AD NUMBER: AD0203571

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

Distribution authorized to US Government agencies and their contractors; Administrative/Operational Use; Aug 1958. Other requests shall be referred to Office of Naval Research, Arlington, VA 22203.

AUTHORITY

ONR memo dtd 24 Aug 1970

A Study of Tests of Report Writing Ability

NORMAN FREDERIKSEN

Office of Naval Research Contract Nonr-2338(00)
Project Designation NR 151-182
Norman Frederiksen, Principal Investigator

Educational Testing Service
PRINCETON, NEW JERSEY

AUGUST 1958

A STUDY OF TESTS OF REPORT WRITING ABILITY

Abstract

A number of experimental tests designed to measure qualities important in report writing were administered to 506 students at a graduate school of business administration. Data used in the study included scores on the experimental tests, the Cooperative English Test, and the Admission Test for Graduate Study in Business, as well as first-year average grades and grades in a report writing course. For a subgroup of 162 students, evaluations of paragraphs and sentences from three reports written by each student were also obtained. Separate "set" and "content" scores were obtained from two of the experimental tests.

Reliabilities of the experimental tests ranged from .43 to .79. The reliability of the sentence-paragraph evaluations was estimated to be only .39.

The set scores, which are thought to measure criticalness, were found to have low negative correlations with measures of ability, while the content scores, particularly the content score of Recognizing Ambiguities, behave like conventional measures of verbal ability. Two other experimental tests, the Interlinear Test and the Selection and Organization Test, are reasonably reliable yet have low correlations with conventional measures of verbal abilities.

The best of the experimental tests for predicting writing ability are Recognizing Ambiguities (content score), Alternative Expressions (set score), and the Interlinear Test.

The Interlinear Test was found predominantly to reflect accuracy. A negative correlation was found between subscores based on items measuring accuracy and organization.

A STUDY OF TESTS OF REPORT WRITING ABILITY*

Introduction

A number of experimental tests intended to measure various aspects of report writing ability have been developed in the Research Division of ETS. If successful, the tests are intended for use eventually as aids in assigning personnel to jobs which require them to write technical reports in various fields. The purpose of the present study is to investigate the relationship of scores on the experimental tests to various measures of success in report writing obtained from students in a graduate school of business.

The Experimental Tests

Before the experimental tests were prepared, the author held a series of discussions with a number of writers and editors of technical reports and with supervisors of such report writers. These discussions resulted in the formulation of several hypotheses regarding characteristics of successful report writers. The experimental tests represent attempts to translate these hypotheses into measuring instruments. Four of the tests developed for this purpose were used in this study.

Recognizing Ambiguities. One characteristic of good report writing which was unanimously ranked high is clarity. Ability to write clearly presumably implies an ability to recognize ambiguous statements. The experimental test of this ability consists of a series of statements, and the task is to judge whether each statement is ambiguous or ~~un~~-ambiguous (ambiguous being defined as allowing more than one grammatically

*This study was supported by the Office of Naval Research, under Contract Nonr-2338(00), and by Educational Testing Service.

defensible interpretation). For example:

> The plot is pure farce, involving a phony British nobleman's quest for the hand of an American heiress, a social climbing American mother, and a visiting English lady named Mrs. Wollope.

Each of the 50 items in the test is to be answered by checking ambiguous or unambiguous. (The sample item above is keyed ambiguous. One interpretation is that the plot involves a quest, an American mother, and Mrs. Wollope; another is that the plot involves the nobleman's quest for an heiress, an American mother, and Mrs. Wollope.)

Alternative Expressions. Good report writers are supposed to be able to sense fine differences in shades of meaning and thus to avoid unintentionally changing meaning in reporting information. Each item in the Alternative Expressions test consists of a sentence in which a word or phrase is underlined, followed by another word or phrase in parentheses. The task is to judge whether the word in parentheses could safely be substituted for the underlined expression. The instructions encourage the examinee to be rigorous, as though a policy decision, administrative action, or even a legal claim might hinge upon the interpretation of the sentence. A sample item:

> He implied that he would resign.
> (It could be inferred.)

The sentence is marked same or different, depending upon whether the consequences are judged to be same or different if the alternative expression is adopted. There are 70 items in this test. (The sample

is keyed Different. "He implied" connotes that he intended to communicate his intention to resign, while "it could be inferred" does not.)

Each of the above tests is scored in such a way as to yield two scores: (1) a "content" score, which presumably measures the ability to make the type of judgment involved in the particular test; and (2) a "set" score, which is thought to measure how critical the candidate is with regard to what constitutes satisfactory writing, or how high his standards of precision in prose writing are. The scores are based upon a model described by Helmstadter [2], in which the examinees' responses are considered to result from a combination of knowledge, guessing, and set. We assume that the respondent answers each item on the basis of knowledge if he can; if not, his answer will result from guessing or from the operation of a response set. In our experimental tests, the candidate who has high standards of precision in the judgment of written statements would presumably have a set toward choosing the alternatives ambiguous or different.

The content score is defined as the proportion of items keyed in one way which are correctly answered, plus the proportion of items keyed the other way which are correctly answered. Content scores may thus vary between 0 and 2. The set score is the difference between the two proportions mentioned above, divided by 2 minus the content score. The set score can vary from -1 (for perfect set toward criticalness) to +1 (for perfect set in the other direction). (The derivations of these scoring formulas are discussed more fully in [1], which also describes relationships of these experimental tests to other measures.)

