NAMI - 1010

EFFECTS OF PERCEIVED SCORING FORMULA ON SOME ASPECTS OF TEST PERFORMANCE

Lawrence K. Waters

June 1967

NAVAL AEROSPACE MEDICAL INSTITUTE
NAVAL AEROSPACE MEDICAL CENTER
PENSACOLA, FLORIDA

EFFECTS OF PERCEIVED SCORING FORMULA ON SOME ASPECTS OF TEST PERFORMANCE

Lawrence K. Waters

Bureau of Medicine and Surgery
MF022.01.02-5001.53

Approved by

Ashton Graybiel, M. D.
Director of Research

Released by

Captain J. W. Weaver, MC, USN
Commanding Officer

22 June 1967

NAVAL AEROSPACE MEDICAL INSTITUTE
NAVAL AEROSPACE MEDICAL CENTER
PENSACOLA, FLORIDA 32512

SUMMARY PAGE

THE PROBLEM

The purpose of this study was to examine the effects on test performance of six different scoring instructions for an objective type test.

FINDINGS

Increases in the penalty for wrong responses were accompanied by consistent increases in the mean number of omitted items, but the mean number correct remained fairly stable over the various penalties. In general, intertest correlations were largest when all items were attempted and lowest when random responses were substituted for omitted items. The scoring formula appropriate to the structure of the items, $(R - \frac{W}{4})$ was generally superior to the scoring formula appropriate to the penalty that examinees were told would be used in scoring the test.

- - - - - - - - - - -

The author is presently at Ohio University, Athens, Ohio.

INTRODUCTION

The instructions given to examinees who are about to take an objective type test may encourage them to guess if they are not sure of an answer, may direct them not to guess, or set up scoring penalties intended to enforce the desired behaivor. For example, examinees may be told not to guess - that the number of wrong answers, or some multiple thereof, will be subtracted from the number of right answers that they obtain. While the test performance of examinees under instructions designed to encourage or discourage guessing has been studied (3-5), there is little empirical information about the differential effects on test-taking behavior of informing the examinees that one or another penalty formula is to be used in the scoring. The purpose of the present study was to examine the effects on test performance of six different scoring instructions.

PROCEDURE

The responses of 123 flight students in the indoctrination week of pre-flight training were used to obtain P-values on a pool of vocabulary items. Each item consisted of a stem word and five alternative words, from which respondents indicated the one most nearly opposite in meaning to the stem word.

From this item pool, two 50-item tests (Forms A and B) were constructed by matching the P-values ($\pm$ .02) of items. The distribution of item difficulties ranged from .03 to .98. Both tests were labeled "Word Knowledge Test."

Form A was administered without time limit to 420 flight students in the early part of the indoctrination week of pre-flight training. Instructions for making responses to the items were printed on the test booklet, but no information was given regarding how the test was to be scored.

Form B was administered to the same sample of flight students near the end of their indoctrination week. The general instructions for responding to the items were the same as for Form A. However, one of six different sets of scoring instructions was printed on each Form B test booklet, and examinees were told that omitted items would not count either way. The six sets of instructions were as follows:

1. Scoring method not discussed.
2. Right answer (R) one point, no deduction for wrong answers.
3. $(R - \frac{W}{4})$ (number of right answers minus one-quarter the number of wrong answers).
4. $(R - w)$.
5. $(R - 2w)$.
6. $(R - 4w)$.

After the examinees had finished Form B, all regular marking pencils were collected, and red pencils were distributed. The examinees were then told to attempt to answer all of the items they had omitted. This procedure provided a means of obtaining performance data on items that were originally attempted and omitted.

Each of the six scoring instructions was used on approximately one sixth of each class tested. This resulted in six groups (one for each scoring instruction), ranging from 61 to 79 cases.

