of freedom where F = 2.67 for significance at the 5 percent level of confidence. This analysis thus provided further evidence that examination pressure "had no effect on the average quality of writing by the groups of students involved" and, in the case of Groups A and B (writing without examination pressure on either day), that there was no significant variation in level of efficiency from day to day in the average quality of writing by these students.

The investigator pointed out that his preceding findings concerned groups of students but that individual variations may be hidden by cancelling each other in group means. Consequently, he also examined differences between scores on papers by the same student. Recognizing that such differences may be produced by the unreliability of ratings as well as by variations in the quality of student writing, and further recognizing that ratings by the same two raters were not always represented in the scores on papers, the investigator used a procedure developed by Robert Ebel for determining the reliability and error for sets of ratings.* First the investigator computed the overall reliability and standard error of the ratings "in order to make more accurate comparisons of a student's obtained scores on different topics".

<table>
<thead>
<tr>
<th>Standard Error</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>3.2031</td>
</tr>
<tr>
<td>Topic 2</td>
<td>3.4563</td>
</tr>
<tr>
<td>Topic 3</td>
<td>2.5416</td>
</tr>
<tr>
<td>Total</td>
<td>3.1203</td>
</tr>
</tbody>
</table>

Then, using a procedure described in detail on pages 75-76 of the study, the investigator determined that differences between two papers greater than certain "theoretical differences" would have to be obtained for the difference in quality of writing to be significant:

*Robert L. Ebel, "Estimation of the Reliability of Ratings," Psychometrika, XVI (December, 1951, 460-474. 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." G.L.K.

SUMMARIES OF SELECTED RESEARCH

<table>
<thead>
<tr>
<th>Topic</th>
<th>Theoretical Difference</th>
<th>Minimum Difference Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.88</td>
<td>9.00</td>
</tr>
<tr>
<td>2</td>
<td>9.58</td>
<td>10.00</td>
</tr>
<tr>
<td>3</td>
<td>7.04</td>
<td>7.50</td>
</tr>
<tr>
<td>Total</td>
<td>8.65</td>
<td>9.00</td>
</tr>
</tbody>
</table>

Each of the following pairings of the four papers by each student was examined to see whether the differences equaled or exceeded the pertinent minimum difference:

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

The difference for each student on each pairing is reported in Appendix C of the study. Since "no more than 16 such variations could be expected as a result of the unreliability of the ratings" (the standard error of the difference between paired scores having been computed at the 5 percent level of confidence), it is striking to note that 86 significant variations in quality occurred out of 320 opportunities (4 pairings times 80 students). The 80 variations were distributed among 47, or over 56 percent, of the students involved.

The 80 students were also ranked according to their writing ability, using the student's highest score (of the four papers) as a basis for the ranking, reasoning that the consequent distortion resulting from error in the ratings was not as serious in this study as distortion resulting from variations in performance. However, as a check against the first type of distortion, the investigator also ranked the students according to the average of all four scores. The reranking resulted in a shift of only four students from the top half to the bottom half of the group. However, while 15 of the 20 students lowest in the first ranking remained in the bottom fourth when reranked according to average score, only 11 of the 20 students highest in the first ranking remained in the top fourth when reranked, suggesting that quality of writing varies more for better than for worse writers. The following "frequencies of significant variations" were observed in the groupings of students:

<table>
<thead>
<tr>
<th>Top 20</th>
<th>Bottom 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using the highest score:</td>
<td>30</td>
</tr>
<tr>
<td>Using the average score:</td>
<td>22</td>
</tr>
<tr>
<td>Using the highest score:</td>
<td>63</td>
</tr>
<tr>
<td>Using the average score:</td>
<td>55</td>
</tr>
</tbody>
</table>
Noting that use of the average score for ranking tended to reduce the frequency of significant variations, the investigator used the highest score of each student as the basis for his subsequent analysis of differences between better and worse writers.

Further observations based on the above data included these:
1. There were more than four times as many significant variations in the quality of writing of the top 20 than in the bottom 20.
2. Of the 50 variations of two or more letter grades, 43 occurred among the upper 40 students while only seven occurred among the lower 40.
3. Of the 31 students having variations of two or more letter grades, 25 were among the upper 40 students, only six among the lower 40. Of the 25 from the upper group, 15 had two or more such variations; of the six from the lower 40, only two had two or more such variations out of the four such variations possible.

As a result of such observations, the investigator concluded “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.”

After examining the validity of evaluating student achievement in writing by a single paper, the investigator compared content factors (from different topics) and day-to-day efficiency as sources of variations in quality of writing. He tabulated the number of significant variations in quality of writing “between days on the same topics and between topics on the same days”:

<table>
<thead>
<tr>
<th></th>
<th>Between Days</th>
<th>Between Topics</th>
<th>Combined Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 20</td>
<td>17</td>
<td>19</td>
<td>36</td>
</tr>
<tr>
<td>Bottom 20</td>
<td>8</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>Upper 40</td>
<td>28</td>
<td>35</td>
<td>63</td>
</tr>
<tr>
<td>Lower 40</td>
<td>11</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>Total Group</td>
<td>39</td>
<td>47</td>
<td>86</td>
</tr>
</tbody>
</table>

If content factors and day-to-day efficiency were equally responsible for significant variations in quality of writing for the total group, the 86 variations would be divided into 43 for each factor. But application of the chi-square test indicated, at the 5 percent level of confidence, that the observed frequencies of 47 and 39 were not significantly different from 43. On the other hand, the differences observed between the top and bottom, and between upper and lower students, were significant at the 1 percent level. Hence the investigator concluded that content factors and day-to-day efficiency “were responsible for comparable variations in the quality of student writing” for the total group but not for comparable variations in the quality of writing by high and low students.