Selection and Organization Test. Other hypotheses about the characteristics of good writers have to do with ability to discriminate relevant from irrelevant material, to pick out salient facts, and to present the selected materials in a logical and readable order. The Selection and Organization Test is designed to measure such abilities. It has two parts. In Part 1, three pages containing ten dispatches more or less related to air power in the Far East are presented, together with instructions to plan a newspaper article, based on the dispatches, dealing with the relative strength of Communist and non-Communist air power in the Far East. Further instructions call for writing an outline of the article as planned, writing a headline for the article, and indicating what use, if any, is to be made of each dispatch in the article. The lead paragraph, it is made clear, is to present the most important conclusion. Part 1 of the test is not scored.

The items of Part 2 of the Selection and Organization Test are mainly statements of ideas contained in the dispatches. The candidate is instructed to choose the option which best describes the use he made of the idea in his story. Part 2 is thus a device which enables the candidate to report what ideas he spontaneously chose to use and how he planned to use them.

One item, for example, is as follows:

> General Howe said the first objective of his command in the event of war would be to gain control of the air.

The multiple-choice options are the same for all such items, viz.:

a. Used as the main idea in the lead paragraph.

b. Used in the lead paragraph to support the main idea.

c. Used as the main idea of a paragraph other than the lead paragraph.

d. Used to support the main idea of a paragraph other than the lead paragraph.

e. Omitted because the passage is not relevant to the assignment.

f. Omitted because the passage may not be true.

g. Omitted for some other reason.

Since there are many satisfactory ways of selecting and organizing ideas for the assignment, the scoring key ordinarily recognizes more than one of the options as correct.

Interlinear Test. This test represents another attempt to get at some of the hypotheses already described, and in addition it presumably measures a quality which might be called accuracy in reporting. It consists of a badly written story based on the same dispatches and assignment described above. The task is to make any corrections that are deemed necessary, the only restrictions being that new ideas are not to be added and ideas already included are not to be removed. Brief instructions on how to use ordinary proofreaders' signs are included.

In scoring the Interlinear Test a guide is used which is based on 38 points in the story where corrections are appropriate. These points cover a wide variety of corrections, from inaccuracies in reporting ("top commander" instead of "top Air Force commander") to poor organization. For example, two paragraphs of the passage lack unity. Credit is given for moving the first sentence of one of the paragraphs to the

beginning of the second, breaking this second paragraph into two paragraphs, and combining the second part with the remaining part of the first paragraph. None of the scoring points are concerned merely with grammar, punctuation, or spelling; all have to do with the accuracy and clarity of the reporting.

Studies of the reliability of scoring the test show that correlations between scorers working independently are high. Correlations of .96 and .97 have been found in two separate studies of scorer agreement. The first correlation was based on 62 cases. One of the scorers developed the scoring instructions and later scored the 62 papers. The second scorer had no training in scoring other than to read the scoring instructions. The correlation of .97 was based on 100 cases from the present study. The two scorers worked from the scoring instructions; they scored a set of sample papers and discussed their disagreements before independently scoring the 100 papers on which the correlation of .97 is based.

Other Tests Used as Predictor Variables

For purposes of comparison with the experimental tests and for use as reference tests to aid in interpreting the findings, scores from two other tests were used in the study.

Cooperative English Test. The Cooperative English Test (Higher Level) yields three scores -- Mechanics of Expression, Effectiveness of Expression, and Reading Comprehension -- as well as a total score.

Admission Test for Graduate Study in Business. A composite score on the Admission Test for Graduate Study in Business, which had been taken by all the students in connection with their application for admission, was used as another reference test.

Criterion Variables

First-year students at the graduate school of business administration take a course in report writing; evaluations of work done in this course provided criteria of writing ability. During the year a number of reports, based on careful analyses of problem situations, are written by each student. In the written analysis the student is expected to put together pieces of information about a case, make deductions, appraise the wisdom of various courses of action, and set forth a persuasive defense of the action he proposes. The grading of these reports is done by a special staff of graders, some of whom are employed full-time in that capacity while others are doctoral candidates who work part-time. Each grader marks the papers of a particular group of students and prepares written comments. Normally a reader interviews the student once during the first term; interviews are optional during the second term. The graders hold meetings to discuss each assignment, at which time they independently grade a few sample papers and discuss the results in order to reach agreement on grading standards. The quality of the reasoning and the solution proposed in the paper are given more weight in grading than the quality of the writing.

While the grading practices used have many advantages from the standpoint of teaching, they leave much to be desired from the point of view of a validation study. Therefore a special evaluation procedure was set up for purposes of this study. The procedure emphasized writing skills rather than the reasoning processes exhibited in the paper. Three of the case reports were used in this evaluation. The grader selected three paragraphs from each report for careful analysis. These paragraphs were (1) the second paragraph, (2) the first paragraph

beginning on page three, and (3) the last paragraph of the report. Each paragraph was evaluated on four defined characteristics: unity, emphasis, development, and coherence. Also, each sentence in each of the three paragraphs was evaluated on the basis of criteria presented in the instructions. These criteria include such characteristics as connection with preceding sentence, grammatical correctness, clarity, emphasis, freedom from redundancy, etc. One judgment was made about each sentence. All judgments were expressed on a scale of 1 to 9. One grader evaluated the reports dealing with one report, a second grader the second report, and a third grader the third report; since no two graders evaluated the same report, a study of grader reliability could not be made. The three graders were not those employed on the regular grading staff. The instructions to raters are reproduced in the Appendix.

