REPORT RESUMES

ED 016 646 24 TE 000 137

VALIDITY OF THE REARRANGEMENT EXERCISE AS A PREDICTOR OF ESSAY WRITING ABILITY.

BY- CONRY, JULIANNE JOYCE
WISCONSIN UNIV., MADISON

REPORT NUMBER BR-5-0216-TR-23 PUB DATE MAY 67
CONTRACT OEC-5-10-154
EDRS PRICE MF-$0.25 HC-$1.00 23P.

DESCRIPTORS- *ENGLISH, *ESSAYS, *PARAGRAPH COMPOSITION, *PREDICTIVE ABILITY (TESTING), *SCORING, EDUCATIONAL RESEARCH, ENGLISH, RESEARCH METHODOLOGY, RESEARCH TOOLS, TEST CONSTRUCTION, TESTING PROBLEMS, TEST RESULTS, TEST VALIDITY, WRITING

DATA FROM THE PARAGRAPH ORGANIZATION PORTION OF THE CEEB ENGLISH COMPOSITION TEST (ECT) WERE CONVERTED TO THE ORIGINAL RANK-ORDER AND WERE THEN RESCORED BY THREE SYSTEMS USING SPEARMAN'S RHO TO DETERMINE WHICH METHOD YIELDED SCORES THAT CORRELATED BEST WITH TOTAL ESSAY SCORES. TWO OF THE METHODS INVESTIGATED, ONE IN WHICH THE NUMBER OF SCORES WAS MULTIPLIED BY THE RANK CORRELATION SQUARED, AND THE OTHER IN WHICH THE NUMBER OF SCORES WAS MULTIPLIED ONLY BY THE RANK CORRELATION, YIELDED CORRELATIONS SLIGHTLY BELOW THE ECT VALUE FOR THE REDUCED SAMPLE. THE THIRD METHOD, USING FISHER'S R-Z TRANSFORMATION AND COMPUTING NZ, YIELDED A CORRELATION WITH THE TOTAL ESSAY SCORE WHICH WAS SLIGHTLY HIGHER. THE IMMEDIATE CONCLUSION DRAWN WAS THAT UNDER THE BEST PREDICTIVE CONDITION, MULTIPLE REGRESSION WEIGHTINGS, THERE IS LITTLE OR NO DIFFERENCE AMONG THE SCORING PROCEDURES. USING TOTAL REARRANGEMENT SCORES TO PREDICT ESSAY SCORES, THE ORIGINAL SCORING IS SUPERIOR. IT WAS DEMONSTRATED, HOWEVER, THAT FACTORS OTHER THAN THE SCORING TECHNIQUES PER SE, INHERENT IN THE DATA AND TESTING CONDITIONS, COULD HAVE SEVERELY RESTRICTED THE PREDICTIVE VALIDITY OF THE TEST. (AUTHOR)

BR 5-0216
PA 24

VALIDITY OF THE REARRANGEMENT EXERCISE AS A PREDICTOR OF ESSAY WRITING ABILITY

RESEARCH
COGNITIVE LEARNING
DEVELOPMENT

WISCONSIN RESEARCH AND DEVELOPMENT
CENTER FOR
COGNITIVE LEARNING

Technical Report No. 23

VALIDITY OF THE REARRANGEMENT EXERCISE AS A PREDICTOR OF ESSAY WRITING ABILITY

Julianne Joyce Conry

Based on a master's thesis under the direction of

Julian C. Stanley, Professor of Educational Psychology

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

May 1967

The preparation of this paper was supported under a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program.

Center No. C–03 / Contract OE 5–10–154

PREFACE

This report is based on the master's thesis of Julianne Joyce Conry. Members of the examining committee were Julian C. Stanley, Chairman; Donald M. Miller; and Bob B. Brown.

The primary goal of the Wisconsin R & D Center for Cognitive Learning is to extend knowledge about, and to improve educational practices related to, cognitive learning in children and youth. Controlled experimentation is requisite for achieving this objective. The Laboratory of Experimental Design, part of the technical section of the R & D Center, provides valuable assistance to project directors in the design of experiments and also in the analysis of data. Further, the staff of the LED are charged with extending knowledge about experimental designs, scaling procedures, data analysis and the like.

This technical report is the third in a series describing new developments in the methodological area. In it, variations in scoring the rearrangement exercise, in which the student is required to order a series of events, are investigated empirically. Data from the Paragraph Organization portion of the CEEB English Composition Test, generously loaned to the author by Educational Testing Service and the College Entrance Examination Board, were rescored to determine which method yielded scores that correlated best with total essay score. The results indicate that there was little difference in scoring systems derived from variations of rank correlation methods.

Herbert J. Klausmeier
Co-Director for Research

CONTENTS

LIST OF TABLES

I
INTRODUCTION AND BACKGROUND OF THE PROBLEM

THE REARRANGEMENT EXERCISE

Of theoretical and practical interest to educators in diverse subject matter fields is the rearrangement exercise. It may require the student to order a series of events or procedures, to determine cause and effect, or to organize sentences logically into a meaningful paragraph. A primary theoretical problem has been determining a satisfactory method of scoring the sequences. The literature was searched for investigations of several methods; considerable methodological work but little empirical evidence was found. The possibilities of the rearrangement exercise appear to be extensive, and the scoring problems are not prohibitive, yet little practical significance has been attributed to it.

DESCRIPTION OF EDUCATIONAL TESTING SERVICE RESEARCH

Included in the battery of College Entrance Examination Board achievement tests is an English Composition Test (ECT)[1] of demonstrated validity on an exceptionally reliable criterion of students' essay writing.

A 24-school sample of high school juniors and seniors was required to write five essays—two were each forty minutes long, and three were each twenty minutes long. All students were asked to write on identical topics: the forty-minute essays were expository - argumentative; the twenty-minute essays were descriptive, narrative, and expository, respectively. Each of these essays was read by five readers who rated the papers on a general-merit scale of three points—superior, average, or inferior. The resulting criterion score was a pooled estimate of each student's writing ability based on the judgments of 25 readers. The reading reliability of each essay (i.e., average agreement among readers) was ascertained to be from .74 to .78; the total essayreading reliability was .92. Each essay, treated as one item, yielded a total essay reliability of .84 (Katz, 1963).

If it could be demonstrated that an easily administered objective test (such as the English Composition Test) was closely related to the reliable criterion measure, the total essay score, large-scale testing programs could well benefit from this knowledqe. The staff of Educational Testing Service undertook a study to examine the predictive power of eight objective item types. These subtests included Prose Groups, Error Recognition, Interlinear (Expository), Interlinear (Narrative), Construction Shifts, Sentence Correction, Usage, and Paragraph Organization. It became clear that the objective items were effective predictors of students' writing ability, yielding correlations ranging from .548 to .707 for the first seven groups. The eighth, Paragraph Organization, was correlated with the criterion only .458 for the total group and .426 for a more restricted group of students who had taken the PSAT as juniors. Essentially on the basis of this study, the Paragraph Organization, or rearrangement item, was then eliminated from the item pool available for the new forms of the English Composition Test. Certain idiosyncrasies associated with the rearrangement item in general, and the paragraph organization items in particular, however, justified further study by the present author.

