ED 017 960 CG 001 803

THE EFFECTS OF ITEM ORDER AND ANXIETY ON TEST PERFORMANCE AND STRESS.

BY- HAMBLETON, RONALD K.
ONTARIO INST. FOR STUDIES IN EDUC., TORONTO

PUB DATE 8 FEB 68

EDRS PRICE MF-$0.25 HC-$1.28 30P.

DESCRIPTORS- *TEST CONSTRUCTION, RESEARCH, MATHEMATICS, *STRESS VARIABLES, ITEM ANALYSIS, *PERFORMANCE FACTORS, ANXIETY,

THIS RESEARCH INVESTIGATED THE EFFECT OF ITEM ORDER ON THE PERFORMANCE OF A MATHEMATICS TEST, ON THE AMOUNT OF STRESS GENERATED DURING A TEST, AND ON THE PERFORMANCE OF HIGH AND LOW TEST ANXIOUS SUBJECTS. SOME 106 HIGH SCHOOL STUDENTS COMPLETED THE ACHIEVEMENT ANXIETY TEST. THEY WERE RANDOMLY ASSIGNED TO ONE OF TWO TREATMENT GROUPS TWO WEEKS LATER. SUBJECTS IN ONE GROUP WERE ADMINISTERED A STANDARDIZED MATHEMATICS ACHIEVEMENT TEST WITH THE ITEMS ORDERED FROM EASY TO DIFFICULT. SUBJECTS IN THE SECOND GROUP TOOK THE SAME WITH THE ORDER OF ITEMS REVERSED. A PHYSIOLOGICAL INDICANT OF STRESS, HEART-RATE, WAS MEASURED THREE TIMES DURING THE TEST USING A PULSEMETER. RESULTS CONFIRMED THE FINDING OF OTHER RESEARCHERS THAT THE MEAN NUMBER OF CORRECT ANSWERS FOR TEST QUESTIONS ARRANGED IN THE DIFFICULT-TO-EASY ORDER WERE SIGNIFICANTLY LOWER THAN THE MEAN NUMBER OF TEST QUESTIONS ARRANGED IN THE REVERSE ORDER. THIS STUDY GENERALIZES THE RESULT TO THE CONTENT DOMAIN OF MATHEMATICS. THIS STUDY PROVIDES TENTATIVE SUPPORT FOR THE HYPOTHESIS THAT ITEM ORDER HAS AN EFFECT ON THE STRESS GENERATED DURING A TEST. THIS POINT DESERVES TO BE RESEARCHED ADDITIONALLY TO ACHIEVE MORE CONCLUSIVE EVIDENCE THAN WAS OBTAINED IN THIS STUDY. LASTLY, THE DATA REVEALED NO INTERACTION BETWEEN ITEM ORDER AND LEVEL OF TEST ANXIETY. THIS PAPER WAS PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (CHICAGO, FEBRUARY 8-10, 1968). (AUTHOR)

THE EFFECTS OF ITEM ORDER AND ANXIETY ON TEST PERFORMANCE AND STRESS[1]

Ronald K. Hambleton

The Ontario Institute for Studies in Education

[1] The author acknowledges the assistance of Dr. Ross E. Traub in the preparation of this paper.

Paper presented at the annual meeting of the American Educational Research Association, February 8-10, 1968, in Chicago, Illinois.

THE EFFECTS OF ITEM ORDER AND ANXIETY ON TEST PERFORMANCE AND STRESS

Introduction

It is a generally accepted practice for test constructors to arrange the items in a test in order of increasing difficulty. The rationale behind this practice is quite simple—it increases the probability that an examinee will succeed on the early items and thereby gain confidence for the more difficult items later in the test. However tests are not always constructed in this way. For example, to reduce the chance of cheating, examiners sometimes make the order of presentation of items in a test different for different examinees. There is some evidence that the order of items in a test has an effect on performance.

MacNicol (1956) found that when items were ordered from difficult-to-easy, the mean number of correct answers on the test was significantly lower than the mean number of correct answers obtained when the items were ordered in one of two other ways: from easy to difficult and at random. There was no appreciable difference between average performance on the easy-to-difficult and the random orders. These results were obtained for a test administered under essentially power conditions.