For each student, the four judgments made about each paragraph were added to obtain paragraph score, and the judgments about the sentences in each paragraph were averaged to make a sentence score. From these paragraph scores and sentence scores the following twelve variables were obtained for use in the analysis:

Sum of paragraph and sentence scores, all judges
Sum of paragraph ans sentence scores, Judge 1 (Report 1)
Sum of paragraph and sentence scores, Judge 2 (Report 2)
Sum of paragraph and sentence scores, Judge 3 (Report 3)
Sum of paragraph scores, all judges
Sum of paragraph scores, Judge 1 (Report 1)
Sum of paragraph scores, Judge 2 (Report 2)
Sum of paragraph scores, Judge 3 (Report 3)
Sum of sentence scores, all judges
Sum of sentence scores, Judge 1 (Report 1)
Sum of sentence scores, Judge 2 (Report 2)
Sum of sentence scores, Judge 3 (Report 3)

In addition, three other criterion measures were obtained from the school. One is the first-year average grade, another is the final grade in the report-writing course, and the third is a dichotomous measure, warned-not-warned that the student has a writing deficiency.

Other Variables

The instructions for evaluating paragraphs in the reports contain phrases which suggest the possibility that length of paragraphs might be related to the judgments. For example, the question is asked, "Is the paragraph clearly unified about a single, central subject or thought?" Logically it would seem that unity is easier to achieve in a few sentences than in a large number of sentences. Similarly, with respect to the question, "Does it (the paragraph) make a clearly observable central point about its subject, and does it make that point emphatically?" it would seem that emphasis is more easily attained in few sentences than in many. Similar arguments can be applied to judgments of development and coherence.

In order to test the hypothesis that paragraph ratings are inversely related to paragraph length, the following four variables were included:

Total number of sentences, all reports
Total number of sentences, Report 1 (Judge 1)
Total number of sentences, Report 2 (Judge 2)
Total number of sentences, Report 3 (Judge 3)

The Subjects

The class entering the graduate school of business in the fall of 1956 provided the subjects of the study. All were given the experimental tests and the Cooperative English Test at the time of entrance in the

fall. The Admission Test for Graduate Study in Business had been taken earlier, at the time the subjects were candidates for admission. Scores on this test were used, along with other information, in admitting students; therefore some curtailment in the distribution of scores may be expected.

The special evaluations of reports were obtained for a sub-sample of the subjects, since considerations of time and budget did not permit making these evaluations for the entire class. The judges were selected and trained for this work by the staff of the report writing course. Since this aspect of the grading got under way late in the school year, and since the special evaluations could not be permitted to interfere with the usual routines of grading and reporting, it was necessary to call back from students previously written reports. Student participation in this part of the study was largely voluntary; therefore it seemed likely that papers receiving low scores in the initial grading would be underrepresented.

A total of about 610 students were tested, but about 104 were dropped from the study because they omitted too many items from the Ambiguities and Expressions tests for the scoring formulas to be applicable. That left 506 students, of whom 162 returned reports for evaluation while 344 did not. Complete data on all 30 variables were therefore available for 162 students, and data on 14 variables, excluding the special evaluations, for 506.

Analysis

The primary purpose of the study was to investigate the relationships of the experimental tests to the measures of ability to write

which were obtained at the school. In order to make these relationships more meaningful, however, it is also necessary to investigate the reliability of the tests and their correlations with other ability measures, and to examine the interrelationships and reliability of the criterion measures. We also need to find out the effects on sampling of allowing students to be self-selected for membership in the subsample whose reports were given special evaluation.

Accordingly, intercorrelations of all 30 variables were computed for the 162 cases with complete data, and intercorrelations of 14 variables (all except those resulting from the special evaluation) were computed for 506 cases. Means and standard deviations were obtained for both of these groups and also for the 344 cases who had not returned reports for evaluation. Significance tests of differences between means for the 162 who returned reports and the 344 who did not were made. Reliability estimates for the experimental tests and for the criterion ratings were determined.

Comparison of Students Who Did and Who Did Not Return Reports. The 162 students who returned reports for use in the criterion study were compared with the 344 who did not, with respect to the 14 variables. One would suppose that the students who chose to return their reports for use in the validity study would tend to be more able, on the average, than those who did not. It turned out that the advantage favored those who returned reports in only 9 of the 14 comparisons, and the differences were very small. Only one of the differences was significant at the 5% level; the means for first year average grade were 75.2 and 74.5 for those returning and those not returning reports, respectively. The

method of obtaining cases for use in the careful evaluation of reports did not produce any important bias in ability level of students as measured by the 14 variables available.

Reliability of Experimental Tests. The reliability of the experimental tests was estimated by obtaining scores for odd- and even-numbered items and applying the Spearman-Brown prophecy formula. The results were as follows:

Recognizing Ambiguities (Content)	.48
Alternative Expressions (Content)	.43
Recognizing Ambiguities (Set)	.54
Alternative Expressions (Set)	.73
Interlinear Test	.79
Selection and Organization Test	.68

Although these reliabilities are not high enough to justify use of the tests in individual selection, they do suggest that satisfactory reliabilities could be attained by application of suitable techniques of test refinement. None of the experimental tests had been item-analyzed prior to their use in this study.