DESCRIPTION OF THE ENGLISH COMPOSITION TEST

Section II of the English Composition Test consists of four prose paragraphs and one short poem, in scrambled order. The student is asked to read each group and decide what would be

[1] English Composition Test, Form JCBQ2, 495002. Princeton, N.J.: Educational Testing Service, 1958.

the best order in which to arrange the parts to produce a well-organized unit. For example, for a five-part exercise, the following six options are used:

a) Which sentence did you put first?
b) Which sentence did you put after A?
c) Which sentence did you put after B?
d) Which sentence did you put after C?
e) Which sentence did you put after D?
f) Which sentence did you put after E?

Thus, for an exercise requiring the ranking of n options, the total possible number of points is n + 1. Given this set of responses on answer sheets it was then possible to reconstruct the original rank order.

A systematic attack of this kind of item would seem to require the student to:

1) determine an acceptable order,
2) write it down,
3) convert it to comply with ECT instructions, and
4) record the sequence on IBM answer sheet.

It is obvious that this procedure is inefficient. When compared with other subtests of the ECT, it would appear to be a poor predictor, for the very reasons that the inefficiency would yield fewer items completed in the allotted 20-minute period and probably cause more clerical errors. This would result in lower reliability, contributing to lower validity. Short-cutting this method often leads to non-reconstructable sequences. In the sample of the responses of 533 students to five such exercises, obtained from Educational Testing Service, 116 (i.e., 22%) were eliminated for this reason.

The student is given one sample paragraph of three sentences to attempt before beginning the examination. The Paragraph-Organization subtest itself contains 5 rearrangement exercises:

1) Set A consists of five sentences to be arranged and six questions to be answered: The content to be organized requires a comparison of the personal reactions of Churchill and Hitler to wartime safety precautions.

2) Set B consists of six sentences to be rearranged and seven questions to be answered. The content is a description of the physical and chemical properties of a new product.

3) Set C consists of a narrative poem of six stanzas to be rearranged and seven questions to be answered.

4) Set D consists of five sentences to be rearranged and six questions to be answered. The content is a development of arguments for a choice of alternative number systems.

5) Set E consists of eight sentences to be rearranged and nine questions to be answered. The context describes the inevitable results of a belief in power for power's sake.

According to how well the items on the answer sheet correspond to the key, the student is awarded up to a possible 6, 7, 7, 6 or 9 points for each of the five sections. However, it may be noted that, using the ECT scoring system, each rearrangement exercise is treated as being composed of n separate items, rather than as a complete unit in which the parts are not independent of one another.

The most critical deficiency of the ECT scoring system can be noted from Table 1. Assuming that a rank correlational method is, indeed, most desirable from the theoretical considerations, the ECT is inconsistent in the manner in which it awards points. For five things ranked, an individual receiving 0 points could have a sequence correlating between -1 and +.8 with the key. For one point the range is -.9 to +.8; for two points the values are -.8 to +.6; for three points the values are -.5 to +.9. Only six points are uniquely determined by a correlation of 1.0. The worst possible arrangement, a reversal of the key, yields a rank correlation of -.1 and receives 0 ECT points. However, by merely interchanging adjacent ranks and retaining a correlation of .8, again 0 points are awarded. This seems to be inefficient as a predictor simply because all patterns are treated alike. A similar dispersion would be expected for six things ranked and eight things ranked.

It will also be noted that, on the basis of random rank orders, it would be easier to obtain one point than 0 points (probability 48/120 as compared with 36/120).

Consequent to the procedure of converting a five-rank sequence to a six-option question is the absence of five and four as possible point values. For three things ranked, 0, 1, and 4 points are permitted by the ECT method; for four things ranked, 0, 1, 2, and 5 points are possible. In general, for n + 1 categories of ECT scoring, values of n - 1 and n are impossible.

An example will clarify the reason for this. Suppose the examinee is asked to rank-order the four-item sequence A, B, C, D. The resulting pattern on the answer sheet will be

a) Which sentence did you put first? — A
b) Which did you put after A? — B
c) Which did you put after B? — C
d) Which did you put after C? — D
e) Which did you put after D? — O

If the keyed order were DBAC, the correct pattern would be DCAOB; the individual would receive no credit.

TABLE 1

Comparison of Spearman's Rho and Points Awarded on ECT for 120 Combinations of Five Things Ranked

ECT	Values of r_R																				
Points	-1.0	-.9	-.8	-.7	-.6	-.5	-.4	-.3	-.2	-.1	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0	Total
6																					1	1
3						1					2		2		4			6		4		19
2			3		3	1[a]						6					3					16
1		4			4			8		8	2		4	8		4	4[a]		2			48
0	1[a]			6[a]		4	4	2[a]	6	2	2	4		2		2[a]			1[a]			36
TOTAL	1	4	3	6	7	6	4	10	6	10	6	10	6	10	4	6	7	6	3	4	1	120

[a] Indicates pairs of correlations and point values not observed in the data

The ECT method, a comparison with a key of five options, would actually be considering all possible permutations of five things ranked, rather than of four things ranked. One of the pattern possibilities yielding three points by this method could be DOACB. However, it is impossible to reconstruct a rank order from this sequence.

a) Which sentence did you put first? — D
b) Which did you put after A? — none
c) Which did you put after B? — A
d) Which did you put after C? — C
e) Which did you put after D? — B

D
C follows C
B
A

Therefore, there must be 120 (permutations of ABCDO) - 24 (permutations of ABCD) = 96 permutations lost in this manner. For five items ranked and six orders, the value is 720 - 120 = 600. The number would increase with the number of items to be ranked. By not permitting certain point values to exist at the high end of the point scale, there would be an appreciable effect in markedly restricting the discrimination among the better students.

It is difficult to rationalize the addition of the superfluous item in the ECT. It would, perhaps, spuriously raise the reliability of the test by increasing the number of items, when, in fact, the order is determined without it. At most, it creates confusion on the part of the examinee.

As a result of observing these discrepancies and inconsistencies, the present study was devised to investigate the predictive validity of the rearrangement exercise scored by rank correlational methods.

Educational Testing Service generously loaned answer sheets and IBM cards containing the criterion scores for 533 high school juniors and seniors. For each of the five rearrangement exercises of the ECT consistent pattern responses were converted to the original rank-order and were then rescored by three systems. Spearman's rho was the rank-correlation statistic computed rather than Kendall's tau because the greater variability of rho for given ECT scores was expected to be more discriminating. A comparison of tau and rho values and the obtained ECT scores is made in Table 2.

The first method was an nr_R conversion, where n is the number of items to be rank-ordered. The possible point values range from -n for a reversal of the key to +n for the keyed sequence.

The second method was a $\pm nr_R^2$ conversion, with values again ranging from -n to +n. This procedure is advocated by Stanley (1964) for the reason that using the score formula nr_R, the values of r_R are treated as an equal-unit scale. By using nr_R^2, or n times the predictable portion of the variance, an attempt is made to compensate for the inequality.

To create a scale of more equal-appearing intervals, Fisher's r-to-z transformation was used in the third scoring situation.

The following questions were asked:

1. a) Will the total rearrangement-test scores differ in the extent of predictive validity under the four methods of scoring?