One explanation of this phenomenon is suggested by Flaugher, Melton and Myers (1966). They found that when easy items appeared later in a test they were not reached by some subjects. In other words, if the test is speeded, it is clear that the difficult-to-easy order would disadvantage slow students since they would not have a chance to answer the easier items. This explanation is inadequate for MacNicol's results however since her test was administered under power conditions.

Another possible explanation has been offered by Mollenkopf (1950). He argued that fatigue and pressure to finish could account for poorer performance on items when they appeared later in the test than when they appeared earlier in the test.

Another and perhaps more interesting possibility is that personality characteristics of individual subjects hinder their performance of items in the difficult-to-easy order. One such personality characteristic which has been found to influence test performance is anxiety. This is a variable which has been studied extensively in test situations (I. G. Sarason, 1960; Ruebush, 1963) but apparently never in connection with item order.

Test anxiety is considered to be specific anxiety associated with test situations. It is measured by instruments such as the Alpert-Haber Achievement Anxiety Test (1960). This type of anxiety is generally found to be negatively related to test performance (Alpert and Haber, 1960; Carrier and Jewell, 1966; Grooms and Endler, 1960; Mandler and Sarason, 1952; I. G. Sarason, 1956b, 1957, 1959a, 1963). However the size of the correlation between test anxiety and test performance depends on the testing situation. Results of a study by I. G. Sarason (1961) found stronger negative correlations between test anxiety and aptitude test scores than between test anxiety and grade point averages. The aptitude scores were obtained from tests administered in large group sessions and were to support applications to college, a highly desired goal. On the other hand, grade point averages were based on classroom tests administered over the course of a semester and any one test would be unlikely to be perceived as important. Thus Sarason's evidence seems to suggest that the more important a test seems to the student, the greater the negative correlation between test anxiety and performance.

It should be observed at this point that anxiety is a word used in two ways: to refer to a personality trait and to refer to a transitory state. Studies by Cattell and Scheier (1958, 1961) suggest that anxiety questionnaires measure a relatively stable and permanent personality trait of the individual while physiological indicants of anxiety such as heart rate and palmar sweat measure a transitory state of the individual which fluctuates over time. This transitory state has been referred to in the literature as arousal or stress (Spielberger, 1966).

Further evidence that anxiety questionnaires measure a relatively permanent personality trait was provided by Smith (1965); he found that the characteristic level of questionnaire anxiety was unaffected by the stress conditions of the test administration.

The theory distinguishing trait and state anxiety holds that individuals with high anxiety scores as measured by a questionnaire are not anxious all the time; however, such individuals are more likely to emit anxiety responses than less anxious individuals in personally threatening situations such as tests. These anxiety responses interfere with task-relevant activities and lead to a subsequent reduction in performance level. Anxiety responses include heightened physiological activities (e.g. heart rate and sweating) and self-effacing statements (e.g. "I can't pass this test.").

Preliminary support for this theory came from results reported in learning experiments. I. G. Sarason (1956a, 1958) found that under certain instructional conditions, low anxious subjects were superior in performance to high anxious subjects, yet under different instructional conditions there were no differences.

Supporting evidence for the trait-state theory is found also in the literature of anxiety and test performance. Many studies suggest that individuals obtaining high scores on anxiety questionnaires differ from other individuals in the extent to which their performance is disrupted under conditions of stress (I. G. Sarason, 1957, 1959b). Typically the stress has been created by verbal instructions—e.g. informing the subject he is about to take an intelligence test. Wrightsman (1962) found that when a test was seen as important, the scores of anxious subjects were significantly lower than those of non-anxious subjects. When the test was seen as unimportant, anxiety was unrelated to performance.

Similar findings were reported by Sarason and Palola (1960). They found that under neutral or reassuring test instructions (informing the group that they were involved in a research project in which their function was to evaluate the test) high test anxious subjects did not differ from low test anxious subjects in performance. However when the test was administered under stressful conditions (informing the group that the test had been found to predict course grades, success in later life and even personality) the low test anxious subjects scored significantly higher.