ANALYSIS AND RESULTS

Table I shows mean scores for Form A (R_A), the initially attempted items on Form B (R_B), and all the items on Form B (R_{BT}). Comparison of R_A and R_{BT} shows a consistent difference of about one item in the mean numbers of right answers; this indicates that Form B was probably a little more difficult than Form A. The differences

among groups on R_A and on R_{BT} are not statistically significant. On R_B, however, the numbers of right answers for all groups whom were told that wrong answers would be penalized differed significantly ($P < .05$) from both R_A and R_{BT} means. Further, it was established by using Duncan's Multiple Range Test (1) that mean R_B scores for groups with unspecified or zero weights for wrongs differed significantly from mean R_B scores for all groups with specified weights greater than zero. Mean number of rights for groups with specified weights greater than zero did not differ significantly from each other. The proportion of right responses (rights/attempts) increased consistently, but not significantly, with increases in scoring weights.

Table I

Mean Rights for Six Groups on Two Forms of a Vocabulary Test

Type of Score	Groups, With Weight for Wrongs					
	1 (Unspec)	2 (0)	3 (-1/4)	4 (-1)	5 (-2)	6 (-4)
R_A (Form A)	26.79	26.23	26.37	26.45	25.95	27.55
R_B (Form B)	25.20	24.97	22.15	20.46	20.46	20.84
R_{BT} (Form B)	25.59	25.33	25.37	25.26	24.96	26.22
P (R_B/attempts)	.53	.51	.59	.62	.64	.68

Table II compares the mean numbers of answers omitted on Form A (O_A) with those omitted on Form B (O_B), and the differences between these as the penalty for wrong responses is increased. The groups were roughly comparable in terms of number of

omitted items on Form A ($F < 1.0$), but differed substantially on Form B as a function of the specific scoring sets ($F = 34.73$, $n_1 = 5$, and $n_2 = 414$). For each increment in the specified scoring weight for wrong responses, there was a significant ($P < .05$) increase in the mean number of omits, except for the difference between the unspecified and the zero conditions. Apparently, examinees regarded no specified scoring instructions in much the same manner as they did zero weights for wrongs.

Table II

Mean Omits for Six Groups on Two Forms of a Vocabulary Test

Type of Score	Groups, With Weight for Wrongs					
	1 (Unspec)	2 (0)	3 (-1/4)	4 (-1)	5 (-2)	6 (-4)
O_A (Form A)	1.23	1.33	1.44	1.32	1.37	0.43
O_B (Form B)	2.05	1.23	12.31	16.88	18.56	19.36
$O_B - O_A$	0.82	-0.10	10.87	15.56	17.19	18.93

For groups with scoring weights greater than zero, the more able examinees (higher R_A scores) tended to omit fewer items on Form B (r's ranged from -.33 to -.42), and to get somewhat more of the originally omitted items correct when forced to answer (r's ranged from .18 to .26). There was no consistent pattern of relationships with variations in scoring weights. For either unspecified or zero weight groups, essentially no relationship was found between R_A scores and number right of the originally omitted Form B items (r's = .01 and .09). Form A rights and Form B omits correlated -.24 and -.01, respectively, for these groups.

Examinee test performance in terms of omitted items was affected in a consistent manner by variations in the penalty they were told would be applied to wrong responses. The more able examinees, as might have been expected, omitted fewer items but still got more of the originally omitted items correct when forced to answer.

Table III gives the correlations between R_A scores and 1) rights on attempted Form B items (R_B), 2) total rights on Form B (R_{BT}), and 3) total rights on Form B when random responses were assigned to omitted items (R_{BR}). The random responses were assigned by use of a table of random numbers. After the responses had been assigned to each omitted item, these were scored and the number right added to the R_B score.

Table III

Intertest Correlations for Six Groups with Three Types of Rights Scores on Form B

Correlation Between Form A Rights and Given Form B Scores	Groups, With Weight for Wrongs					
	1 (unspec)	2 (0)	3 (-1/4)	4 (-1)	5 (-2)	6 (-4)
$r_{R_A R_B}$*	813#	779	816	751	811	676
$r_{R_A R_{BT}}$	804	785	840	756	828	723
$r_{RA R_{BR}}$	807	790	804	743	790	563

*See text for definition of three Form B rights scores.
#Decimal points omitted.

The differences in intertest correlations involving the three Form B rights scores were generally very small, in large part due to the fact that R_B was a large component of both R_{BT} and R_{BR}. However, it was of interest to compare rights scores under the various scoring sets with rights scores when examinees had been forced to answer every item under both actual "informed" guessing and simulated "random" guessing conditions.