The above tabulation for “between topics” included both similar and dissimilar topics. To ascertain whether or not these types of topics have comparable effects on the quality of student writing, the investigator made a new tabulation of the significant variations observed “between topics”:

<table>
<thead>
<tr>
<th></th>
<th>Dissimilar Topics</th>
<th>Similar Topics</th>
<th>Combined Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 20</td>
<td>9</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>Bottom 20</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Upper 40</td>
<td>15</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>Lower 40</td>
<td>10</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Total Group</td>
<td>25</td>
<td>14</td>
<td>39</td>
</tr>
</tbody>
</table>

If similarity and dissimilarity of topics were equally responsible for significant variations in quality of writing for the total group, the 39 variations would be divided into 19½ for each set of topics. But application of the chi-square test indicated, at the 5 percent level of confidence, that the observed frequencies of 25 and 14 were not significantly different from 19½. However, although the differences observed between similar and dissimilar topics was negligible for the top 20 and upper 40 groups, it was significant for the lower 40 students. Hence, the investigator concluded that although both the similarity and dissimilarity of topics were about equally responsible for variations in the quality of writing by better writers, only the dissimilarity of topics was responsible for significant variations in the quality of writing by worse writers. (The reader should be reminded here that the “similar topics”—Topics 1 and 2—both demanded a sort of low-level argumentative mode of discourse, and Topic 3—involved in the pairing of “dissimilar topics”—demanded the expository mode. In other words, the variations in the quality of writing of worse writers may result more from a change in mode of discourse than from a change in subject matter. The subject matter of all three topics was strongly similar in that in each case it involved matters commonplace for Michigan State freshmen.)
Finally, in order to determine whether or not examination pressure had an adverse effect on student writing, the investigator tabulated the gains and losses in the scores of the "pressure" and "nonpressure" students from the first day of writing to the second.

<table>
<thead>
<tr>
<th>Nonpressure Groups A and B</th>
<th>Gains</th>
<th>Losses</th>
<th>Combined Gains and Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper 40</td>
<td>6</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Lower 40</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Total Group</td>
<td>9</td>
<td>13</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pressure Groups C and D</th>
<th>Gains</th>
<th>Losses</th>
<th>Combined Gains and Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper 40</td>
<td>10</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Lower 40</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Total Group</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Grand Total</td>
<td>24</td>
<td>23</td>
<td>47</td>
</tr>
</tbody>
</table>

Although more losses than gains were observed among the nonpressure groups and more gains than losses among the pressure groups, application of the chi-square test indicated, at the 5 percent level, that the differences could be attributed to chance variations. The frequencies of significant gains and losses for the lower half of the students were too small to permit statistical analysis, but no pronounced trends were observed. In short, the investigator concluded 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."

In summary, where a group of college freshmen is concerned, the average quality does not seem to be affected by variations in the day-to-day efficiency of individuals or by the pressure of the examination situation. Although the study indicated that the average quality was not affected by the nature of the topics assigned, there is basis for discounting this conclusion in favor of the notion that variation from expository to argumentative mode of discourse does not seem to affect the average quality of the writing of a group of freshmen.

On the other hand, variations in the quality of writing of individual freshmen did seem to result from variations in the day-to-day efficiency of students, especially of the better writers, but the pressure of final examination conditions did not appear to affect adversely the writing of either the better or the worse writers. Although the writing of the better writers did not seem to vary significantly in quality according to whether the students were writing expository or argumentative papers, the quality of writing by the worse writers did reveal a significant number of variations according to mode of discourse.

Among the implications which the investigator drew from his findings were these:

1. The justification seems very doubtful for grading a college freshman's achievement at the end of a writing course on the basis of one writing sample, especially for better writers.
2. An evaluation of the overall, or average, group improvement resulting from a writing course may be obtained from a single pretest theme and a single post-test theme.
3. An evaluation of "the effectiveness of different teaching methods with students who vary in writing ability" should be based on several samples of writing from each student, writing on different topics (modes of discourse) and during different days, the same procedure being followed both with the pretest and the post-test themes, the nature of the assignment being similar if not the same on both these occasions.
4. The post-test papers may be used as a final examination without seriously affecting the results.

The Smith Study

Smith, Dora V. *Class Size in High School English: Methods and Results.* Minneapolis: University of Minnesota Press, 1931. 309p.

Problem:

The purposes of this study, conducted in the ninth grade at the University High School of the University of Minnesota, were "(1) to discover the effect of the size of the class upon the efficiency of instruction in ninth grade English and (2) to devise techniques whereby a large group may be handled effectively with a minimum of waste of time and activity and a maximum of attention to and response from the individual." Although the investigation was conducted during 1925-1927, the problem and procedures are meaningful today, especially in view of current explorations in the use of lay readers to help English teachers handle composition marking for large classes. In this study, however, the investigator, who taught the large and small classes involved, had "no assistance in the actual conduct of the class nor in the handling of individual pupils," including the marking of compositions. Moreover, although it is not explained in the published report, the investi-
gator was not concentrating most of her efforts on teaching the experimental and control classes; in addition to them, she was teaching a class of senior English with 32 pupils, sharing substantially the teaching of a two-hour graduate seminar containing 21 experienced teachers, supervising the work of nine student teachers, acting as adviser to the freshman class, and chairing a faculty curriculum committee. In short, the study is a report of what can be achieved with a large class by a teacher generally known to be highly proficient and energetic but teaching with at least a normal instructional load.

When the study was being considered for inclusion in this report, a question was raised about how representative the University High School ninth graders would be of school pupils in general. Her published study does not bring out this point, but the investigator has informed the writers of this report that the ninth graders had come from thirty different Minneapolis schools for this, their first year in the University school. The fact that pupils applied for admission to the University High School, where undoubtedly tuition was charged, and the pupils' average IQ of 116 suggest that the investigation was conducted with groups of superior intelligence and socioeconomic background. Consequently, the conclusions of this investigation have relevance to superior ninth grade pupils rather than to ninth graders representative of the American school population as a whole.