TABLE 2

Values of Rho and Tau by ECT Points, Five Things Ranked

Values of Tau	0	1	2	3	6
1.0					1.0
.8				.9 .9 .9	
.6	.8	.8 .8		.7 .7 .7 .7 .7 .7	
.4		.5 .5 .6 .6 .6 .5 .6 .5	.6 .6 .6	.4 .4 .4	
.2	.5 .3 .3 .5	.3 .3 .3 .3 .2 .3 .2 .2 .3 .2 .3 .3		.2 .4 .0 .2 .0	
0.0	-.1 .1 .0 .1 .1 .0 .1 -.1	-.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1	.1 .1 .1 .1 .1 .1		
-.2	-.2 -.2 -.3 -.2 -.2 -.3 -.2 -.2	.0 -.3 -.3 -.3 .0 -.3 -.3 -.3 -.3 -.3	-.5	-.5	
-.4	-.4 -.5 -.5 -.4 -.4 -.5 -.5 -.4	-.6 -.6 -.6 -.6	-.6 -.6 -.6		
-.6	-.7 -.7 -.7 -.7 -.7 -.7		-.8 -.8 -.8		
-.8		-.9 -.9 -.9 -.9			
-1.0	-1.0				

b) Will treating each of the five rearrangement exercises as one item rather than as six or more items increase the predictive validity?

2. Will differential weighting of the five rearrangement exercises increase the predictive validity?

3. Are some of the five rearrangement exercises significantly more valid than others, regardless of the scoring procedures?

The multiple regression RGR computer program was employed to compute a correlational table and multiple regression equation for each of the three sets of scores.

II
LITERATURE REVIEW

THE HISTORY OF THE REARRANGEMENT EXERCISE

The history of the scoring of the rearrangement exercise evolved from the appearance of an article by Wilson (1926). He was one of the first to recognize the comprehensive nature of this type of objective test item and at the same time recognized problems dealing with adequate scoring. It was noted that the procedure of giving credit ıere options were placed in the correct positions failed to distinguish between small and large errors. This factor increases in importance as the number of options increases.

Since then, advocated scoring methods have stemmed from variations of four approaches. Points are awarded based on

a. Number of correct sequences
b. Number of correct relations perceived
c. Sums of differences between the student's order and the keyed order
d. Rank correlation between student's order and the keyed order

The distinctions are generally made between number of errors and extent of errors.

Wilson introduced the rearrangement exercise as ideal for testing continuity of events in history. He proposed that the length be no longer than eight or twelve items, with each item having a definite location in time, with no overlap between items. A distinct cause and effect relationship should exist between items, taken into account by a scoring method which would measure the relation of each item to all other items.

In a later article (Wilson, 1930), an error of calculation was corrected and the following method was provided.

Given an exercise requiring the ranking of nine items with the correct order keyed as 1, 2, ..., 9, there are a total of 36 time relations among the items.

1 has a proper time relation to 2,3,4,5,6,7,8,9..8
2 has a proper time relation to 3,4,5,6,7,8,9..7
3 has a proper time relation to 4,5,6,7,8,9..6
4 has a proper time relation to 5,6,7,8,9..5
5 has a proper time relation to 6,7,8,9..4
6 has a proper time relation to 7,8,9..3
7 has a proper time relation to 8,9..2
8 has a proper time relation to 9..1
Total 36

For n items, the total number is $\frac{1}{2}n$ (n - 1). This is to be considered the perfect score for an exercise of a length of n items. In order to score a given permutation, the number of correct time relations are tabulateo and summed.

Nesmith (1929) recommended a similar method which would scale the number of correct relations to a maximum total score of 100. Given the same exercise of nine items to be rearranged, the following questions are asked:

a. Is 2 below 1—anywhere, not necessarily immediately below?
b. Is 3 below 2—anywhere, not necessarily immediately below?

⋮

h. Is 9 below 8—anywhere, not necessarilv immediately below?

For nine items, there are eight questions to be asked, and for each affirmative answer, 100/(n - 1) or $12\frac{1}{2}$ points are awarded. A maximum of 100 points is possible for a perfect sequence. Wilson (1930) points out, however, that this method doesn't consider all relations between items, but only those which immediately precede and follow it. Neither does it proportionally weight errors in pupil markings.

Worcester (1930) added a note of finality as he proposed that the item type be abandoned for lack of a suitable technique of scoring. Attacking the rationale of the previous articles, he commented that many wide displacements are caused by complete ignorance, so that the position is assigned by chance. By the Wilson method, these chance errors (and poor guesses) are overly penalized. Furthermore, the pupil would rarely compare an item with every other item before placing it within the sequence, but rather would establish certain "landmarks" for comparisons. As an alternative, he suggested a series of multiple choice questions requiring the pupil to compare each item with each of the others and indicating which should appear first in the sequence.

Cureton and Dunlap (1930) advocated the use of the actual rank correlation as a basis for determining point values. Applying the formula for Spearman's rho, $\text{rho} = 1 - (6\Sigma d^2)/(n^3 - n)$, values would be obtained ranging from a minimum of -1 to a maximum of +1. In order to conform with standard grading practices, he proposed a conversion to a percentage score so that scores range from 0 to 100%, with most of the scores in the upper 50%. The formula, $100 - (300\Sigma d^2)/(n^3 - n)$, yields scores above 50% associated with positive correlations and scores below 50% associated with negative correlations.

Odell (1928, 1930) and Russell (1926) advocated a method of sums of absolute differences between the student's ranking and the keyed ranking. Russell provided a conversion table for sums of differences: Odell subtracted the individual's difference sum from the maximum possible sum (the difference being denoted as $\Sigma D - \Sigma d$).

John (1930) compared four methods of scoring the continuity test: Spearman's rho, Wilson's method, Nesmith's method, and the Sangren-Woody method. The Sangren-Woody (1927) method is essentially the same as the Odell method with $S = \frac{1}{2}(\Sigma D - \Sigma d)$. The significance of this study lay in the use of Spearman's rho as the criterion for evaluation of the other techniques. The assumption had been made that rho is too complicated to be used as a practical scoring method. However, results of a study by McNamara (1936) indicate that scoring the exercise making use of tables especially prepared to convert sums of squared differences to rho is only slightly more time consuming than the other variations.

The following product-moment correlation coefficients were determined.

Nesmith—Spearman	.377
Wilson—Spearman	.978
Sangren—Spearman	.938
Wilson—Sangren	.930
Wilson—Nesmith	.509
Sangren—Nesm.:h	.307

As alternatives to rho, the Wilson and Sangren-Woody methods were rated satisfactory for classroom use. A study by Odell (1931) essentially confirmed these results. When compared with rank correlation scores and in terms of speed of scoring, the Wilson, Sangren-Woody, and Odell methods were advocated.

Sims (1934) pointed out that although the three methods advocated by McNamara may correlate well with rho, he objected to the fact that the maximum possible score depends on the length of the set rather than the number of exercises. Rearrangement tests of five, ten, and fifteen item exercises were evaluated and found to intercorrelate sufficiently high to allow freedom in determining length of the test for a specific purpose. However, scores on tests of differing lengths of sets cannot be compared. Nevertheless, using a formula $S = \Sigma D - \Sigma d$, he developed a conversion to be made so that a maximum score of n rather than ΣD could be obtained.