The two groups without any specified penalty for guessing, who omitted less than 4 per cent of the Form B items, showed no consistently different relationships among the three sets of Form B rights scores. However, the four groups with a specified penalty for wrongs, who omitted from 25 per cent to 40 per cent of the Form B items, demonstrated higher intertest correlations for R_{BT} than for R_B scores. Also, in terms of the intertest correlations, informed guessing was superior to "random" guessing. In this study, forcing examinees to answer every item led to higher intertest coefficients, but for examinees who guessed at random it would have been better not to force them to respond to every item.

The first three rows of Table IV present the intertest correlations between R_A and three types of Form B scores: R_B, the Form B score obtained by using scoring formula appropriate to the specified category, and the "best" Form B score. It is obvious that the scoring formula appropriate to the set given examinees on Form B yielded progressively smaller coefficients compared to rights only, as the penalty for wrongs given examinees increased. Examinees did not adapt their performance on Form B in relation to the scoring set given them. It can be seen that $R - \frac{W}{4}$ was the best of the scoring formulas used for all groups with specific scoring instructions, except for the most extreme group. In general, the scoring formula appropriate to the structure of the test

was better than the scoring formula appropriate to the scoring set. While the number of items attempted decreased as the weights for wrongs increased, examinees with the more severe penalties did not omit enough items. Whether examinees overestimated their probability of success on several items or did not perceive the severity of the penalty cannot be determined from these data.

Table IV

Intertest Correlations for Six Groups with Rights Only and Formula Scoring

Intertest Correlation Between R_A and Given Score on Form B	Groups, With Weight for Wrongs					
	1 (Unspec)	2 (0)	3 (-1/4)	4 (-1)	5 (-2)	6 (-4)
Rights only	813*	779	816	751	811	676
Appropriate scoring formula		779	817	703	656	482
Best Form B score	813	783	817	779	847	763
Formula for best Form B score	R	$R - \frac{W}{4}$	$R - \frac{W}{4}$	$R - \frac{W}{4}$	$R - \frac{W}{4}$	R - w
Optimal weights for wrongs#	-03	-43	-31	-29	-36	-63

*Decimal points omitted.
#Optimal weights for wrongs when rights are weighted one (2).

REFERENCES

1. Duncan, D. B., Multiple range and multiple F tests. Biometrics, 11:1-142, 1955.

2. Guilford, J. P., Psychometric Methods. New York: McGraw-Hill, 1954.

3. Keislar, E. R., Test instructions and scoring method in true-false tests. J. exp. Educ., 21: 243-249, 1953.

4. Jackson, R. A., Guessing and test performance. Ed. Psychol. Meas., 15:74-79, 1955.

5. Swineford, F. and Miller, P. M., Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. J. ed. Psychol., 44-129-139, 1953.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Naval Aerospace Medical Institute
Pensacola, Florida 32512

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

N/A

3 REPORT TITLE

EFFECTS OF PERCEIVED SCORING FORMULA ON SOME ASPECTS OF TEST PERFORMANCE

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

N/A

5 AUTHOR(S) (First name, middle initial, last name)

Lawrence K. Waters

6 REPORT DATE	7a. TOTAL NO OF PAGES	7b. NO OF REFS
22 June 1967	9	5

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO: MF022.01.02-5001	NAMI - 1010
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report): 53
d.	

10 DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
Author's present address: Ohio University Athens, Ohio	

13 ABSTRACT

This study examined the effects on test performance of systematic variations in the scoring formulas which examinees were told would be used in scoring their tests. Two equivalent 50-item vocabulary tests were constructed. Form A was administered without scoring formula specified to 420 pre-flight students. Three days later Form B was administered to the same groups subdivided six ways, i.e., same instructions as Form A; zero weight for wrongs; 1/4, 1, 2, or 4 points off for wrongs. Increases in the penalty for wrong responses were accompanied by consistent increases in the mean number of omitted items, but the mean number correct remained fairly stable over the various penalties. In general, intertest correlations were largest when all items were attempted and smallest when random responses were substituted for omitted items. The scoring formula appropriate to the structure of the items, $(R - \frac{W}{4})$, was generally superior to the scoring formula appropriate to the penalty that examinees were told would be used in scoring the test.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Testing						
Personnel						
Selection						