These previous findings serve not only to support the trait-state theory of anxiety but they point out also the use of test directions to vary the stress of a test situation. Another test characteristic which could have an effect on the stress of the testing situation is the order of presentation of items.

Objectives

The review of the literature in the previous section indicated that item order, stress of the test situation, and test anxiety have an effect on test performance. This study was designed to investigate the relationships among these three variables.

The first objective of this study was to investigate under power testing conditions the effect of item order on test performance. It was expected that a difficult-to-easy order of items would prove to be more difficult than would the reverse order. This aspect of the study consists of an attempt to replicate the findings of MacNicol (1956) in a different content domain--mathematics.

In order to attain this first objective, a standardized mathematics achievement test was presented to a group of high school students. One group was given the test with the items ordered from easy-to-difficult while the second group was given the test with the items in reverse order. Performance scores for males and females were compared under the two arrangements.

The second objective of this study was to investigate the effect of item order on the stress induced in a test situation. The question asked was whether the stress of a test situation could be increased merely by changing the order of presentation of a set of test questions. It was hypothesized that a difficult-to-easy order of items would generate more anxiety responses and result in a higher level of stress for the subjects than would an easy-to-difficult order of items. A test of this hypothesis was made by measuring a physiological indicant of stress several times during the testing session. The groups working the items in different orders were compared on their average level of stress.

The third objective of this study was to investigate the interaction of item order and anxiety. The question asked was whether there was any difference in performance of the high and low test anxious subjects on the two arrangements of test items. On the assumption that a difficult-to-easy order of items leads to a more stressful test situation than the reverse order, it was expected that the difference in performance between the high and low test anxious subjects would be greater on the difficult-to-easy order than on the reverse order. Anxiety scores were obtained by administering a standardized anxiety questionnaire.

Method

Test Anxiety, Stress, and Achievement Measures

Test anxiety measure. The Achievement Anxiety Test (AAT) was used in this study to obtain a measure of test anxiety. This instrument was developed by Alpert and Haber (1960). It consists of two independent scales: a facilitating scale of nine items and a debilitating scale of ten items. Items on the facilitating scale are of the form--"Anxiety helps me to do better during exams and tests," while items on the debilitating scale are of the form--"Anxiety interferes with my performance during examinations and tests." Alpert and Haber (1960) state that the two scales have both undergone numerous revisions based on the results of item analyses, validity studies and theoretical reformulations. The test-retest reliabilities for each scale over a ten week period are reported to be about .85.

The two scales were combined into one questionnaire with the odd numbered items being from the debilitating scale and the even numbered items from the facilitating scale.

Stress measure. A physiological indicant of stress, heart rate, has been reported to be one of the best indicators of stress (Cattell and Scheier, 1961). In this study, heart rate was measured using a pulsemeter.

The pulsemeter (produced by Fraser Sweatman Incorporated, Pennsylvania) is transistorized and battery operated, and provides an instantaneous reading of the pulse. The pulsemeter has a range of

30 to 200 beats per minute with a needle indicator dial calibrated to provide quick and precise readings. The subject places a finger in a pressure sensitive device which measures pulse rate and displays it on the dial.

The ease with which the pulse rate can be measured makes the pulsemeter a useful piece of equipment. In this study it was possible to obtain the heart rate repeatedly during the administration of a test with a minimum of disruption to the subject.

Achievement measure. The achievement test used in this study was the Cooperative Mathematics Test Algebra II, Form B © 1962 ETS Princeton, N.J. Of the 40 items in the test, 30 were selected for use in this study. The chosen items covered topics taught in the Grade 11 Mathematics Course in Ontario.

The mathematics items were pretested on 250 students in two high schools in Toronto to obtain an index of difficulty for each item. Using these indices, two forms of the test were produced: Form I consisted of items arranged in order from easy to difficult; in Form II the items were arranged in the reverse order.

Subjects

The subjects were 106 eleventh grade mathematics summer school students from two secondary schools in Toronto, Ontario. This total represented 100% of the summer school enrollment in mathematics in the two schools.

The incomplete data of subjects who missed one of the two testing sessions was used wherever possible.

Procedure

Anxiety test administration. The subjects were told that the AAT was a questionnaire to find out how they felt about tests. They were assured that their answers would not be given to their teachers and that their scores would not be placed in their school record.