When the number to be arranged is even, $\Sigma D = n^2/2$, therefore, $\Sigma D/n = n/2$. Dividing S by the ratio $\Sigma D/n$, he obtained the formula

$$(1) \quad S' = (\Sigma D - \Sigma d)n/\Sigma D = n - (2\Sigma d)/n.$$

The worst possible score is still 0. When n is odd, $\Sigma D = (n^2 - 1)/2$ and $\Sigma D/n = (n^2 - 1)2n$.

$$(2) \quad S' = [(n^2 - 1)/(n - \Sigma d)](2n/(n^2 - 1))$$

when simplified is equal to $n - (2n\Sigma d)/(n^2 - 1)$.

Formula (1) is greater by $\Sigma d/(n\Sigma D) \leq 1/n$ since $\Sigma d/\Sigma D \leq 1$.

The error introduced in using formula (1) for n odd is never greater than 1/n, and therefore can be used.

Compared with the testing time of 3-4 recall items or 2-6 recognition items per minute suggested by Ruch (1929), the rearrangement time is not prohibitive. For a 5-item test, 3.5 items per minute was the average rate of completion. For a 10-item test, the average was 3.3 items per minute, and for a 15-item test, 2.6 items per minute.

In order to compensate for the effect of chance or guessing on rearrangement scores, Conrad (1936) used Russell's and Odell's method to consider the frequencies of discredit scores or sums of differences for the n! permutations of n things ranked. The median discredit score and those sums greater than the median would be awarded zero points, requiring the student to achieve a score higher than (or sum of differences lower than) chance expectation in order to receive any credit for the test.

The maximum score is $\frac{1}{2}n$ for n even, and $\frac{1}{2}(n + 1)$ for n odd. No negative point scores were considered. Conversion of discredit scores to point scores are made by means of a table furnished for tests consisting of from two to twenty items. The gradation of credits for given discredits is not uniform. Conrad assumes that as one approaches the zero-credit limit, the possibility of sheer guessing increases rapidly. However, his particular gradation was apparently arbitrarily assigned.

In 1942, Rosander considered the rank order as a measure of discriminal ability, and the number of inversions of elements was the scoring index he proposed. Interchanging any pair of events constitutes one inversion; displacing one element two positions or two elements one position represents two inversions. If I_i is the number of items ranked after x_i which should have been ranked before x_i, then ΣI_i is the number of inversions in the student's sequence. The maximum number of inversions is $n(n-1)/2$; the mean number by chance is $n(n-1)/4$. The frequency distribution of number of inversions approaches the normal probability curve as n increases.

Score points may be assigned as a function of a) the X number of inversions, b) a discriminal score ($X_{mean} - X$), or c) $X_{maximum} - X$. Rosander furnishes a table of the 24 permutations of four things ranked, 10 permutations which may be chosen as the keyed order, and the associated point values. The teacher would be required to locate the pupil's ranking and her key and to record the score. Considering the inconvenience of developing tables for exercises requiring more than four items ranked (e.g., 6 = 720 pe.mutations), this technique would certainly be prohibitive.

Note that this system appears to be nothing more than the reverse of the Wilson method outlined previously. Wilson (1930) tabulated the number of correct "time relations"; Rosander is merely tabulating the number of relationships which are incorrect. Both resemble the approach taken by Kendall (Ferguson, 1959) in developing the tau rank correlation statistic. Except as an illustration of the continuation of attention being directed toward the rearrangement exercise, little more is added to the development of suitable scoring techniques. Operationally more graphic, but yielding similar results, is a method given by Hearley and Venables (1936) for rearranging topic sentences and literature quotes.

Odell (1944) shortened the rank correlational methods for practical use by tabling sums of positive differences (pupil's rank order - keyed order) for an R approach [See Spearman's Footrule (1904): $R = 1 - (3\Sigma d)/(n^2 - 1)$] and sums of squares of differences for a rho approach. He recommends the latter.

Odell acknowledged the three disparate theories of minimum scores which are tabled accordingly:

1. 0 points for reversal (i.e., minimum score) and $\frac{1}{2}$n for chance.
2. Negative points for reversal (i.e., misinformation) and 0 points for chance.
3. 0 points for chance or worse.

It is quite clear that correlational methods of scoring the rearrangement test are far superior to a method of exact order sequences because the former take the extent of misinformation into account. However, with some uses of the rearrangement exercise, a sense of exact relationship might be crucial to the problem, e.g. expression of cause and effect relationships. A method was suggested by Montgomery (1946) to take into account both the number of items in correct position and the distance of each item from its correct position.

First, the highest possible sum of differences is determined by a comparison of the correct and reverse orders. Scores are translated into a ranking grade (denoted by RGr) on a 100% basis: for example, in a 10 item test, RGr = 100 - 2ΣD, since the maximum sum of differences is 50.

Secondly, a score is determined by the summation of the number of events placed in proper sequence (denoted by Sq) to events immediately preceding them. For example, 1, 2*; 5, 6*, 7*. In a 10 item list, then, only 9 correct sequences are possible. The value of each check is, in this case, 100/9 = 11.1. The sequence-rank grade is an average of these two: SqGr = (RGr + 11.1 × Σ Sq)/2. Each number of item series varies 1) the highest possible ΣD yielding a coefficient J, where J = 100/ΣD, and 2) the number of correct sequences possible, (n - 1), and its coefficient, K, where K = 100/Sq.

Although no set procedure is suggested for constructing a continuity test scored in this manner, a frequency table of sequences correctly ordered can indicate item difficulty.

Three years later, Ezell (1949) prepared tables to convert sums of squared differences to point values by substituting rho = 0.0, .2, .4, .6, .8, and 1.0 in Spearman's formula, rho = $1 - (6\Sigma d^2)/n(n^2 - 1)$ to give the lower limit of five equal steps of sums. The method would give more points to the student more closely approximating the correct order, still to the dismay of the statistician who would deplore the non-equality of rho intervals. Ezell dismisses this problem.

Stanley (1964) averted this difficulty by using rho directly and converting to point values by the formula ± nr_R^2, where n is the number of items to be ranked, and rho is denoted by r_R.

r_R^2 is used rather than r_R since r_R^2 indicates the proportion of variance accounted for by the correlation. He also provides a table of sums of squared differences for facility in scoring.

Although a considerable amount of study has been carried out regarding systems of scoring or rearrangement items, Ashburn and Bradshaw (1953) after comparing 10 scoring methods, drew the conclusion that none of the methods were adequate and that the continuity-type question should be discarded altogether.

Students were asked to arbitrarily order ten fictitious countries by size. These rankings were scored by each of the ten following methods:

1. Based on possible sum of differences; score 0 for complete reversal
2. Based on sum of differences; $\frac{1}{2}$ of maximum sum receives 0; greater than $\frac{1}{2}$ receives negative scores
3. Based on sum of differences; $\frac{1}{2}$ or more of the maximum sum receives 0
4. Based on rank correlation of sum of squares of possible differences; chance arrangement ($\frac{1}{2}$ of maximum sum) receives 50 points, perfect negative correlation receives 0
5. Based on rank correlation; $\frac{1}{2}$ of maximum possible sum of squared differences receives 0, negative correlations receive negative scores
6. Based on rank correlation; chance scores and all negative correlations receive 0
7. Based on number of order relationships; reversal receives 0
8. Based on number of order relationships; answer of $\frac{1}{2}$ or less of all possible order relationships receives 0
9. Based on number of exact order sequences
10. Average score based on sum of differences and number of exact sequences (methods 1 and 9)

The range of scores for each student totally ignorant of the ranking was considerable.