This experiment was carried on and evaluated for two years, but many of the techniques for dealing with the large class were tried during the first year for the first time. Furthermore, the teaching included literature and the mechanics of reading as well as written composition. The following summary will, of course, focus on the composition aspects of the experiment during the second year. In some matters, however, this focus cannot be maintained; sometimes statistics were reported only for the two-year period in such a manner that the data for the second year cannot be separated from the data of the first year.

In addition to the above qualifications, the reader's attention is directed to the complex but careful procedures used in matching the pupils in this experiment. Concerned throughout with English in general, not solely with written composition, this study nevertheless investigated the composition in a total English class setting with a care which merits attention today even though, thirty years later, the reader will desire additional studies of class size which may corroborate or refute the findings of this investigation.

### Table 1

<table>
<thead>
<tr>
<th>Test</th>
<th>Range</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average I. Q.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>96-136</td>
<td>115.90</td>
<td>10.52</td>
</tr>
<tr>
<td>Large</td>
<td>97-142</td>
<td>116.43</td>
<td>10.62</td>
</tr>
<tr>
<td>Chronological Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>157-189</td>
<td>167.43</td>
<td>7.26</td>
</tr>
<tr>
<td>Large</td>
<td>155-186</td>
<td>167.14</td>
<td>9.20</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>65-128</td>
<td>96.67</td>
<td>17.09</td>
</tr>
<tr>
<td>Large</td>
<td>75-128</td>
<td>96.71</td>
<td>14.28</td>
</tr>
<tr>
<td>Grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>30-90</td>
<td>63.38</td>
<td>17.15</td>
</tr>
<tr>
<td>Large</td>
<td>30-85</td>
<td>61.86</td>
<td>13.93</td>
</tr>
<tr>
<td>Composition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>3.4-6.7</td>
<td>5.19</td>
<td>.75</td>
</tr>
<tr>
<td>Large</td>
<td>4.4-6.5</td>
<td>5.36</td>
<td>.63</td>
</tr>
<tr>
<td>Spelling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>69-133</td>
<td>118.81</td>
<td>14.06</td>
</tr>
<tr>
<td>Large</td>
<td>97-134</td>
<td>118.00</td>
<td>8.83</td>
</tr>
</tbody>
</table>
based on an initial composition written and rated according to the instructions for the Hudelson TypiCAL Composition Ability Scale, which requires the reproduction of a narrative ("Slatter's Hill Snowball Fight"). The compositions were rated by five teachers, none of whom knew which papers were written by which class (although that was not made clear in the published report). The average score was used in matching the pupils. No data were offered on reader reliability, but the investigator reported that "all tests used in the experiment were rescored by an outside scorer before any computations were made." Scores for each pupil on each pretext are given in the published report, from which the figures in Table I are taken. The investigator further reported that "in no case did the difference exceed the probable error of the means except in composition. From the standpoint of significance, the chances that an equal difference would occur outside of these limits were even for chronological age, reading ability, grammar, and spelling. The chances for the differences in IQ and composition were 1.18 to 1." Thus, in matching pupils for all of the aspects of English studied, the large group was slightly favored with respect to composition.

The course of study, "identical in both classes," consisted of the ninth grade units usual for the University High School at that time:

First Quarter.—Minimum essentials of English grammar and spelling: a short story unit with Rosa Mikel's Short Stories for English Classes as a starting point; composition and the monthly publication of a freshman paper.

Second Quarter.—A unit on punctuation; minimum essentials of grammar (continued); the Merchant of Venice and other plays; composition and the continuation of the freshman paper.

Third Quarter.—A unit in letter writing; a project in the literature of chivalry with Ivanhoe as a basis; a poetry unit based upon Sparkling and Sparkling's Open Gates; composition and the continuation of the freshman paper.

Although the investigator taught the two classes herself, she had three observers sit in each class daily and record "pupil participation in each class, the attitude of teacher and pupils, the physical features of the room, the time consumed in movements to and from the blackboard or in small-group activities, and the time devoted to each section of the lesson in each class." (The instructions given to observers are reproduced on pages 278-279 of the published report.) The observers, all of whom were university seniors training to be English teachers, were shifted every three weeks, at the end of which they summarized their observations. The summaries, taken as a whole, thus "approximated a stenographic report of each day's lesson" throughout the year.

SUMMARIES OF SELECTED RESEARCH

Rather than attempt to employ the usual methods of instruction in both classes, and thus handicap the large class with ill-adapted procedures, the investigator devised a number of techniques for use with the larger group, reporting them in Chapter VII. Sometimes the concomitant procedures for the small group were described or implied in the published report; sometimes they were not.

The large class was divided into six groups, each of which was drawn from the complete range of ability available and was seated together in the classroom. The locations were shifted each six weeks to equalize opportunity for favorable seating. Each group had an especially capable pupil as leader responsible for taking the attendance and collecting and distributing the papers of the eight or nine pupils in his group. Various other techniques were employed to facilitate the teaching of literature, reading, and speech to the large class.

When the large class had written compositions, the investigator had the pupils meet in their groups to read their papers aloud, select the best one to read aloud to the class, and make suggestions to each other for improving their writing. She noted that "every pupil was able to read his composition aloud . . . and, frequently on the same day, the pupils chosen could also give theirs before the class as a whole." A list of questions, mostly concerning rhetorical considerations but also sentence structure and mechanics, was given to the pupils before they wrote to guide them in their selections and suggestions. (See pages 157-158.) The investigator attested that this type of provision was done with especial care before the pupils wrote business and social letters, and "So far as possible, letters were at least begun in class under the direction of the teacher."

The pupils also drilled or recited on grammar problems in their groups, limiting the material "to those portions of English grammar that, in the absence of scientific evidence, are believed to be functional in correct speech and writing." The investigator carefully instructed the pupils, when noting that a fellow group member had erred, "to question him in regard to the tests he applied before making his response," helping him to "think out" the correct response.