The AAT was administered by a researcher during a class period, two weeks prior to the administration of the achievement test. The students were unaware that their questionnaire results would be used in conjunction with the results of any other tests.

The instructions for administering the questionnaire were adapted from the original directions of Alpert and Haber (1960). Most students were able to complete the questionnaire in about 15 minutes; slower students were granted additional time in order that everyone would finish.

Achievement test administration. The 32-page mathematics test was printed in booklet form. Each page in the booklet consisted of a 3½ X 5½ inch sheet of paper. The first two pages in the booklet included space for student identification and general test directions. The remaining 30 pages contained the test questions. The subjects were given 40 minutes in which to complete the test.

To ensure maximum effect of item order on test performance it was considered essential that every subject work through the items in order. To ensure that this was done, three steps were taken: students were instructed to try the items in order; only one item appeared on each page of the test; and written instructions were given at the bottom of each page for the subject to go on to the next question.

Subjects were randomly assigned to one of two treatment groups. Subjects in one group were administered the standardized mathematics achievement test with the items ordered from easy-to-difficult (Form I). Subjects in the second group took the same test but with the order of items reversed (Form II). The subjects were not aware that there were two versions of the test.

To observe any effects item order might have on the stress of the test situation, it was considered essential that the subjects perceived the test as being important. To ensure that the subjects were properly motivated, the following statement was read immediately before the general directions for the achievement test: "This morning you are going to take a Grade 11 Achievement Test in Mathematics. Your results on this test will be sent to the school, so that your teacher can use this information in arriving at your final grade."

The administrators in this part of the study were different from the one used to administer the anxiety questionnaire to reduce the likelihood of an association being made between the two tests.

Stress measurement. At the outset of the experiment, even before the motivating instructions for the achievement test had been read, the subjects were told that they would undergo repeated pulse rate measurements before, during and after the test.

After the test instructions had been read and immediately before a subject started the test, his pulse rate was measured. Additional pulse readings were taken at the 10, 20, 30, and 40 minute marks of the test. The last reading was made as the test was taken from the student. At the same time the pulse reading was taken, the item the subject had reached on the test was recorded.

Scoring the materials. The answer sheet for the AAT provided five choices for each question. These choices, of which the subject was to choose one, were "Never", "Sometimes", "About half the time", "Frequently", and "Always". For the purpose of scoring, these choices were given numerical values of 0, 1, 2, 3, and 4 respectively. The score for each item was totalled with scores for other items on the same scale to arrive at two scores—a debilitating anxiety score and a facilitating anxiety score.

In the Mathematics Achievement Test, the subjects answered the questions by circling one of the five choices available. Subjects responses coded 0, 1, 2, 3 or 4 were key-punched on to IBM cards and scored on an IBM 7094 computer. Since the directions told students to guess, no correction for guessing was employed.

Statistical Derivation of Stress Scores

It was reported in a previous section that five pulse readings were taken on each subject: before he started the test, and then at the 10, 20, 30, and 40 minute marks of the test. The last reading was made as the test was taken from the student.

The stress score for each subject was found by averaging the second, third, and fourth pulse readings (the during-test readings). This was done in all cases except when the subject had finished the test before his third or fourth reading was taken. In the case where the subject finished the test before the third reading was taken, his stress score was simply his pulse reading at the 10 minute mark. In the more frequent case where a subject finished the test after the third and before the fourth reading, his stress score was the average of his second and third pulse readings. This procedure of arriving at stress scores was adopted in an attempt to remove the effect of extraneous factors affecting stress scores for those who finished the test early.

Results

Order of Test Items

For each form of the test a plot was made of the item difficulty level versus the item position in the test. The results revealed that except for a few misplacements, the majority of items on each form were in the desired order. The rank correlation between item position in the test and the position the item should have been in, based on the item difficulty level as estimated from the data of this study, was .71 for the easy-to-difficult order and .52 for the difficult-to-easy order.

Effect of Item Order on Test Performance

Out of 30 questions, the subjects who took the easy-to-difficult form of the test (N = 55) averaged 11.41 correct answers. The subjects taking the difficult-to-easy form of the test (N = 51) averaged 9.96 correct answers.