Intercorrelations of the Predictor Measures. In Table 1 the intercorrelations, means, and standard deviations of the experimental tests and other ability tests are shown. The first entry in each cell of the table is for the group of 506 students, and the second entry is for the subgroup of 162 students who returned reports for use in the special scoring.

Comparison of the pairs of correlations shows that they are very similar. The biggest difference is .11, and 47 of the 55 differences are smaller than .08. There is, however, a consistent tendency for the

(N = 506 for Group; N = 162 for Subgroup)

		1 Amb. C.	2 Alt. Ex. C.	3 Amb. S.	4 Alt. Ex. S.	5 Interl.	6 S. and O.	7 Coop: Mech.	8 Coop: Eff.	9 Coop: Read.	10 Coop: Total	11 Admis. Test
1. Recognizing Ambiguities: Content	Group		.15	.06	-.05	.14	.09	.37	.26	.35	.40	.27
	Subgroup		.23	.00	-.02	.25	.20	.40	.27	.40	.44	.37
2. Alternative Expressions: Content	Group	.15		-.01	-.11	.05	.13	.17	.07	.19	.18	.18
	Subgroup	.23		-.06	-.09	.05	.13	.13	.02	.22	.15	.13
3. Recognizing Ambiguities: Set	Group	.06	-.01		.22	.01	.04	-.09	-.08	-.07	-.09	.01
	Subgroup	.00	-.06		.17	.07	.07	-.10	-.08	-.11	-.11	.02
4. Alternative Expressions: Set	Group	-.05	-.11	.22		-.13	-.02	-.15	-.06	-.04	-.09	-.01
	Subgroup	-.02	-.09	.17		-.16	-.04	-.22	-.08	.02	-.11	-.02
5. Interlinear Test	Group	.14	.05	.01	-.13		.18	.18	.21	.14	.22	.17
	Subgroup	.25	.05	.07	-.16		.20	.28	.26	.21	.31	.22
6. Selection and Organization Test	Group	.09	.13	.04	-.02	.18		.16	.19	.19	.22	.24
	Subgroup	.20	.13	.07	-.04	.20		.18	.16	.28	.25	.34
7. Cooperative English Test: Mechanics	Group	.37	.17	-.09	-.15	.18	.16		.47	.50	.79	.45
	Subgroup	.40	.13	-.10	-.22	.28	.18		.44	.56	.81	.51
8. Cooperative English Test: Effectiveness	Group	.26	.07	-.08	-.06	.21	.19	.47		.42	.78	.39
	Subgroup	.27	.02	-.08	-.08	.26	.16	.44		.45	.79	.41
9. Cooperative English Test: Reading Comprehension	Group	.35	.19	-.07	-.04	.14	.19	.50	.42		.81	.64
	Subgroup	.40	.22	-.11	.02	.21	.28	.56	.45		.84	.65
10. Cooperative English Test: Total	Group	.40	.18	-.09	-.09	.22	.22	.79	.78	.81		.62
	Subgroup	.44	.15	-.11	-.11	.31	.25	.81	.79	.84		.64
11. Admission Test for Graduate Study in Business	Group	.27	.18	.01	-.01	.17	.24	.45	.39	.64	.62	
	Subgroup	.37	.13	.02	-.02	.22	.34	.51	.41	.65	.64	
Mean of Group		1.20	1.31	-.05	-.12	12.5	13.6	64.2	74.1	77.5	73.2	587
Mean of Subgroup		1.21	1.30	-.04	-.15	12.7	13.8	63.9	74.5	77.4	73.2	592
Standard Deviation of Group		.18	.15	.26	.32	4.9	4.3	8.8	9.6	10.0	8.2	74
Standard Deviation of Subgroup		.17	.14	.28	.31	5.1	4.1	8.6	10.0	10.2	8.4	77

correlations from the subgroup to be larger; in fact 41 of the 55 correlations are larger for the subgroup. (This proportion of differences in one direction would be expected to occur by chance less than once in a hundred times.) The tendency for subgroup correlations to be larger cannot be accounted for by greater variability, as can be seen from the standard deviations at the bottom of the table. No other explanation for the generally higher correlations in the subgroup can be offered.

The two content scores correlate positively with each other, as do the two set scores, but the correlations between set and content scores tend to be negative. When the correlation of .22 between the two set scores is corrected for unreliability of both measures, it becomes .35; apparently the two scores are only in part measuring the same quality. The set scores quite consistently have low negative correlations with the other measures of ability. Negative set scores are supposed to indicate greater criticalness; therefore these negative correlations suggest that there is a slight tendency for the more able people to be more critical or to have higher standards of good writing.

If the reliabilities of the experimental tests are taken into account in interpreting their correlations with the Cooperative English Test and the business school test, it would appear that the content score of Recognizing Ambiguities is most like the conventional measures of verbal ability. On the other hand, the Interlinear Test and the Selection and Organization Test do not overlap the conventional tests greatly. As might be expected, the set scores least resemble the conventional tests of ability in the study.