With the large class, the investigator found it more necessary than ever to break the 50-minute class period into three or more different activities and not to spend the entire time on "a procedure of one sentence around the class." She reported that she had to go into the classroom with a "definite program that permitted a maximum of participation for the pupils," a definiteness which was "especially necessary if she was
to make quick transitions from one part of the lesson to another with a minimum of confusion and waste of time." Still it was not possible, of course, to have as much pupil participation in the whole large class as in the small class. While fourteen lessons in grammar were being held during October, an actual count of pupil recitations was kept ("answering or asking an oral question during regular class discussion or volunteering some comment thereupon"). The recitations averaged 1.55 per pupil per day in the large class, 2.55 in the small class, with only one pupil in the large class reciting as often as the average pupil in the small class. The investigator reported that if the questions were to be distributed fairly in the large class, it was necessary to call on pupils most of the time and "pay little attention to volunteers."

Rather than have the large class recite as one group, the investigator often used the group recitation technique described in the paragraph before the preceding one. Then each pupil recited three or four times in 15 or 20 minutes, covering the two dozen or so sentences assigned for homework. The investigator went from group to group, settling disagreements and noting which three or four sentences caused enough problems to discuss with the pupils when the class reconvened as a whole. After this entire procedure (which took from 20 to 25 minutes), the investigator often administered a brief test before turning to new work. She remarked on the increased attention, interest, training in group cooperation, and intergroup competition but made no comment on whether or not the good pupils suffered from continually helping the poor pupils in this procedure. Somewhat similar group procedures were used in seeking more descriptive adjectives to use in writing.

The large class responded with more zeal than the small class to "anything in the nature of a contest," especially a spelling contest. To prevent the possibility of "harm to a nervous child with a special disability in spelling," the investigator saw to it that "No lesson . . . contained more than ten words selected from the minimum essentials list for the ninth grade in Minnesota."

During a two-week letter writing unit, the groups became business firms, each with a manager, overseer, and messenger having his own special duties. Although the pupils occasionally marked and corrected each other's papers, the investigator always gave the final score. The group envelope contained the list of pupils, the types of letters written, and a system of accounts: ¹⁰ for an A letter, ⁸ for a B, ⁶ for a C, ⁴ for a D, a ⁴ deduction for each late letter, and a dollar bonus for each firm having every letter finished on time. Each pupil wrote one or two letters each day during this unit. The amount of paper correction was so heavy at this time that the investigator could not return to each pupil in the larger group all his letters the day after he wrote them.

Various other composition projects were pursued in both classes, but they did not seem different because of the size of the class except in one respect: when compositions were being chosen to be read at a holiday program, more good stories could be found in the larger class but that number was still a smaller proportion of the large class than the proportion of papers used from the small class.

A major problem was providing for individual differences in the large class. One technique was to have each pupil record his scores from quarter to quarter on examinations on various aspects of the work, stimulating self-competition. Another was to group pupils for drill according to their needs, the investigator working with weaker pupils drawn from each group while the others proceeded as usual under the direction of their regular leaders. A third device was to seat in the front row the pupils who had experienced difficulty the previous day and to call on these pupils for review questions. Volunteers were occasionally permitted to take these seats to benefit from the review—and to help the class to welcome the device. The very nature of the composition work permitted individual tastes to operate in the selection of topics and individual needs to be met in the marking of papers. At the end of each quarter, the investigator had individual conferences with the pupils, reviewing their compositions (and an error check sheet occasionally used) and pointing out constant repetitions of errors. Pupils with similar repetitions were assigned to the same row of seats for laboratory work, permitting the investigator to give them special help.

Brief tests and exercises were given very frequently to compensate for less opportunity to participate in the large class. When occasionally such tests were written on the blackboard, the investigator filled the blackboard with writing large enough for all pupils to see; however, the constant looking up and down at the blackboard seemed to introduce errors not found when tests were mimeographed and was conducive to wandering glances at other pupils' papers. The pupils corrected most of these tests themselves and took interest in tabulating class errors and suggesting the next steps in teaching. The pupils also corrected each other's homework exercises in punctuation.
Results

The results of the experiment were tested at the end of the year by a number of measures, but only those results relating to composition are given here. Only the matched pupils in the large class are considered in comparison to the pupils in the small class.

Five objective tests were given in capitalization, punctuation, and spelling. The large class was slightly superior to the small one in capitalization and more clearly so in spelling, but the two classes were practically even in punctuation. Another five objective tests were administered to measure results in grammar, with no clear difference in favor of either class.

The investigator reported another measure which, at first examination, seems to be a more valid test of composition ability than the objective tests—mean total points for the approximately 30 papers per pupil written during the letter writing unit. Unfortunately, however, the investigator herself was the sole rater of the letters and the ratings of all the letters were used (not merely several of the latest for each pupil, reflecting the end result of differentiated instruction), largely discounting the significance of her finding that the small class was distinctly superior in this measure.

Double the most valid measures of composition ability were the Hudsonel Typical Composition Ability Scale (reproduction of a narrative) and the Hudsonel English Composition Scale (original narrative on "The Most Exciting Ride I Ever Had"). The results indicated that the small group scored higher (not merely the mean scores but the distribution of scores) as a whole than the large group, the difference being statistically significant. The investigator noted that the small group was superior in both grammar and vocabulary, with the large group stronger in organization and style.

In Table 2, a slightly higher proportion of the pupils in the small classes surpassed their mates, but the pupils in the large class who surpassed their mates did so by a wider margin, especially in the reproduction of a story. It seems clear that, considered in their respective groups, these superior ninth grade pupils showed little difference in their proficiency at writing narrative compositions as a result of being in a larger or smaller English class when concerted efforts were made by a superior teacher to assure individual participation and attention in the larger class.

In addition to measuring the writing ability of the pupils, the investigator explored their attitudes toward small class and large class instruction. Six months after the experiment had begun, and without advance warning which would have permitted them to be influenced by each other or their parents, they were given a questionnaire to learn their initial and present reactions. The pupils were told that the questionnaire was being given to provide data for another investigator, and one of the pupils collected the completed papers and took them to his office without examination by the teacher.