In a second experiment, students were given information pertaining to relationships between the countries. By correct application, they should have perceived the relative size of all but two of the ten (the information enables the student to establish 32 of 45 order relationships, or 71%). Since those two positions would be due to chance arrangement, none of the students should receive a score over 80 or below 71.

None of the methods assigned a majority of scores between these values. 38% of the scores were made by students who failed to correctly make use of the information supplied and were given credit only by chance. Variation of scores was as much as 97 points on the same paper for a student who did not correctly apply the information.

It was assumed that the results were a function of a) inadequacies of scoring, b) chance, or c) failure to apply the given information. An unduly heavy emphasis was placed on the first, with a casual dismissal of the other two.

It would seem entirely reasonable to attribute a fair share of the unexplained discrepancy in scoring to the second and third. The point might be made that perceiving the relationships when given information and being able to apply it correctly is no simple task for students trained only in the memorization of absolute sequences.

In several academic areas, it is desirable to test students' ability to perceive relationships between items: chronologically, preferentially, sequentially. The rearrangement exercise would seem to be the most straightforward method of testing this ability. The conclusion drawn by Ashburn and Bradshaw would seem unwarranted.

One obvious example is the rearrangement of historical events. Ideally, success on such a test should be commensurate with knowledge of relationships rather than sheer memorization of dates which are soon forgotten. Leichty (1954) attempted to ascertain "Student Thinking on Items Involving Chronology" by interviewing students during an oral retake of a History of Civilization exam. The type of reasoning displayed was assigned to categories including "acceptable grounds," "unacceptable grounds," "grounds vague," "outright guesses," "bad reasoning," and so forth.

In contrast to none of the A students, he found that one-third of the answers given by D students on this exam represented outright guesses, and none were correct guesses. Even the better students were uncertain of chronology, and the weaker ones had practically no concept of it whatsoever. On the surface, it would appear that such an ability is not developed automatically as a student learns a series of events.

The 1936 study by McNamara sought to determine the relationship between total score and the construction of the rearrangement test. Specifically, it was questioned whether there was a tendency for percentage of correct responses to decrease as a function of displacement of the item from its correct position.

Students were given three forms of four tests with statements varying in the number of steps away from their correct positions. These were scored by three methods:

a) number of single items correct
b) the Odell method, and
c) the rho method of rank correlation

Rho was used as a criterion against which the other methods were rated. The correlations between the Odell method and the rho scores ranged from .86 to .94; the correlations between single-items-correct scores and rho ranged from .59 to .81. The expected relationship was not found. In fact, even when items were presented in their correct orders, there was no consistent tendency for the pupils to be more successful in any of the three methods of scoring. Leichty does not recommend this as a standard test construction practice, however.

An additional consideration, the average scoring time required in items per minute for 8- and 10-item series was also ascertained for the three methods.

Single-items-correct:	55/min
Odell Method:	31/min
rho score—writer's tables:	21/min
rho score—nomograph:	17/min

Following 34 years of sporadic interest in the rearrangement test, a recent attempt to revive the exercise was made by Cureton (1960). He presented a table to convert sums of differences to point values on the basis of the Spearman Footrule. A score of zero was awarded when agreement was random or worse, the correlation being zero or negative.

It would seem reasonable to suppose that the difficulty of the rearrangement exercise could be varied by increasing or decreasing the distance between statements; logically, in the case of English composition, or chronologically, in the case of historical events. There would seem to be an optimal limit to the number of items in an exercise as a requirement for ease of scoring by any of the aforementioned methods, and for a minimum of student confusion. After a certain point, the length of the test would be a correlate of general intelligence and clerical aptitude rather than an organizational ability in history, science, or English. Each item must be distinct and have a definite location in time, with no items overlapping. However, research in this aspect of the continuity problem is lacking.

Throughout the chronological development of scoring schemes for the rearrangement exercise, the trend has been toward a rank correlational method:

1) the use of sums of absolute differences,
2) sums of squared differences, and lastly
3) Spearman's rho as a criterion against which the other methods were evaluated.

In order to interpret the English Composition Test's method of scoring the Paragraph Organization subtest, Spearman's rho and variations of it were used.

III
RESULTS

A significant problem influencing the results of this study was the 116 individuals whose answer sheets contained one or more pattern responses which were non-reconstructable and could not be rescored. In a sense, therefore, the data used in this analysis were contaminated insofar as the particular items which could not be reconstructed, the individuals who were eliminated, characteristics of the ECT scoring method, and immediate effects on the students taking the exam interacted to obscure the answers to the specific questions being investigated. Nevertheless, the tedious task of attempting to reconstruct the 533 individuals' Paragraph Organization subtests revealed recording deficiencies in the ECT method.

Many of the non-reconstructable items obviously represented clerical errors. A sequence would appear incomplete on the answer sheet, yet by omitting one of ECT's n + 1 questions for an n option exercise, the rank order could still be completely reconstructed. Students who had established the keyed sequence were therefore penalized by the ECT method of omitting one question, since they would receive n points instead of the n + 1 to which they were entitled by this system.

Other recording errors were multiple responses to the same item. Often the sequence was reconstructable without considering that particular response. Presumably by machine scoring the ECT answer sheets, the multiple response would not be credited and the student would again be penalized.

On the other hand, unreconstructable sequences were often awarded one or more points when several options matched the key.

Similarly, students often received one or more points by the ECT method for having established the first members of the sequence, leaving other options blank, with no complete order in mind. The rank-correlation methods of scoring have no provision for omitted options in a rank order; these were then rescored zero. In the context of the English Composition Test (Paragraph Organization), not only is a beginning topic sentence required, but essential is a logical follow-through to the completing sentence.

Another set of scores was eliminated when the student apparently had an order in mind, but the sequence could not be reconstructed, perhaps through confusion of directions. An example of this is the "random" permutation illustrated on page 3, appearing more than once among the 116 eliminated individuals.

When all five exercises were reconstructable the student's scores were retained in the sample. If one or more exercises were omitted, zero points were awarded and the scores were included in the analysis.

In spite of the theoretical considerations regarding the comparisons between ECT points awarded and a corresponding range of correlations associated with each value, it was argued that although the range of correlations is possible for any given ECT point value, in practice deviant patterns producing this range would rarely, if ever, occur. This was not the case, however. Although a table was not prepared for the permutations of six things ranked or eight things ranked, a tabulation was made of the permutations of five things ranked by ECT points (see Table 1). Note that, in this table, the ECT scores are not corrected for chance as they would be on the actual answer sheets.