Intercorrelations of Judgments. It will be recalled that 162 students returned the reports they had written on three particular assignments for use in the special evaluation. The regular grading procedures, having been set up in part as teaching techniques, were unsatisfactory from the standpoint of producing criteria for evaluating the experimental tests. The careful evaluation of the reports which were written by the 162 students would, we believed, provide more reliable measures of ability to write. The instructions to the graders were written with a view to limiting the evaluations to characteristics of writing which are relevant to the experimental tests. Table 2 shows the intercorrelations of the measures which resulted from the special grading.

Correlations between judges are of most interest. It will be remembered that Judge A read the reports from one assignment, Judge B those from a second assignment, and Judge C those from a third. The correlations therefore reflect both the unreliability of the judging process itself and the unreliability in the performance of the writers. The intercorrelations show that the attempt to develop reliable measures failed; the intercorrelations of the composite of all paragraph and sentence judgments were .17, .18, and .18. If we consider the average of these intercorrelations to represent the reliability of a measure one-third the total length of the measure, we find by use of the Spearman-Brown formula that the reliability of the composite for all three judges is only .39.

The correlation between the total paragraph score and the total sentence score is .76. This figure cannot be used for estimating reliability, however, because of the experimental dependence of paragraph

(N = 162)

	1 P&S, All	2 P&S, A	3 P&S, B	4 P&S, C	5 P., All	6 P., A	7 P., B	8 P., C	9 S., All	10 S., A	11 S., B	12 S., C	13 N., All	14 N., A	15 N., B	16 N., C
1. Paragraphs and Sentences, All Judges		.75	.67	.56	.99	.75	.65	.53	.84	.67	.62	.38	-.06	-.10	.11	-.12
2. Paragraphs and Sentences, Judge A	.75		.17	.18	.73	.99	.15	.15	.69	.90	.22	.18	-.08	-.10	.00	-.08
3. Paragraphs and Sentences, Judge B	.67	.17		.18	.67	.18	.99	.17	.52	.12	.84	.14	-.05	-.08	.12	-.12
4. Paragraphs and Sentences, Judge C	.56	.18	.18		.58	.16	.17	.97	.45	.19	.19	.56	.03	-.01	.14	-.03
5. Paragraphs, All Judges	.99	.73	.67	.58		.73	.66	.57	.76	.63	.58	.30	-.05	-.09	.12	-.11
6. Paragraphs, Judge A	.75	.99	.18	.16	.73		.16	.14	.66	.85	.23	.17	-.08	-.10	.01	-.08
7. Paragraphs, Judge B	.65	.15	.99	.17	.66	.16		.16	.47	.11	.76	.12	-.03	-.06	.12	-.10
8. Paragraphs, Judge C	.53	.15	.17	.97	.57	.14	.16		.32	.16	.15	.35	.04	.01	.13	-.04
9. Sentences, All Judges	.84	.69	.52	.45	.76	.66	.47	.32		.71	.64	.67	-.09	-.14	.06	-.10
10. Sentences, Judge A	.67	.90	.12	.19	.63	.85	.11	.16	.71		.14	.19	-.08	-.10	-.03	-.05
11. Sentences, Judge B	.62	.22	.84	.19	.58	.23	.76	.15	.64	.14		.21	-.12	-.16	.07	-.18
12. Sentences, Judge C	.38	.18	.14	.56	.30	.17	.12	.35	.67	.19	.21		.03	-.03	.09	.02
13. Number of Sentences, All Reports	-.06	-.08	-.05	.03	-.05	-.08	-.03	.04	-.09	-.08	-.12	.03		.79	.76	.81
14. Number of Sentences, Report A	-.10	-.10	-.08	-.01	-.09	-.10	-.06	.01	-.14	-.10	-.16	-.03	.79		.42	.41
15. Number of Sentences, Report B	.11	.00	.12	.14	.12	.01	.12	.13	.06	-.03	.07	.09	.76	.42		.44
16. Number of Sentences, Report C	-.12	-.08	-.12	-.03	-.11	-.08	-.10	-.04	-.10	-.05	-.18	.02	.81	.41	.44	
Mean	228.8	61.4	84.8	82.0	184.6	50.0	68.4	66.2	44.7	11.8	16.8	16.2	34.5	11.9	11.2	11.4
Standard Deviation	24.6	14.7	12.1	9.3	20.5	12.2	10.3	8.3	5.0	2.8	2.3	2.3	12.0	5.3	4.4	5.6

Intercorrelations of Criterion Variables. Table 3 shows the intercorrelations of various criterion data, including the composite of the sentence and paragraph judgments. This measure appears to have little in common with any of the measures obtained from the school records, in part because of its low reliability.

The correlation of .56 between grades in the report-writing course and first-year averages is of course spuriously high because it is in part a self-correlation. There is only a moderate correlation between the course grade and warned-not-warned.

Correlation of Experimental Tests and Other Ability Tests with Criterion Variables. Table 4 presents the correlations between predictors and criteria. Coefficients are given for both the group and the subgroup, except that the composite of the paragraph and sentence judgments was available only for the subgroup.

We will consider first the criterion shown in the first column, final grade in the report-writing course. We do not know the reliability of this measure, but we are probably safe in assuming that it is not high; we should therefore not expect to find high correlations with it. The highest correlations with this final grade are those for the Cooperative English Test and in particular its subtest, Mechanics of Expression. Since this subtest contains items measuring grammatical usage, punctuation, capitalization, and spelling, it may be inferred that the graders were sensitive to these aspects of writing. Such emphasis could come about because mechanics comprise relatively objective characteristics of writing which graders are likely to see in common. The highest correlation in the column is .41, the correlation between Mechanics of Expression and final grade in report writing for the subgroup.