The test for the significance of differences in test performance under the two item orders was carried out using a two factor analysis of variance design (Lindquist, 1953, pp. 127-132). The two factors were item order and sex. Sex was introduced as a factor in the design to make the test for the significance of the item-order effect more sensitive. Because the assignment of students to test forms was done at random, and because there were more boys than girls in the sample, the number of observations per cell of the analysis of variance table varied from 19 to 32. A modification of the analysis of variance procedure was used to take into account the unequal numbers (Winer, 1962, pp. 222-224).

The analysis of variance is summarized in Table 1. The main effect due to item order was significant at the .05 level. The main effect due to sex did not reach the conventional level of significance although it approached significance ($.05 < p < .10$). There was no interaction between item order and sex.

The conclusion that may be drawn on the basis of this analysis is that scores on the easy-to-difficult order were on the average significantly higher than scores on the difficult-to-easy order and this difference was independent of the sex of the examinee.

A check was made to see whether the difficult-to-easy form of the test was more speeded than the easy-to-difficult form since if it was, it would be possible to explain the observed difference in performance, at least in part, by the failure of subjects working the difficult-to-easy test form to attempt the easy items appearing later in the test. An analysis of subjects' responses revealed that 8 out of 55 subjects did not complete the easy-to-difficult form, whereas 13 out of 51 subjects did not finish the difficult-to-easy form. (The difference in the number of subjects completing each form was not statistically significant when tested by a χ^2 test for contingency : $\chi^2 = 2.00$, d.f. = 1, $p > .15$.) A more detailed analysis of the data for the subjects who did not complete the test revealed the following: on the easy-to-difficult form the 8 subjects who failed to finish did not attempt a total of 38 items; on the difficult-to-easy form the 13 subjects who did not finish left a total of 71 untried items. (An untried item

TABLE 1

ANALYSIS OF VARIANCE TABLE FOR
MATHEMATICS TEST SCORES
(Item Order X Sex)

Source of Variance	d.f.	MS	F Ratio
Item Order	1	53.53	4.06*
Sex	1	51.40	3.90
Interaction	1	15.12	1.15
Error	102	13.17	

Note.--The number of subjects in each cell of the design varied from 19 to 32.

*Significant at the .05 level.

was defined as an item not reached by a subject and is indicated by the fact that the subject had not reached the page of the test on which the item appeared. Because the items appeared in the test, one per page, it was possible to determine quite accurately which items were not reached. Although almost twice as many items were not reached on the difficult-to-easy order as on the reverse order, it is unlikely that this difference affected the obtained results. The argument that may be advanced in support of this assertion is the following: the average difficulty of the items not reached on the easy-to-difficult order (as estimated from data on the group who finished the easy-to-difficult form) was .25. Had the 8 subjects performed to the average level of the group finishing the form, the average score on the easy-to-difficult order would have been increased from 11.41 to 11.59. On the difficult-to-easy form the average difficulty of the items not reached (as estimated from data on the group who finished the difficult-to-easy form) was .50. If the 13 subjects had performed to the average level of the group finishing the form, the average score on the difficult-to-easy order would have increased from 9.96 to 10.66. Thus the differences in performance on the two forms would be reduced from 1.45 correct answers to .93 correct answers. It is likely however that the 21 subjects who failed to finish the test were the poorer mathematics students than the ones who finished; hence their chances of performing as well as on the unfinished items as those subjects who did finish the test was remote. Thus the difference in performance on the two forms would probably have been very much closer to 1.45 correct answers than to .93 correct answers had sufficient time been allowed for all subjects to complete the test. Thus the speededness of the difficult-to-easy test form does not appear to be a plausible explanation of the obtained difference in average performance between the easy-to-difficult and difficult-to-easy forms.