Although the investigator noted that "the experiences of the 1925-1926 group may have influenced the attitude of incoming students the next year," such an influence doubtless would exist in any school where large classes were used year after year. Consequently, the results only for the 1926-1927 group are given here (when they can be isolated), and, since the published study reports for all pupils were the subject of the investigation, the pupils in the large group, the reactions of the small class are compared to those of the entire large class. Because of absences on the day the questionnaire was filled out, only 46 pupils in the large class and 17 in the small class are included.

When asked on the questionnaire which class they had hoped to be placed in the previous September, 24 percent of the pupils in the large class indicated a preferred preference for the large class, 50 percent said they would have chosen the small class then, and 26 percent voiced no preference. After six months of experience in the class, 43 percent of the large class preferred it, 34 percent would still choose the small class, and 22 percent indicated no preference. The pupils whose preference had shifted from the small to the large class gave as their chief reasons (1) that they learned just as much, (2) that they heard more opinions expressed and shared more varied talents, (3) that, contrary to their expectations, they received just as much help in the large class in English as they did in the small classes in
other subjects, and (4) that they became acquainted with more members of the freshman class." (The third reason, of course, is suspect here; we have no evaluations of the teaching proficiency of the teachers in other subjects—whose abilities may not have equalled the investigator's—and hence we cannot attribute the help necessarily to the large class per se.) Of the reasons given by the large-class pupils for disliking the larger group, the chief were that there was (1) less "opportunity for participation and individual attention and (2) more noise and confusion from many pupils working in the same room at the same time." Four students missed a feeling of intimacy within the larger group. Only one student referred to difficulty hearing during the group activities, although more mentioned such a problem when the class met as a whole. Thirty-five percent of the large class preferred that class specifically for composition work, but 9 percent preferred the small class for that activity; over half of the large class evidently did not think the difference important enough to mention.

The investigator also studied the responses of her observers to twelve questions each was asked to answer. When asked in which class the teacher and pupils knew each other better, 65 percent of the observers stated that there was no difference and 20 percent noted that, although there was no difference between teacher and pupils, the pupils in the large class did not know each other as well as those in the small class knew each other. Although one-fourth of the observers noted more shyness in the small class, one-third saw none in either class, and one-third saw the same amount in both. Forty percent of the observers thought the pupils in the large class asked more questions, but 60 percent indicated that the number of questions was equal in the two classes. The most striking agreement among the observers related to the spirit of each class, four-fifths making a strong endorsement of the superior interest and enthusiasm of the large class, one-fifth noting that the spirit of the two classes was the same. Citing the difficulty of summarizing evidence on the discipline in the two classes "because of varying views as to what constitutes 'good discipline,'" the investigator nevertheless offered her own observation that "Restlessness and inattention when the class met as a unit for general discussion were on the whole more prevalent in the large class than in the small one, and, once started, were harder to quell." Doubtless it was not entirely such situations, however, which led all of the observers but one to state emphatically that the large class was more difficult than the small one to manage, even though the investigator herself affirmed that she "entered the large class
with more of a feeling of responsibility and with a greater sense of adventure than she did the small one." The investigator's candid revelation of such feelings may lead some readers to believe that her objectivity in reporting was extended to an objectivity in treating both classes the same insofar as possible; other readers, however, may believe that a natural interest in the success of the large class would contribute, even with a person as fair-minded as this investigator, to the Hawthorne effect—the tendency of any experimental group to surpass its usual efforts because an experiment heightens its interest in learning.

Finally, the investigator made some of her own observations on the teacher's load and the expense to the school system. She thought that it did not take her quite two and a half times as long to correct 50 papers as to do 20, or to plan "group activities and assignments where a topic is arranged for each individual member of the group." She also noted that a teacher would seldom wish to meet a class of 50 without special planning, though he might occasionally do that with but 20 pupils in a class. Although she paid tribute to the value of "great quantities" of mimeographed materials in the large class, she also stated that, when not in an experimental situation necessitating much testing, she probably could dispense with half of what she used and, at much less expense than the salary of another teacher, obtain clerical help in checking diagnostic check sheets and objective tests. In fact she estimated, using the pay rates of 1926, that a school could save 28 percent of the teacher expense if it assigned teachers three classes of 50 pupils each and clerical assistance for three hours a day rather than five classes of 20 pupils each and no clerical assistance, but she questioned "how many classes of 50 a teacher may teach in one day without harm to herself or to the pupils."

In conclusion, the Smith study, although probably affected by the over-involvement of this obviously remarkable teacher-investigator with the actual conduct of the experiment, still offers strong indication that the teaching of composition to superior ninth graders may be generally as effective in large classes as in small classes if the teacher has the requisite ability and is given adequate time for preparing lessons and adequate assistance in handling the routine aspects of class management. Whether large class teaching may be further made practicable for more typical teachers and pupils by the judicious use of competent lay readers and modern devices like the overhead projector remains to be demonstrated in a carefully controlled experiment. But when further investigations are made of the effectiveness of teaching composition to large classes, the teaching techniques and experimental procedures of the Smith study will offer many suggestions and a few noteworthy cautions.

The Becker Study

Problem
In this University of Iowa freshmen program* integrating the teaching of reading, writing, speaking, and listening with a staff largely composed of part-time graduate assistants, an experimental comparison was made in 1957-58 of the effectiveness of three methods of teaching: (1) the Normal method, making one instructor responsible for the training of his students and meeting them in class four times each week, (2) the Bibliography method, making the student more "self-reliant and independent" by reducing from four to three the number of formal class meetings per week, by presenting the basic principles of the course . . . in a bibliography of assigned and collateral readings, and by devoting most of the class hours to performances under the guidance and criticism of a classroom instructor, and (3) the Kinescope method, presenting "the basic principles of the course by 'experts' through the use of film recordings of television programs (kinescopes), supplemented by discussion of these principles and their application in performances under the guidance and criticism of the regular classroom instructor." The Kinescope classes met four times each week, at one meeting of which a 29-minute kinescope usually was shown and discussed. Essentially the same material was presented under each method. (See the Schedule of Kinescopes on a later page.)