TABLE 3

Intercorrelations of Criterion Variables

(N = 506 for Group; N = 162 for Subgroup)

		1 Final Grade	2 Warned Vs. Not Warned	3 First-Year Average	4 P&S, All
1. Final Grade, Report Writing	Group		.42	.56	
	Subgroup		.41	.56	.23
2. Warned vs. Not Warned	Group	.42		.24	
	Subgroup	.41		.24	.21
3. First-Year Average Grade	Group	.56	.24		
	Subgroup	.56	.24		.22
4. Paragraphs and Sentences, All Judges	Group				
	Subgroup	.23	.21	.22	
Mean of Group		74.8	1.91	74.7	
Mean of Subgroup		75.2	1.91	75.2	228.8
Standard Deviation of Group		4.7	.29	2.9	
Standard Deviation of Subgroup		4.2	.28	2.9	24.6

TABLE 4

Correlations of Predictor Variables with Criterion Variables
(N = 506 for Group; N = 162 for Subgroup)

		1 Final Grade	2 Warned Vs. Not Warned	3 First-Year Average	4 P&S, All
1. Recognizing Ambiguities: Content	Group	.15	.10	.14	
	Subgroup	.19	.15	.19	.30
2. Alternative Expressions: Content	Group	.02	.02	.04	
	Subgroup	.03	.01	.08	-.01
3. Recognizing Ambiguities: Set	Group	-.02	-.04	-.07	
	Subgroup	.03	-.00	-.04	.09
4. Alternative Expressions: Set	Group	-.21	-.02	-.13	
	Subgroup	-.30	.05	-.17	-.15
5. Interlinear Test	Group	.14	.07	.14	
	Subgroup	.25	.11	.21	.19
6. Selection and Organization Test	Group	.11	.08	.18	
	Subgroup	-.02	.03	.10	.08
7. Cooperative English Test: Mechanics	Group	.28	.15	.15	
	Subgroup	.41	.24	.25	.27
8. Cooperative English Test: Effectiveness	Group	.25	.10	.23	
	Subgroup	.18	.10	.23	.22
9. Cooperative English Test: Reading Comprehension	Group	.18	.12	.27	
	Subgroup	.14	.15	.20	.21
10. Cooperative English Test: Total	Group	.28	.13	.27	
	Subgroup	.29	.19	.28	.29
11. Admission Test for Graduate Study in Business	Group	.19	.15	.36	
	Subgroup	.22	.22	.33	.20

Among the experimental tests, the content score of Recognizing Ambiguities, the set score of Alternative Expressions, and score on the Interlinear Test appear to be potentially useful. If we correct for the unreliability of the experimental tests, the validity coefficients for the group and subgroup, respectively, are as follows: Ambiguities content, .22 and .28; Alternative Expressions set, -.25 and -.35; Interlinear, .16 and .28. The results for the set score are particularly interesting, since set scores tend to measure qualities that conventional aptitude tests do not measure. (The negative coefficients are in the expected direction: high criticalness tends to be associated with high grades.)

The dichotomous criterion warned-not-warned (of a deficiency in writing ability) has very low correlations with all the predictors. One would guess that the reliability of this measure is quite low. It is interesting to note that the best predictor of this criterion appears to be Mechanics of Expression, which again suggests that inability to spell, capitalize, and punctuate are likely to be identified in common by judges of written material.

First-year average grade is a criterion which of course is much broader than mere writing ability. The best predictor is the one intended to predict such a criterion, the Admission Test for Graduate Study in Business. Among the experimental tests, the highest correlations are those for Recognizing Ambiguities (content score) and the Interlinear Test. Corrected for unreliability of the predictor the correlations are .20 and .28 for Ambiguities and .16 for .24 for Interlinear.

The correlations for the composite of paragraph and sentence judgments are considerably attenuated by unreliability. Even so, the Cooperative English Test correlates .29 with this criterion, and the Mechanics of Expression part correlates .27. The experimental tests, being less reliable than the Cooperative English Test, yield lower correlations with the composite criterion. If the correlations are corrected for unreliability of the predictor, we find correlations of .30 for Recognizing Ambiguities (content score), .19 for the Interlinear Test, and -.15 for Alternative Expressions (set score).

In general, the set score for Recognizing Ambiguities and the content score for Alternative Expressions were found to be unrelated to any criterion and to have low correlations with other ability measures. Furthermore, their reliabilities were among the lowest of the experimental tests. It is suggested that the level of item difficulty may account for the poor performance of these measures. The rationale of the set score is that if an examinee cannot answer an item through knowledge, he will answer it by guessing or through the operation of a set or bias. Therefore, if the items in a test can be answered through knowledge, set will have relatively little influence; but if the items are too hard to answer through knowledge the candidate will have to resort to guessing or to the influence of a set. This hypothesis about level of item difficulty in relation to quality of set and content scores could be tested through the use of item analysis procedures.