Subtests A and D of the Paragraph Organization tests require the rearrangement of five sentences. Of the 120 possible permutations of five things ranked, only the seven starred pairs of correlations and point values did not appear among the response patterns. Within the sample studied, it was observed in subtest A, for example, that after having correctly identified the first element of the sequence, among the other four sentences to be ranked, all 24 possible permutations appeared. Similarly on subtest D, 17 of the 24 possible permutations appeared. The six- and eight-item exercises did not produce the same proportionate number of patterns.

Tables 3—7 represent conversions of sums

TABLE 3

Rescoring of Exercise A, Five Things Ranked

D^2	r_R	$\pm nr_R^2$	nr_R	nz
0	1.0	5	5	15*
2	.9	4	4	7
4	.8	3	4	5
6	.7	2	4	4
8	.6	2	3	3
10	.5	1	2	3
12	.4	1	2	2
14	.3	0	2	2
16	.2	0	1	1
18	.1	0	0	0
20	0.0	0	0	0
22	-.1	0	0	0
24	-.2	0	-1	-1
26	-.3	0	-2	-2
30	-.5	-1	-2	-3
32	-.6	-2	-3	-3
36	-.8	-3	-4	-5
38	-.9	-4	-4	-7

*Note: z truncated at $r = .995$.

TABLE 4

Rescoring of Exercise B, Six Things Ranked

D^2	r_R	$\pm nr_R^2$	nr_R	nz
0	1.00	6	6	18
2	.94	5	6	10
4	.89	5	5	8
6	.83	4	5	7
8	.77	4	5	6
10	.71	3	4	5
12	.66	3	4	5
14	.60	2	4	4
16	.54	2	3	4
18	.49	1	3	3
20	.43	1	2	3
22	.37	1	2	2
24	.31	1	2	2
26	.26	0	2	2
30	.14	0	1	1
32	.09	0	0	0
36	-.03	0	0	0
46	-.31	-1	-2	-2
48	-.37	-1	-2	-2
50	-.43	-1	-2	-3
54	-.54	-2	-3	-4
56	-.60	-2	-4	-4

TABLE 5

Rescoring of Exercise C, Six Things Ranked

D^2	r_R	$\pm nr_R^2$	nr_R	nz
0	1.00	6	6	18
2	.94	5	6	10
4	.89	5	5	8
6	.83	4	5	7
8	.77	4	5	6
10	.71	3	4	5
12	.66	3	4	5
14	.60	2	4	4
16	.54	2	3	4
18	.49	1	3	3
20	.43	1	3	3
22	.37	1	2	2
24	.31	1	2	2
26	.26	0	2	2
28	.20	0	1	1
30	.14	0	1	1
32	.09	0	0	0
34	.03	0	0	0
36	-.03	0	0	0
38	-.09	0	0	0
40	-.14	0	-1	-1
42	-.20	0	-1	-1
44	-.26	0	-2	-2
46	-.31	-1	-2	-2
52	-.49	-1	-3	-3
56	-.60	-2	-4	-4
66	-.89	-5	-5	-8

TABLE 6

Rescoring of Exercise D, Five Things Ranked

D^2	r_R	$\pm nr_R^2$	nr_R	nz
0	1.00	5	5	15
2	.90	4	4	7
4	.80	3	4	5
6	.70	2	4	4
8	.60	2	3	3
10	.50	1	2	3
12	.40	1	2	2
14	.30	0	2	2
16	.20	0	1	1
18	.10	0	0	0
20	0.00	0	0	0
22	-.10	0	0	0
26	-.30	0	-2	-2
28	-.40	-1	-2	-2
32	-.60	-2	-3	-3
36	-.80	-3	-4	-5

TABLE 7

Rescoring of Exercise E, Eight Things Ranked

D^2	r_R	$\pm nr_R^2$	nr_R	nz
0	1.00	8	8	24
2	.98	8	8	18
4	.95	7	8	15
6	.93	7	7	13
8	.90	6	7	12
10	.88	6	7	11
12	.86	6	7	10
14	.83	6	7	10
16	.81	5	6	9
18	.79	5	6	8
20	.76	5	6	8
22	.74	4	6	8
24	.71	4	6	7
26	.69	4	6	7
28	.67	4	5	6
30	.64	3	5	6
32	.62	3	5	6
34	.60	3	5	6
36	.57	3	4	5
38	.55	2	4	5
40	.52	2	4	5
42	.50	2	4	4
44	.48	2	4	4
46	.45	2	4	4
48	.43	1	3	4
50	.40	1	3	3
52	.38	1	3	3
54	.36	1	3	3
56	.33	1	3	3
58	.31	1	2	2
60	.29	1	2	2
62	.26	0	2	2
64	.24	0	2	2
66	.21	0	2	2
68	.19	0	2	2
70	.17	0	1	1
72	.14	0	1	1
78	.07	0	0	0
80	.05	0	0	0
84	0.00	0	0	0
86	-.02	0	0	0
88	-.05	0	0	0
90	-.07	0	0	0
92	-.10	0	-1	-1
94	-.12	0	-1	-1
96	-.14	0	-1	-1
98	-.17	0	-1	-1
102	-.21	0	-2	-2
104	-.24	0	-2	-2
108	-.29	-1	-2	-2
110	-.31	-1	-2	-2
112	-.33	-1	-3	-3
114	-.36	-1	-3	-3
116	-.38	-1	-3	-3
128	-.52	-2	-4	-5
140	-.67	-4	-5	-6
142	-.69	-4	-6	-7
144	-.71	-4	-6	-7
148	-.76	-5	-6	-8
162	-.93	-7	-7	-13

of squared differences and rho equivalents to the three scoring systems used in the present study. Only those sums of squared differences actually observed in the data are recorded. It will be noted that there were frequent instances of negative subtest scores. Several negative total test scores were assigned. That a considerable range is represented may be noted as the correlation of student orders with the key varies from a perfect positive correlation to near-perfect negative correlation: +1.0 to -.9 for Exercise A, +1.0 to -.6 for Exercise B, +1.0 to -.89 for Exercise C, +1.0 to -.8 for Exercise D, and +1.0 to -.93 for Exercise E. Associated point values range from +n to -n for the nr_R and nr_R^2 scoring. The nz scoring is more differentiating at the extremes. This contrasts with the ECT method, which fails to differentiate fully at the high end of the scale because it does not provide for a complete range of points.

Five subscores and a total score were obtained for 417 individuals. The five exercises became variables in three multiple regression equations computed to predict the criterion. Of genuine concern was the effect of the elimination of 116 individuals from a sample which was initially restricted (individuals who had taken the PSAT-V). The assumption was made that conceivably individuals with non-reconstructable response patterns were generally less able students if these sequences reflected an inability to successfully organize an order and apply the ECT scoring system. Their mean total essay score was compared with the mean total essay score of the individuals retained in the sample. The eliminated students were found to have significantly ($\alpha = .01$) lower essay scores than the group used in the study. It was then anticipated that the correlation between the paragraph organization scores and essay scores would be attenuated due to restriction of range. The original correlation obtained by ETS was .426 for the PSAT group. With 116 individuals eliminated from the sample the correlation was only slightly reduced to .4156. With respect to the present study, other effects of eliminating individuals because of non-reconstructable responses are unknown, though it seems reasonable to surmise that the comparative conclusions are not affected seriously.