As he reads this summary, the reader is invited to note the careful statistical procedure used and the possibility that results which are not statistically significant may be meaningful for that reason. He is also advised to observe that some of the unsatisfactory aspects of this experi-

---

*The name of the program was changed from the Communication Skills Program to the Rhetoric Program in 1961.
ment might have been prevented if a pilot investigation had preceded the regular study.

Procedures

All three methods were used in the regular two-semester course (10:1 and 10:2), but only the Normal and Bibliography methods were used in the accelerated one-semester course (10:3). Classes were apportioned equally among the methods at each hour, and students were randomly assigned to the classes at the hour for which they registered. Except that "instructors teaching more than one section of the course were assigned to classes all within one treatment group" and "an attempt was made to achieve a balance of experienced and inexperienced instructors," instructors were randomly assigned to classes. If a class did not have the same instructor in both 10:1 and 10:2, it was eliminated from the experiment. If an instructor taught more than one section, only one was counted in the experiment, facilitating control of instructor differences. If complete pretest and post-test scores could not be obtained for a student, his scores were eliminated from consideration. In the final data, there were 758 students represented, 585 in the two-semester course and 178 in the one-semester course, divided among the methods approximately as explained above.

Detailed, day-by-day syllabi were distributed to the instructors and students of the Bibliography classes (10:1, 10:2, and 10:3) and the Kinescope classes (10:1 and 10:2). These five syllabi were prepared by the supervisors of the course, guided by the suggestions of Communication Skills staff committees which met during the spring and summer preceding the year of the experiment. (Although it is not reported in the published study, these syllabi were not distributed to the staff until a few days before instruction began, making it difficult for the instructors to familiarize themselves with the course plan or the bibliographic references.) The Bibliography syllabi included the bibliographies of collateral readings on the topics covered in the course. At least one copy of each reference was placed on "reserve" circulation at the University Library. "Methods of stimulating these sections were left to the instructors of these classes." Excerpts from the syllabi are given in Appendix B in the full report. In 10:1, the emphasis was on expository writing and speaking, and reading and listening to exposition. In 10:2, the emphasis shifted to argument and, to a lesser extent, criticism. In 10:3, the emphases of both 10:1 and 10:2 were covered. The textbooks were the same for all three courses:

---

SUMMARIES OF SELECTED RESEARCH


Twenty-eight half-hour kinescopes were used in the kinescope classes, all but two prepared for this purpose at the University of Iowa. Simultaneous showing of the kinescopes in two different classrooms was made possible by printing two copies of each film. No more than three classes (with a maximum of 25 students per class and an average less than 24) gathered in each screening room for showings, and after the film was shown they returned to their regular classrooms for discussion of the topic. (Although it is not mentioned in the published report, it is true that the instructors rarely previewed the kinescopes before their students viewed them. Before some kinescopes were shown, the instructors were supplied with outlines of the content but rarely with suggested questions to use in guiding class discussion. It is not surprising, then, to learn that a few instructors considered the kinescopes little more than an interruption of their regular teaching.) The topics of the kinescopes ranged from the theoretical to the practical, but the larger proportion was practical. Some 16 of the kinescopes may be said to concern communication in general or to relate to speech as well as to writing; 5 relate directly to writing; 7 relate mostly to speaking, reading, or listening. The schedule for the kinescopes follows:

**Schedule of Kinescopes 1957-58**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Date Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Ideas in Communication&quot;</td>
<td>Oct. 8, 1957</td>
</tr>
<tr>
<td>&quot;Framework of Ideas&quot;</td>
<td>Oct. 15, 1957</td>
</tr>
<tr>
<td>&quot;Adjustment to Communication&quot;</td>
<td>Oct. 22, 1957</td>
</tr>
</tbody>
</table>

**Topic**

- "The Nature of Communication"
- "Ideas in Communication"
- "Framework of Ideas"
- "Adjustment to Communication"

**Date Shown**

- Oct. 2, 1957
- Oct. 8, 1957
- Oct. 15, 1957
- Oct. 22, 1957
RESERCH IN WRITTEN COMPOSITION

<table>
<thead>
<tr>
<th>Topic</th>
<th>&quot;Expert&quot;</th>
<th>Date Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Concepts of Purpose&quot;</td>
<td>Dr. John G. Gerber, Professor, English, SUI</td>
<td>Oct. 29, 1957</td>
</tr>
<tr>
<td></td>
<td>Dr. Carl A. Dallinger, Associate Professor, Speech;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coordinator of Communication Skills, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Bodily Action&quot;</td>
<td>Dr. Clark Griffith, Assistant Professor, English, SUI</td>
<td>Nov. 1, 1957</td>
</tr>
<tr>
<td>&quot;Paragraphing&quot;</td>
<td>Dr. Harry H. Crosby, Assistant Professor and Writing Supervisor,</td>
<td>Nov. 12, 1957</td>
</tr>
<tr>
<td></td>
<td>Communication Skills, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Basic Patterns of Organization&quot;</td>
<td>Dr. William Eller, Associate Professor and Director of</td>
<td>Nov. 19, 1957</td>
</tr>
<tr>
<td></td>
<td>Reading Laboratory, Education, Reading Supervisor, Communication</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skills, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Effective Reading&quot;</td>
<td>Dr. John C. Gerber, Professor, English, SUI</td>
<td>Dec. 3, 1957</td>
</tr>
<tr>
<td></td>
<td>Mr. Cleo Martin, Instructor, Communication Skills, SUI</td>
<td>Dec. 10, 1957</td>
</tr>
<tr>
<td>&quot;The Support of Ideas&quot;</td>
<td>Dr. John C. Gerber, Professor, English, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Research Techniques&quot;</td>
<td>Dr. William Eller, Associate Professor and Director of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Laboratory, Education, Reading Supervisor, Communication</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skills, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Effective Sentences&quot;</td>
<td>Dr. William Eller, Associate Professor and Director of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reading Laboratory, Education, Reading Supervisor, Communication</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skills, SUI</td>
<td></td>
</tr>
<tr>
<td>&quot;Documentation and Manuscript Techniques&quot;</td>
<td>Mr. William J. Holmes, Instructor, Communication Skills, SUI</td>
<td>Jan. 7, 1958</td>
</tr>
<tr>
<td>&quot;The Process of Revision&quot;</td>
<td>Dr. Clark Griffith, Assistant Professor, English, SUI</td>
<td>Jan. 10, 1958</td>
</tr>
</tbody>
</table>