Part Scores from the Interlinear Test

After the findings described so far became known, the problem arose as to precisely what kinds of skills are measured by the Interlinear Test. Of the 38 scoring points (items), it was found that 25 could be classified as items concerned with the accuracy of reporting information, six were concerned with organization (rearrangement of sentences, etc.), and the rest seemed to measure various other things. The papers of all 610 students who took the test were rescored to yield an accuracy score and an organization score. The split-half reliabilities of these new measures were found to be .83 for accuracy and .36 for organization. The intercorrelations shown in Table 5 were obtained.

The results show that the Interlinear Test in its present form should be interpreted primarily as a measure of the accuracy with which an editing task is performed. The reliability of the 25-item accuracy score is higher than the reliability of the 38-item total score. An interesting result is the significant negative correlation between accuracy and organization scores. While it would be desirable to do another study, in which an attempt is made to verify the negative correlation from samples of behavior which are experimentally independent, the finding suggests that our difficulties in predicting writing ability may stem from criteria which are too inclusive. We need to know whether writing ability is a unitary characteristic or a composite of relatively independent abilities.

Summary and Conclusions

A number of experimental tests designed to measure qualities important in report writing were administered to 610 entering students

TABLE 5

Intercorrelations of Interlinear Test Scores
(N = 610)

	Accuracy	Organization	Total Interlinear	Cooperative English total
Accuracy		-.13	.93	.18
Organization	-.13		.17	.15
Total Interlinear	.93	.17		.24
Cooperative English total	.18	.15	.24	
Mean	8.29	2.35	12.39	72.29
Standard Deviation	4.4	1.5	4.9	8.8

at a graduate school of business administration. Data used in the study included scores on the experimental tests, the Cooperative English Test, and the Admission Test for Graduate Study in Business; grades in a report-writing course and first-year average grades; and (for a subgroup of 162 students) composite evaluations, made by three judges, of paragraphs and sentences from reports written by the students.

Two of the experimental tests, Recognizing Ambiguities and Alternative Expressions, require judgments to be made about short samples of writing. Each test yields a content score, which is supposed to measure ability to make the discriminations required, and a set score, which is supposed to measure criticalness with respect to writing.

Another experimental test, the Selection and Organization Test, provides a measure of how appropriately the candidate selects material for an assigned topic and how well he organizes it. The last experimental test, called the Interlinear Test, requires the candidate to revise a poorly written story based on the same material and assignment used in the Selection and Organization Test.

Complete data were available for 162 students; all but the special evaluations of reports were available for 506. The 162 students did not differ appreciably in mean scores from those who did not volunteer to provide reports for the special evaluation.

Reliabilities of the experimental test scores ranged from .43 to .79. The special evaluations of reports, which it was hoped would provide a more reliable criterion than is usually available for writing skills, were disappointing; the reliability of the composite of the paragraph and sentence judgments was estimated to be only .39.

Intercorrelations of the predictors show that the two set scores tend to have low negative correlations with other measures of ability. (The negative relationship means that "criticalness" tends to be associated with high ability.) The content score of Recognizing Ambiguities is the experimental test score most like conventional measures of verbal ability. The Interlinear Test and Selection and Organization Test are reasonably reliable, yet have low correlations with the conventional measures.

Grades in the report-writing course were best predicted by the Cooperative English Test, particularly the subtest Mechanics of Expression (for which the r's are .28 and .41 for the entire group and the subgroup respectively). The best of the experimental tests for predicting this criterion are Recognizing Ambiguities (content score), Alternative Expressions (set score), and the Interlinear Test. Corrected for unreliability of the tests, the r's are .22 and .28 for the content score of Ambiguities, -.25 and -.35 for the set score of Expressions, and .16 and .28 for the Interlinear Test. These correlations presumably are still considerably attenuated by the unreliability of grades.

All correlations involving the dichotomy warned-not-warned (of a writing deficiency) are very low, probably because of unreliability of this criterion.

First-year average grades are best predicted by the Admission Test for Graduate Study in Business (r's are .36 and .33). Of the experimental tests, the best predictors are Recognizing Ambiguities (content score) and the Interlinear Test. The r's , corrected for the unreliability of the test, are .20 and .28 for Ambiguities content and .16 and .24 for the Interlinear.

In spite of the estimated low reliability of the composite of paragraph and sentence judgments (.39), the Cooperative English Test correlated .29 with this criterion. If the correlations of the experimental tests are corrected for attenuation due to unreliability of the predictor, the best test is Recognizing Ambiguities (content score) with an r of .30. Corrected correlations for the Interlinear and for Alternative Expressions (set score) are .19 and -.15 respectively.

The Interlinear Test was rescored to yield part scores for items reflecting accuracy and for organization. The accuracy score is highly reliable (.83) and correlates .93 with total Interlinear score (since most items were of this type). The correlations involving the Interlinear Test should therefore be interpreted as measures of accuracy in editing a report. The correlation between accuracy and organization scores was -.13, which suggests the possibility that writing ability may be a composite of relatively independent abilities rather than a unitary trait.

REFERENCES

1. Frederiksen, Norman, and Messick, Samuel. Response Set as a Measure of Personality. ONR Technical Report. Princeton, N. J.: Educational Testing Service, 1958.

2. Helmstadter, G. C. Procedures for obtaining separate set and content components of a test score. Psychometrika, 1957, 22, 381-393.