Correcting the multiple r for degrees of freedom (i.e. for shrinkage when cross validated), it was found that the nr_R^2 scoring correlated with essay scores .4127, and the nr_R scoring correlated .4026, both slightly less than the .4156 obtained by the original ECT method. The nz scoring correlated .4319 with the essay scores, or slightly higher than the original scoring. The simple sums of the points from the five rearrangement exercises correlate with the total essay scores .29, .23, and .32 respectively for the three scoring methods (see Correlation Table 8). So it is obvious that differentially weighting the scores on the five exercises raised the correlation with the essay-test criterion considerably.

For precise comparisons, the answer sheets should be rescored by the original method as five separate scores, and also correlated with total essay in a multiple regression equation. Preferably, they should be included as 35 separate items since the ECT treats each response independently of the 5, 6, or 8 item exercise of which it is a part. The additional rescoring was considered too time consuming to be feasible for this analysis.

That none of the three rank correlation rescoring methods was particularly more effective as a predictor than the others may be explained by the high intercorrelations. Among total paragraph organization scores, r ranged from .862 to .9552. All were greater than correlations with the original scoring, with r ranging from .650 to .8616 (see Correlation Table 8).

It will be noted from Table 9, Regression Coefficients, that Exercises A and B of the Paragraph Organization add essentially nothing to the prediction. The relative magnitude of the regression coefficients remained the same over the three scoring procedures.

On the basis of only one three-sentence practice problem, the first two items may be poor predictors because the student has not acquired the mental set required to attack the rest of the examination. That is, he cannot apply any essay-related organizational logic until he has mastered the recording technique.

The two items may be poor predictors in the sense that they are incomprehensible or confusing and do not differentiate relative abilities to organize a meaningful paragraph, but only dichotomously differentiate students who are either able or unable to decipher any meaning out of the sentences.

From the mean exercise scores (Table 10), it will be noted that Exercise A produced a lower mean score than the other five-item exercise, D. Similarly, Exercise B produced a lower mean score than the other six-item exercise, C. If the second of the two hypothesized reasons is correct, then the exercises evidently have not been scaled from easy to difficult within this section of the English Composition Test.

The fifth exercise, requiring the ranking of eight sentences, produced the lowest mean score, and contributed the least of the remaining three exercises. Presumably, this is the result of a speed factor involved in the test. Relatively few students completed this exercise.

Frequencies of omissions were tabulated by exercise for the total sample of 533 students. Comparing the five items, approximately 2% of the omissions were in Exercise A, 1% in B, 11% in C, 14% in D, and 72% in E.

It would seem that the content of Exercises C and D may have been more familiar and therefore better reflect an ability to logically organize the paragraph. If the student is unable to comprehend the more obscure paragraph, no matter how logically he is able to think, he will not be able to apply those abilities which are presumably being tested by this exercise, and which are one of the criteria for successful essay writing. Apparently a speed factor and practice effects may also have influenced achievement on this test.

Unknown is the relative effectiveness of the characteristics of the content, for example, causal chains as opposed to flexible comparisons or pure description, or, on the other hand, the use of particular connectives relating items. Exercises C and D, stanzas of a poem and exposition on number systems, may have more clearly forced certain patterns.

This may be related to the previously referenced comments by Worcester questioning whether the student compares an item with every other item before determining its position within the sequence. Perhaps this is controlled by the content of the exercise.

Wilson's list of desirable characteristics of rearrangement items acquires added importance for evaluating the test: 1) each item has a definite location in time, 2) items should not overlap, and 3) a distinct cause and effect relationship should exist between items.

The reader is referred back to the descrip-

TABLE 8

Correlation Coefficients

nr_R^2 Scoring						
A	B	C	D	E	Total	
1	2	3	4	5	6	
1.0000	.0069	-.0211	.0866	.1593	.4753	1
	1.0000	.0084	-.0435	.0275	.4962	2
		1.0000	.0737	.0359	.4667	3
			1.0000	.0905	.4418	4
				1.0000	.5251	5
					1.0000	6

nr_R Scoring						
A	B	C	D	E	Total	
7	8	9	10	11	12	
.9258	.0098	-.0227	.0578	.1496	.4629	1
.0009	.9648	-.0026	-.0439	.0369	.5373	2
-.0449	-.0121	.9544	.0897	.0388	.3732	3
.0507	-.0672	.0659	.9109	.1113	.3491	4
.1075	.0305	.0428	.0691	.9343	.5206	5
.3997	.4637	.4409	.3965	.5084	.9400	6
1.0000	.0137	-.0399	.0251	.0971	.4549	7
	1.0000	-.0190	-.0723	.0313	.5446	8
		1.0000	.0913	.0419	.3902	9
			1.0000	.0936	.3657	10
				1.0000	.5541	11
					1.0000	12

nz Scoring						ECT	PSAT-V	ESSAY	
A	B	C	D	E	Total				
13	14	15	16	17	18	19	20	21	
.9703	.0212	-.0131	.0675	.1449	.4495	.3230	.0648	-.0032	1
.0214	.9459	.0339	-.0420	.0315	.4407	.2403	.0195	-.0209	2
-.0135	.0300	.9505	.0843	.0182	.5002	.4984	.2332	.2304	3
.1058	-.0267	.0833	.9381	.1167	.4422	.4590	.3436	.3289	4
.1671	.0261	.0624	.0923	.9608	.4706	.3999	.1934	.1979	5
.4843	.4898	.4752	.4181	.5076	.9552	.7843	.3369	.2859	6
.8805	-.0007	-.0561	.0356	.0954	.3562	.2058	-.0065	-.0508	7
.0245	.8777	.0113	-.0580	.0290	.3941	.1703	-.0242	-.0486	8
-.0171	.0190	.8615	.0637	.0243	.4460	.4368	.2133	.2333	9
.0765	-.0263	.0856	.8443	.0981	.3925	.4019	.3084	.3004	10
.1575	.0396	.0643	.1094	.9024	.4596	.4032	.2078	.2176	11
.4640	.5028	.3573	.3250	.5029	.8620	.6502	.2549	.2324	12
1.0000	.0391	.0002	.0884	.1540	.4857	.3597	.0902	.0060	13
	1.0000	.0545	-.0260	.0273	.4856	.3212	.0275	-.0071	14
		1.0000	.0996	.0473	.5555	.5523	.2621	.2554	15
			1.0000	.1160	.4655	.4873	.3344	.3230	16
				1.0000	.4808	.4137	.2014	.2415	17
					1.0000	.8616	.3609	.3199	18
						1.0000	.4610	.4156	19
							1.0000	.6259	20
								1.0000	21

TABLE 9

Beta Coefficients, Normal Equation, and Multiple Correlation Coefficients

	Beta Coefficients, Normal Equation					Multiple Correlation Coefficients	
	Item 1	Item 2	Item 3	Item 4	Item 5	R	R - df Corrected
nr^2_R	-.05254	-.01378	.2009	.3024	.1721	.4246	.4127
nr_R	-.06782	-.0308	.1979	.2638	.1922	.4149	.4026
nz	-.05006	-.01535	.2185	.2813	.2066	.4431	.4319