Effect of Item Order on Stress Scores

The means and standard deviations of stress scores and pre-test stress scores under the two item orders are summarized in Table 2. The group administered the easy-to-difficult order (N = 24) had an average stress score of 75.20 whereas the group administered the reverse order (N = 24) had an average stress score of 76.48. (The number of subjects in this part of the study was reduced from 106 to 48 because of a malfunction in the pulsemeter during the testing of students in one of the two schools. When using 106 subjects with a difference of 2 between the stress score means, the power of the F test for rejecting the hypothesis of no differences between the means was .70. However because only 48 subjects were used in the analysis, for the same difference in stress score means, the power of the F test dropped to .45. In order to maintain the power at approximately .70, the .10 significance level was adopted for testing the difference between stress scores under the two forms.)

It is clear from Table 2 that prior to the administration of the test there were differences in the average pre-test stress scores of the experimental groups. For this reason the difference in the stress scores of the experimental groups was tested using the method of analysis of covariance with the pre-test stress score used as the

TABLE 2

MEANS AND STANDARD DEVIATIONS OF STRESS SCORES

Test	N	Stress Scores		Pre-Test Stress Scores	
		Mean	S.D.	Mean	S.D.
Easy-to-Difficult Order	24	75.20	6.70	80.29	9.33
Difficult-to-Easy Order	24	76.48	5.00	77.71	8.14

covariate. The procedure followed was that outlined by Gulliksen and Wilks (1950).

The results of the analysis of covariance are summarized in Table 3 and may be explained in the following way.

The analysis of covariance method of Gulliksen and Wilks (1950) tests three statistical hypothesis.[1] The first is a homogeneity of variance hypothesis. It states that the variance of stress scores about the regression line of stress scores on pre-test stress scores is the same for the two experimental groups. In Table 3, the statistic that tests this hypothesis, χ^2 (H_1) , is less than the critical value at the .10 level, thus the observed data do not contradict the homogeneity of variance hypothesis.

The second statistical hypothesis was then tested. Hypothesis two is that the slope of the regression of stress scores on pre-test stress scores is the same for each experimental group. In effect, hypothesis two asserts that there is no interaction between the effect of item order and the level of pre-test stress scores. In Table 3, the statistic that tests this hypothesis is F (H_2) . Since the observed value F (H_2) is less than the critical value at the .10 level, the data fail to contradict the hypothesis of equal regression slopes.

The analysis of covariance was completed by testing the third statistical hypothesis. It states that the intercept of the regression of stress scores on pre-test stress scores is the same for each

[1] A prior assumption underlying the use of the analysis of covariance techniques is that in each experimental group, the regression of criterion on predictor scores is linear. The results of a linearity of regression test failed to reject the hypothesis that the regression of stress scores on pre-test stress scores was linear.

TABLE 3

ANALYSIS OF COVARIANCE

Hypothesis	d.f.	Significance Test
1. Homogeneity of variance	1	X^2 $(H_1) = 2.12$
2. Equal slopes of regression	1, 44	F $(H_2) = .09$
3. Equal intercepts of regression	1, 45	F $(H_3) = 2.38$*

*$.05 < p < .10$

experimental group. This hypothesis asserts that there is no effect of item order on stress scores. In Table 3, the statistic that tests this hypothesis is $F\ (H_3)$. The probability of the occurence by chance of an F value greater than the one observed using a 1-tailed test[1] is approximately .06. Thus the hypothesis that the regression lines had equal intercepts can be rejected at the .10 level of significance. The adjusted stress scores mean on the easy-to-difficult item order was 74.71. For the reverse order, the adjusted stress score mean was 76.97.

The very tentative conclusion is drawn here that the difficult-to-easy item order produced a more stressful test situation than the reverse item order. This conclusion is stated tentatively because of the failure to observe a conventional level of significance. What is obviously required is additional research to establish the conclusion more firmly.

Interaction of Item Order and Test Anxiety

The Achievement Anxiety Test gives both a facilitating and a debilitating anxiety score. Since this part of the study was concerned with the negative effects of test anxiety on test performance under varying degrees of stress, only the latter score was used in the analysis.

On the basis of the debilitating anxiety scores the 100 subjects were divided into two groups, high test anxious (HTA, N = 25) and low

[1] Since the purpose of the analysis of covariance was to determine whether the hypothesized directional difference between stress scores of the experimental groups is supported by the data, a one-tailed test of significance was appropriate (Jones, 1954).

test anxious (LTA, N = 25). The HTA and LTA groups included the upper and lower 25% of the sample.