APPENDIX

Directions Regarding Evaluation of Analyses

1. Each reader will be given all of the papers on one of the cases (students returned to us papers they had written on three cases, November-February), and will thus be reading and rating only one of the student's three papers that we are using. This procedure helps insure objectivity in the separate ratings.

2. Each reader will receive a copy of the case on which the papers he is reading were written. He will thus be able to familiarize himself in some measure with the facts and problems being treated.

3. It will not be necessary to read all of any of the papers. Instead, attention should be given to

   the second paragraph of the paper

   the first paragraph that begins on page three (if none does, use from the beginning the paragraph that covers page 3, wherever that paragraph begins)

   the final paragraph of the paper

4. On the rating scale described below, make four separate judgments on each of the three paragraphs. These judgments will be as follows:

   a) Is the paragraph clearly unified about a single, central subject or thought?

   b) Does it make a clearly observable central point about its subject, and does it make that point emphatically? so that there can be no doubt of the argument being offered?

   c) Does it follow a logical sequence of thought? Does it develop its material clearly about the subject so as to reach its conclusion convincingly and forcefully? (One subquestion under this point: does it seem reasonably and satisfyingly complete, or does it seem inadequately developed?)

   d) Are the parts of the paragraph coherent? Do the sentences or ideas stick together (through use of a common point of view, the use of transitional words, effective repetition of key words of preceding sentences in succeeding sentences), instead of seeming to be points existing in a vacuum?

5. Also on the rating scale described below, make one judgment on the effectiveness of each sentence in each of the three paragraphs. Points to look for in determining the effectiveness of a given sentence include the following (all or any one or combination of them, as appropriate):

a) Is the sentence clearly connected to the thought of the preceding sentence?

b) Is the sentence clearly correct: i.e., does it have a subject, verb, object or predicate; do the parts (subject, verb) agree as to person and number; do pronouns have clear antecedents? are the verb tenses consistent within the sentence? etc.

c) Is it completely clear? no ambiguities reasonably possible?

d) Are modifiers properly used and placed? Check the following:

no dangling participles (participial phrases without subjects close to them);

no loosely placed adverbial modifiers, i.e., modifiers confusingly placed apart from the words and phrases they govern?

e) Is there a proper indication, according to what seems the logic of the sentence, or according to the relation that exists between the ideas, of the difference between major and subordinate ideas in the sentence? Are conjunctions and relative pronouns, etc., used to bring out which are the major and which are the subordinate elements or ideas in the sentence.

f) Is the sentence emphatic? Does it make clear, and forcible, what is its major point? It could do so by careful anticipating of the most important topic, placing the important idea last, selecting particularly forceful words, etc. The method used is by itself less important than that the meaning be made clearly emphatic.

g) The sentence should not be overloaded with too many subordinate or minor considerations or ideas. The parts should further be carefully disposed, and clearly integrated (for example by use of parallel structure) so that the relationships being established will be perfectly clear.

h) The sentence should contain no useless or unnecessary words, redundant phrases, etc. Words and phrases should be very precise, sharply pointed in meaning, to eliminate confusion.

i) As far as possible the sentences should avoid excessive use of the passive voice, especially avoiding it where there is no effort to emphasize that the subject is being acted upon.

6. Papers should be rated as follows:

Use a scale of 1 through 9 --

9 outstandingly effective, clear and exceptionally forceful or apt in statement

8 distinctly above average on the stated criterion (or criteria) and especially forceful

7 competent and correct, though not outstanding, on the stated criteria

6 through 1 showing deficiencies on one or more of the criteria. The number of weaknesses, and their seriousness, will determine how low the individual paragraph or sentence will be weighted.

Procedures in Conducting the Evaluation

(Please keep papers in numerical order by student, and rate them in that order.)

1. Record the student's number (from the upper right-hand corner of the paper) in the appropriate space on the paragraph rating form and on each form for sentence rating.

2. Read the second paragraph of the paper.

3. Make an individual judgment, on the rating scale of 1 - 9, of each of the:

Unity

Emphasis

Development

Coherence

of the paragraph, and record the rating assigned on each point in the appropriate box on the "Rating Sheet, Paragraphs" under the inclusive heading "Second Paragraph."

4. On the scale 1 - 9, and using the criteria listed on the general "Directions," make a judgment of each sentence in the paragraph. Record these judgments consecutively on the rating sheet headed "Rating Sheet, Sentences, Second Paragraph." Record the score for the first sentence under "1," the second sentence under "2," and so on from left to right across the sheet.

5. Read the first paragraph on page three of the paper, or the paragraph that covers page three from the beginning of that paragraph, wherever it may be.

6. Record the four separate judgments you make, as described under #3, in the appropriate box on the "Rating Sheet, Paragraphs" under the inclusive heading "First Paragraph on Page 3."

7. Make a judgment of each sentence in the paragraph, recording these judgments for each sentence consecutively, on the sheet headed "Rating Sheet, Sentences, First Paragraph on Page 3."

8. Read the last paragraph of the paper.

9. Record judgments on that paragraph, using the four categories mentioned in #3, on the "Rating Sheet, Paragraphs," under the inclusive heading "Final Paragraph."

10. Make a judgment on each sentence in the paragraph, recording the judgments consecutively according to the number of the sentence, on the sheet headed "Rating Sheet, Sentences, Final Paragraph."

P L E A S E MAKE ENTRIES L E G I B L Y ! ! !