TABLE 10

Means with Respect to Maximum Obtainable Value and Standard Deviations

	MEANS WITH RESPECT TO MAXIMUM OBTAINABLE VALUE							
	$\pm nr^2_R$	$nr^2_{R\ Max}$	$\pm nr_R$	$nr_{R\ Max}$	nz	nz_{Max}	ECT	ECT_{Max}
A	1.269	5	1.681	5	3.348	15	-	-
B	2.621	6	3.029	6	5.722	18	-	-
C	3.887	6	4.388	6	10.230	18	-	-
D	2.096	5	2.731	5	5.141	15	-	-
E	1.108	8	1.693	8	2.621	24	-	-
Total	10.980	30	13.520	30	27.060	90	12.980	35

	STANDARD DEVIATIONS			
	$\pm nr^2_R$	$\pm nr_R$	nz	ECT
A	2.105	2.483	5.796	-
B	2.701	3.361	6.713	-
C	2.345	2.236	7.244	-
D	1.976	1.932	5.368	-
E	2.194	2.729	5.223	-
Total	5.461	6.053	15.110	6.083

tions on page 2. Exercises C, D, and E of the Paragraph Organization subtest appear to best meet these criteria. They were found to be the substantial contributors to the regression. Therefore, the specific items themselves rather than the item type in general could account for the mediocre validity of the rearrangement exercise as a predictor of essay scores. Characteristics of discriminating rearrangement exercises have not been empirically studied.

IV
SUMMARY AND CONCLUSIONS

From theoretical considerations of the present ECT scoring method of the Paragraph Organization, it was anticipated that by reconstructing and rescoring the rearrangement exercises by variations of rank correlation methods the predictive validity of the examination would exceed the value of .426 obtained by the Educational Testing Service staff from their analysis. Two of the methods investigated, nr_R^2 and nr_R, yielded correlations slightly below the ECT value of .4156 obtained on the reduced sample. The third method, nz, yielded a correlation with the total essay score which was slightly higher. Without optimally weighting the five exercises, the product-moment correlations of total rearrangement scores with essay scores were reduced to .29, .23, and .32 for the nr_R^2, nr_R, and nz systems respectively.

The immediate conclusion that could be drawn would be that under the best predictive condition, multiple regression weightings, there is little or no difference among the scoring procedures. Using total rearrangement scores to predict essay scores, the original scoring is superior.

It was demonstrated, however, that factors other than the scoring techniques per se, inherent in the data and testing conditions, could have severely restricted the predictive validity of the test.

These results in no way deter from the value of the rearrangement exercise as a valid and valuable measuring device in a context other than this Paragraph Organization example and should not negate the principles and considerations of scoring procedures as explicated in previous chapters. Future study may be directed towards examination of characteristics of the ranking items themselves which are conducive to manipulation into a meaningful sequence. It may very well be that the rearrangement exercise is a more successful instrument when employed in an area such as history or science, rather than in English composition.

APPENDIX
PSAT-V PREDICTOR OF ESSAY SCORES

Although not directly related to the problem of scoring procedures, but concerned with predicting essay writing ability, is the use of PSAT-V scores in addition to the paragraph organization subtests. Using the three scoring methods used in the original analysis, the five scores were combined with PSAT-V in a multiple regression equation to predict total essay scores. The original correlation of PSAT-V with total essay was .6259 for the reduced group. This correlation was raised to .6450 using the nr_R^2 method, to .6458 using the nr_R method, and to .6519 using the nz transformation. All correlations were corrected for shrinkage. Regression coefficients are given in Table 11. It is significant that the PSAT-V scores appear to be one of the single best available predictors of essay writing ability.

TABLE 11

Beta Coefficients for Paragraph Organization and PSAT-V Scores

	1	2	3	4	5	PSAT
nr_R^2	-.06191	-.02865	.08877	.1302	.0876	.5481
nr_R	-.05499	-.02745	.09915	.1124	.09497	.5493
nz	-.07203	-.02461	.09670	.1236	.1253	.5412

BIBLIOGRAPHY

Ashburn, R., & Bradshaw, J. Experiment in the continuity type equation. *Journal of Educational Research*, 1953, *47*, 201-209.

Conrad, H. S. The scoring of the rearrangement test. *Journal of Educational Psychology*, 1936, *27*, 241-252.

Cureton, E. E. The rearrangement test. *Educational and Psychological Measurement*, 1960, *20*, 31-34.

Cureton, E. E., & Dunlap, J. W. Scoring the rearrangement or continuity test. *School Review*, 1930, *38*, 613-616.

English Composition Test, Form JCBQ2, 495002. Princeton, N.J.: Educational Testing Service, 1958.

Ezell, L. B. A device for scoring chronology tests. *Social Education*, 1949, *13*, 329-331.

Ferguson, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1959, pp. 183-186.

Hearley, M.J., & Venables, F.I. A new English and its marking. *Journal of Education* (London), 1936, *68*, 523-525.

John, L. A comparison of four methods of scoring the continuity test. *School Review*, 1930, *38*, 617-621.

Katz, M. (Ed.) *ETS Developments*, 1963, *11* (1), 1 & 4.

Leichty, V. E. Student thinking on items involving chronology. *Journal of Educational Research*, 1954, *48*, 187-194.

McNamara, W.J. Construction and scoring of the continuity or rearrangement test. *School Review*, 1936, *44*, 50-57.

Montgomery, E. O. Construction of the sequence rank test. *Journal of Educational Research*, 1946, *39*, 523-527.

Nesmith, R. W. Scoring the continuity test. *School Review*, 1929, *37*, 764-766.

Odell, C. W. *Traditional examinations and new type tests*. New York: Century, 1928, pp. 406-408.

Odell, C.W. *Educational measurement in high school*. New York: Century, 1930, pp. 495-597.

Odell, C. W. Still more about scoring rearrangement or continuity tests. *School Review*, 1931, *39*, 542-546.

Odell, C. W. The scoring of continuity or rearrangement tests. *Journal of Educational Psychology*, 1944, *35*, 352-356.

Rosander, A. C. A simple method of scoring and interpreting sequential responses. *Journal of Educational Research*, 1942, *36*, 168-177.

Ruch, G. M. *The objective or new type examination*. Chicago: Scott-Foresman, 1929, pp. 202-203.

Russell, C. *Classroom tests*. Boston: Ginn, 1926, pp. 95-102.

Sangren, P. V., & Woody, C. *Sangren-Woody reading test*. Yonkers-on-Hudson, N.Y.: World Book, 1927.

Sims, V.M. An evaluation of five-, ten-, and fifteen-item rearrangement tests. *Journal of Educational Psychology*, 1934, *25*, 251-257.

Spearman, C. The proof and measurement of association between two things. *American Journal of Psychology*, 1904, *15*, 72-101.

Stanley, J.C. *Measurement in today's schools*. Englewood Cliffs, N.J.: Prentice-Hall, 1964, 379-381.

Wilson, H. E. The continuity test in history teaching. *School Review*, 1926, *34*, 679-684.

Wilson, H. E. Further comments on the scoring of the continuity test. *School Review*, 1930, *38*, 115-123.

Worchester, D. A. Still further comments on scoring continuity tests. *School Review*, 1930, *38*, 462-466.