The results for HTA and LTA subjects on the two forms of the test were somewhat surprising. The HTA group taking the difficult-to-easy form of the test (N = 12) averaged 10.33 correct answers on the test, whereas the LTA group on the same form (N = 12) averaged only 9.08 correct answers. On the easy-to-difficult form of the test, the HTA group (N = 13) averaged 10.00 correct answers. The LTA group on the same form (N = 13) averaged 11.77 correct answers.

The test for the significance of differences in test performance was carried out using a two factor analysis of variance design. In this analysis, the two factors were item order and anxiety. Of special interest in the analysis was the interaction between the two factors. The analysis of variance is summarized in Table 4. The main effects due to item order and anxiety were not significant. The interaction effect also failed to attain the .05 level of significance. The conclusion must be that the data of this study provides no evidence to support the hypothesis that the difference in performance between high and low test anxious subjects would in general be greater on the difficult-to-easy order than the reverse order.

TABLE 4

ANALYSIS OF VARIANCE TABLE FOR MATHEMATICS TEST SCORES (Item Order X Anxiety)

Source of Variance	d.f.	MS	F Ratio
Item Order	1	17.27	1.92
Anxiety	1	.84	.09
Interaction	1	28.44	3.16
Error	46	9.00	

Note.—The number of subjects in each cell of the design was either 12 or 13.

Discussion

This study provides additional clear-cut support for the contention that item order has an effect on test performance. This study generalizes the conclusion of previous research to the content domain of mathematics. It was found that scores on the easy-to-difficult item order were on the average significantly higher than scores on the difficult-to-easy order.

In view of the failure to find an interaction between item order and test anxiety it seems clear that the personality characteristic of test anxiety cannot be used to explain the difference in performance on the two item orders. However it is possible to speculate that the concept of 'response sets' in testing provides an explanation.

Cronbach (1950) stated that when a person takes a test, he brings to the test a number of test-taking habits or response sets which affect his score. Response sets such as the tendency to work for speed rather than accuracy and the tendency to guess when uncertain are well known for their effect on test scores. Although the expectancy that any achievement test will begin with easy items was not conceived as a 'response set' by Cronbach (1946), it is possible to regard this explectancy as such. Moreover because it is common to order items from easy to difficult, a set to expect items to be ordered in that way may be present in grade 11 students. When a subject with such an expectancy encounters difficult items early in a test, he expects even more difficult items later on which makes him more anxious with the likely result that test performance is adversely affected. This explanation gains support from the second part of the study in which it was found that a difficult-to-easy order of test items produced a more stressful test situation for subjects than the reverse order of test items.

It was suggested by Cronbach (1946) that if a particular 'response set' exists in a test situation, the test directions can be revised to reduce the effect of the set. Further research is needed to determine if manipulation of test directions would reduce the observed decrement in test performance produced by administering items in the order difficult to easy.

The tentative conclusion that stress scores were higher for subjects on the difficult-to-easy order than the reverse order provides some indication of the importance of item order on the stress generated during a test. Clearly this point deserves to be researched additionally to achieve more conclusive evidence than was obtained in the present study.

Finally, the importance of this study to test constructors seems to be the evidence it provides for the discontinuation of the practice of making the order of presentation of items in a test different for different examinees to reduce the chance of cheating. It is clear on the basis of currently available evidence that reordering the items of a test in effect produces a test with different properties than the original. Hence it may be impossible to make valid comparisons of the scores obtained by students who take the same test items in a different order. This conclusion is in agreement with the conclusion reached by Flaugher, Melton and Myers (1966).

Summary

The objectives of this research were to investigate the effect of item order on the performance of a mathematics test; on the amount of stress generated during a test; and on the performance of high and low test anxious subjects.

106 high school students completed the Achievement Anxiety Test. Two weeks later, they were randomly assigned to one of two treatment groups. Subjects in one group were administered a standardized mathematics achievement test with the items ordered from easy to difficult. Subjects in the second group took the same test with the order of items reversed. A physiological indicant of stress, heart-rate, was measured three times during the test using a pulsemeter. The three heart-rate measures for each subject were averaged to obtain a stress score.