Second Semester, 10:2 Classes

<table>
<thead>
<tr>
<th>Topic</th>
<th>&quot;Expert&quot;</th>
<th>Date Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Difference Between Words and Things&quot;</td>
<td>Dr. Irving J. Lee, Professor, Speech, Northwestern University</td>
<td>Feb. 13, 1958</td>
</tr>
<tr>
<td>&quot;Voice in Communication&quot;</td>
<td>Dr. Harold G. Shiffer, Assistant Professor, Dramatic Art, SUI</td>
<td>Feb. 18, 1958</td>
</tr>
<tr>
<td>&quot;Why People Misunderstand Each Other&quot;</td>
<td>Dr. Irving J. Lee, Professor, Speech, Northwestern University</td>
<td>Feb. 25, 1958</td>
</tr>
<tr>
<td>&quot;Articulation and Pronunciation&quot;</td>
<td>Dr. Harold G. Shiffer, Assistant Professor, Dramatic Art, SUI</td>
<td>Feb. 28, 1958</td>
</tr>
<tr>
<td>&quot;Making Language Appropriate and Effective&quot;</td>
<td>Dr. Clark Griffith, Assistant Professor, English, SUI</td>
<td>March 7, 1958</td>
</tr>
<tr>
<td>&quot;Mass Media as Background in Communication&quot;</td>
<td>Dr. Richard Braddock, Assistant Professor, Communication Skills, SUI</td>
<td>March 18, 1958</td>
</tr>
</tbody>
</table>

SUMMARIES OF SELECTED RESEARCH

<table>
<thead>
<tr>
<th>Topic</th>
<th>&quot;Expert&quot;</th>
<th>Date Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Obstructions to Critical Listening and Reading&quot;</td>
<td>Dr. William Eller, Associate Professor and Director of Reading Laboratory, Education, Reading Supervisor, Communication Skills, SUI</td>
<td>March 20, 1958</td>
</tr>
<tr>
<td>&quot;The Argumentative Situation&quot;</td>
<td>Dr. Orville A. Hitchcock, Professor, Speech, SUI</td>
<td>March 25, 1958</td>
</tr>
<tr>
<td>&quot;Evidence, the Foundation of Reasoning&quot;</td>
<td>Dr. Orville A. Hitchcock, Professor, Speech, SUI</td>
<td>April 1, 1958</td>
</tr>
<tr>
<td>&quot;Processes of Reasoning&quot;</td>
<td>Dr. Carl A. Dallinger, Associate Professor, Speech; Coordinator of Communication Skills, SUI</td>
<td>April 10, 1958</td>
</tr>
<tr>
<td>&quot;Discussion, Its Basis and Forms&quot;</td>
<td>Dr. Robert F. Ray, Professor and Director, Institute of Public Affairs, SUI</td>
<td>April 15, 1958</td>
</tr>
<tr>
<td>&quot;Advocacy, Its Basis and Forms&quot;</td>
<td>Dr. Orville A. Hitchcock, Professor, Speech, SUI</td>
<td>April 22, 1958</td>
</tr>
<tr>
<td>&quot;Common Fallacies&quot;</td>
<td>Dr. William Eller, Associate Professor and Director of Reading Laboratory, Education, Reading Supervisor, Communication Skills, SUI</td>
<td>April 29, 1958</td>
</tr>
<tr>
<td>&quot;Communication Which Evaluates and Criticizes&quot;</td>
<td>Dr. John C. Gerber, Professor, English, SUI</td>
<td>May 13, 1958</td>
</tr>
</tbody>
</table>

Although the report of the experiment was concerned with reading, writing, speaking, and listening, only the writing will be treated in detail here.

The pretests and post-tests each included a 45-minute objective test in English composition and a 450-word minimum, two-hour expository theme. The themes were rated from 1 to 7 on each of these categories: purpose, content, organization, sentences, diction, and mechanics. (An analysis of each of these categories is given on the Theme Book cover, reproduced on page 40 of the original report, and samples of the instruction sheets given to the students are reproduced on pages 38-39.) The six ratings were then added to yield a total score. Pretest themes were rated by one experienced member of the Communication Skills staff. The 10:3 post-test themes were rated by three members; the 10:2 post-test themes were rated by two members with another reader added in cases of marked discrepancy. The final rating was the mean of these ratings. In each theme test, the student was given a choice of four topics—different on each testing occasion: "The Cold War," "Profit from Friendship," "My High School," etc., for the pretest; "The Significance of
an Incident in History," "Special Knowledge I Possess," "Family Relationships," etc., for the post-test. (Although it is not reported in the published study, the principal investigator has informed the writers of this report that the reliability of the rating was .83.)

In addition, several attitude tests were administered to students and instructors. At the beginning and conclusion of the experiment, a "Study of Educational Attitudes" test was administered to the students apart from their Communication Skills classes and dissociated from the course. This test included 66 multiple-choice items asking how important the student felt various aspects of the communication skills were, and, interspersed among them, 12 similar items about other college subjects. The student was also asked to rate himself on each of the skills of communication and to reveal certain matters about himself, his family, his interests, and his experiences.