Results of this study confirmed the finding of other researchers that the mean number of correct answers for test questions arranged in the difficult-to-easy order were significantly lower than the mean number of test questions arranged in the reverse order. This study generalizes the previous result to the content domain of mathematics. In addition, this study provides tentative support for the hypothesis that item order has an effect on the stress generated during a test. This point deserves to be researched additionally to achieve more conclusive evidence than was obtained in this study. Lastly, the data failed to support the hypothesis of an interaction between item order and level of test anxiety.

References

Alpert, R., and Haber, R. N. Anxiety in academic achievement situations. J. abnorm. soc. Psychol., 1960, 61, 207-215.

Carrier, N. A., and Jewell, D. O. Efficiency in measuring the effect of anxiety upon academic performance. J. educ. Psychol., 1966, 57, 23-26.

Cattell, R. B., and Scheier, I. H. The nature of anxiety: a review of thirteen multivariate analyses comprising 814 variables. Psychol. Rep., 1958, 4, 351-388.

Cattell, R. B., and Scheier, I. H. The meaning and measurement of neuroticism and anxiety. New York: Ronald Press, 1961.

Cronbach, L. J. Response sets and test validity. Educ. psychol. Measmt., 1946, 6, 475-494.

Cronbach, L. J. Further evidence on response sets and test design. Educ. psychol. Measmt., 1950, 10, 3-31.

Flaugher, R. L., Melton, R. S., and Myers, C. T. A study of the effects of item rearrangement. Paper presented at the 74th Annual Convention of the American Psychological Association, 1966, New York City.

Grooms, R. R., and Endler, N. S. The effect of anxiety on academic achievement. J. educ. Psychol., 1960, 51, 299-304.

Gulliksen, H., and Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Jones, L. V. A rejoiner on one-tailed tests. Psychol. Bull., 1954, 51, 585-586.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

MacNicol, Katharine. Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, Educational Testing Service, Princeton, N.J., 1956.

Mandler, G., and Sarason, S. B. A study of anxiety and learning. J. Abnorm. soc. Psychol., 1952, 47, 166-173.

Mollenkopf, W. G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika, 1950, 15, 291-315.

Ruebush, B. K. "Anxiety." Child Psychology. Sixty-Second Yearbook, Part 1, National Society for the Study of Education. Chicago: University of Chicago Press, 1963. Chapter II, 460-516.

Sarason, I. G. Effect of anxiety, motivational instructions, and failure on serial learning. J. exp. Psychol., 1956, 51, 253-260. (a)

Sarason, I. G. The relation of anxiety and "lack of defensiveness" to intellectual performance. J. consult. Psychol., 1956, 20, 220-222. (b)

Sarason, I. G. Test anxiety, general anxiety, and intellectual performance. J. consult. Psychol., 1957, 21, 485-490.

Sarason, I. G. Effects on verbal learning of anxiety, reassurance, and meaningfulness of material. J. exp. Psychol., 1958, 56, 472-477.

Sarason, I. G. Intellectual and personality correlates of test anxiety. J. abnorm. soc. Psychol., 1959, 59, 272-275. (a)

Sarason, I. G. Relationships of measures of anxiety and experimental instructions to word association test performance. J. abnorm. soc. Psychol., 1959, 59, 37-42. (b)

Sarason, I. G. Empirical findings and theoretical problems in the use of anxiety scales. Psychol. Bull., 1960, 57, 403-415.

Sarason, I. G. Test anxiety and the intellectual performance of college students. J. educ. Psychol., 1961, 52, 201-206.

Sarason, I. G. Test anxiety and intellectual performance. J. abnorm. soc. Psychol., 1963, 66, 73-75.

Sarason, I. G., and Palola, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. J. exp. Psychol., 1960, 59, 185-191.

Smith, C. P. The influence on test anxiety scores of stressful versus neutral conditions of test administration. Educ. psychol. Measmt., 1965, 25, 135-141.

Spielberger, C. D. Anxiety and behavior. New York: Academic Press, 1966.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Wrightsman, L. S. The effects of anxiety, achievement motivation, and task importance upon performance on an intelligence test. J. educ. Psychol., 1962, 53, 150-156.