Directly associated with the Communication Skills classes was a questionnaire administered near the close of the course, inquiring which of the three methods the students preferred, which they believed easiest, and which most effective. A similar questionnaire was filled out by the instructors.

Results

The investigators reported that before beginning the analyses of data they decided to use the 5 percent level of confidence as the level at which to reject the hypothesis of no difference.

Simple analyses of variance were made to test the homogeneity of the 10:1 and 10:3 classes on each of the pretests and the placement test Composite Score (indicating the percent of students the sum of whose percentile ranks on the English, Vocabulary, Reading Comprehension, and Mathematics tests falls below the student's sum—a score which had been "found to correlate highly with academic achievement in college"). The classes did not differ significantly on any of the pretest measures considered here.

A groups-within-treatments analysis of covariance design, with the individual as the unit of measurement, was planned to test the differences among methods of instruction. (This results in a more precise analysis when individual scores are used.) But before making that analysis, a test for differences among classes within each method was conducted to determine whether or not there were extreme or systematic differences among instructors.
Analysis of variance utilizing a treatments-by-level design (four levels of approximately equal numbers of students, established by the Composite Score) was used to test whether or not students of different levels of academic aptitude achieved more or less in theme writing under each method. A significant interaction being found between levels and methods in both the two-semester and the one-semester courses, differences among methods of instruction were tested for each level independently for each course, and differences among levels were tested for each method independently for each course.

Within the next to the top level in the two-semester course, the Kinescope group averaged significantly lower theme scores than the Normal or Bibliography groups, which did not differ significantly. No significant differences among methods were found at any other level in the two-semester course. Although significant differences among some levels were found for each of the methods of instruction, the differences between pairs of levels were somewhat unsystematic in the two-semester course. In no case did a group at one level average a significantly higher theme score than a group at a higher level.

In the one-semester accelerated course, no significant interaction was found between levels and methods for the theme post-test, but significant differences were found among all levels and methods of instruction. The Normal group again scored significantly higher than the Bibliography group, and the top-level group scored significantly higher than each of the other levels. The other levels did not differ significantly among themselves.

On the "Study of Educational Attitudes" test, administered near the end of the experiment, analyses of variance showed no significant differences among instructors (classes) within each method or among methods of instruction.

When the students in the two-semester course filled out the questionnaire on their attitudes toward the three methods of instruction, on all three questions the Normal method was obviously the most favored among both the Normal students and the Bibliography students, while the Kinescope students clearly favored the method which they had experienced. The Bibliography method was least favored by the Normal and Kinescope students, but the Kinescope method was least favored by the Bibliography students. In the one-semester accelerated course, both the Normal and Bibliography groups favored the Normal method, the Normal students reacting overwhelmingly in that fashion. Summing up the results of the attitude questionnaire, the investigators reported,

"Though students, in general, seem to prefer the Normal method of instruction, experience with either of the other methods tends to make them much more favorable to the method which they have experienced."

Near the end of each course, 44 instructors who had taught in the two-semester course and 8 instructors who had taught in the one-semester course filled out questionnaires indicating their attitudes toward the three methods. Although theoretical cell frequencies were not great enough to permit the calculation of a chi-square, it was very clear that these instructors preferred the Normal method of instruction and overwhelmingly rejected the Bibliography method. Their primary reason was that the Normal method provided them the most opportunity for discussion, and many noted that the other methods afforded too little time to adjust to the individual needs of students. A few liked the standardization and better coverage of lecture material through the kinescopes.

In conclusion, there seemed to be no consistently significant difference in the effectiveness of the three methods in the two-semester regular course. The Normal method may have been more effective than the Bibliography method for teaching theme writing to freshmen in the one-semester accelerated course, but the observed difference may well have been the result of chance. It is possible that the Kinescope method would have fared better had lecturers more experienced with television made the kinescopes and that the Bibliography method would have fared better had the bibliographies been more selective. It is also possible that both the experimental methods would have been more successful if the instructors had had more experience with them, but the attitudes of the instructors seem to work against the likelihood. Although the Bibliography method (classes meeting one less time per week) permits the saving of classroom space and of instructor time in class, it may be less effective as a method of teaching composition to college freshmen who are generally of above average academic ability.

If the Becker experiment had been preceded by a pilot experiment—permitting the development of more satisfactory kinescopes and bibliographies and perhaps a more supportive attitude on the part of the instructors, the Kinescope and Bibliography methods might have been shown significantly superior to the Normal method. The real conclusion of the experiment is that—despite the factors working against both experimental methods—the Normal method was not significantly superior. That in itself is significant, in a different sense, because it suggests that college freshmen, perhaps especially those of average ability, can learn to improve their writing as well when new concepts are presented by
film or by a bibliography of references as when new concepts are presented by individuals on a staff composed largely of part-time graduate assistants among whom there is a fairly high annual turnover.

V.

REFERENCES FOR FURTHER RESEARCH

Summaries and Bibliographies

A "classic" source to consult for research in the twenties and before is Bullo L. Lyman's *Summary of Investigations Relating to Grammar, Language, and Composition*, published in 1923 as a *Supplementary Educational Monograph* by the University of Chicago. Lyman gave summaries of more than 250 studies and offered many excellent suggestions for the improvement of research. Periodic summaries or bibliographies published in the intervening period include the following:


*English Journal*, Yearly reviews of research in secondary English, prepared by the NCTE Committee on Research. Journal.


Since 1950, a number of bibliographies and summaries have been published in addition to the many helpful bibliographies included with other books, articles, and unpublished manuscripts. Some titles are offered here:


REFERENCES FOR FURTHER RESEARCH


RESEARCH IN WRITTEN COMPOSITION


REFERENCES FOR FURTHER RESEARCH


RESEARCH IN WRITTEN COMPOSITION


