DOCUMENT RESUME

ED 050 168 TM 000 563

AUTHOR Hendrickson, Gerry F.
TITLE The Effect of Differential Option Weighting on Multiple-Choice Objective Tests.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J.
REPORT NO R-93
BUREAU NO BR-61610-0321
PUB DATE Jan 71
GRANT OEG-2-7-061610-0207
NOTE 53p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Answer Keys, Correlation, Factor Structure, *Guessing (Tests), High School Students, Mathematical Models, *Multiple Choice Tests, Objective Tests, *Scores, Scoring, Sex Differences, Statistical Analysis, Test Bias, *Testing Problems, Test Reliability, Test Validity, *Weighted Scores

ABSTRACT

The purpose of this study was to determine whether option weighting improved the internal consistency and intercorrelation of the subtests. The differential option-weighting scheme employed in this study is based on one devised by Guttman. The tests were first scored with Guttman-type weights and then with conventional correction-for-guessing weights. The internal-consistency of the tests increased markedly when Guttman-type weights were used. The correlation of the two verbal subtests increased somewhat when Guttman weights were used, but the correlation of the two mathematics subtests as well as the intercorrelation of all verbal and mathematics subtests decreased. Differences in the factor structure of the Guttman-weighted and the conventionally weighted subtests were used to explain the result. (Author)

THE JOHNS HOPKINS UNIVERSITY

REPORT No. 93

CENTER FOR THE STUDY OF SOCIAL ORGANIZATION OF SCHOOLS

THE EFFECT OF DIFFERENTIAL OPTION WEIGHTING ON MULTIPLE-CHOICE OBJECTIVE TESTS

BY

GERRY F. HENDRICKSON

JANUARY 1971

STAFF

John L. Holland, Director

James M. McPartland, Assistant Director

Virginia Bailey
Thelma Baldwin
Zahava D. Blum
Judith P. Clark
James S. Coleman
Robert L. Crain
David DeVries
Keith Edwards
Doris R. Entwisle
Gail Fennessey
James Fennessey
Catherine J. Garvey
Ellen Greenberger
Rubie Harris
Edward J. Harsch
Robert T. Hogan
John H. Hollifield
Michael Inbar
Nancy L. Karweit
Judith Kennedy
Steven Kidder
Hao-Mei Kuo
Samuel Livingston
Edward L. McDill
Rebecca J. Muraro
Jeanne O'Connor
Suzanne K. Pieper
Meredith A. Prell
Martha O. Roseman
Peter H. Rossi
Leslie Schnuelle
Aage B. Sørensen
Julian C. Stanley
Keith F. Taylor
Mary C. Viernstein
Diana F. Ward
Murray A. Webster
Barbara J. Williams
Phyllis K. Wilson

THE EFFECT OF DIFFERENTIAL OPTION WEIGHTING ON MULTIPLE-CHOICE OBJECTIVE TESTS

GRANT NO. OEG-2-7-061610-0207

BR-61610-0321

Gerry F. Hendrickson

REPORT NO. 93

JANUARY, 1971

This research was supported in part by the College Entrance Examination Board and the Educational Testing Service. Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education, and no official endorsement by the Office of Education should be inferred.

The Johns Hopkins University

Baltimore, Maryland

Acknowledgement

This paper is based on part of a thesis submitted to The Johns Hopkins University in fulfillment of a requirement for the Ph.D. degree. I extend my thanks to Professor Julian C. Stanley, my advisor, for his helpful advice and criticism of the thesis and to Professor Bert F. Green, Jr. for his critical reading of this manuscript.

The Effect of Differential Option Weighting on

Multiple-Choice Objective Tests

Abstract

The purpose of this study was to determine whether option weighting improved the internal consistency and intercorrelation of the subtests. The differential option-weighting scheme employed in this study is based on one devised by Guttman. The tests were first scored with Guttman-type weights and then with conventional correction-for-guessing weights. The internal-consistency of the tests increased markedly when Guttman-type weights were used. The correlation of the two verbal subtests increased somewhat when Guttman weights were used, but the correlation of the two mathematics subtests as well as the intercorrelation of all verbal and mathematics subtests decreased. Differences in the factor structure of the Guttman-weighted and the conventionally weighted subtests were used to explain the result.

Table of Contents

Tables and Figures

The Effect of Differential Option Weighting on Multiple-Choice Objective Tests

Many tests constructed by teachers for use in their own classrooms and virtually all commercially published tests follow a multiple-choice format. These tests are often scored with the familiar "correction-for-guessing" formula, whereby the score for a particular individual is

$$\text{Score} = \text{Right} - \frac{\text{Wrong}}{\text{options} - 1} \quad .$$

Thus, for a five-option test, an examinee receives 1 point if he marks the keyed option, 0 points if he marks nothing, and -1/4 point if he marks any incorrect option (these are called "distracters"). The examinee's total score on the test is simply the algebraic sum of the points he receives on each item. In the model which underlies the derivation of this formula, it is assumed that if the examinee does not know the correct answer, he guesses randomly among the options or omits the item altogether.

Most people would agree, however, that students rarely guess among the options in a strictly random manner. If an examinee is not sure of an answer, he will usually make an educated guess. The more knowledge an examinee possesses regarding the question, the more informed his guess will be, and the greater his probability of marking the correct option. The examinee in this case is said to have partial knowledge of an answer.

On the other hand, sometimes an examinee may feel fairly cer-

tain that an incorrect option is actually the correct one. In this case, an incorrect option was chosen because of misinformation. Thus, misinformation decoys the examinee into marking an incorrect option, whereas partial information increases the probability that he will choose the correct one.

Therefore, information about the examinee's ability is revealed by the alternative he chooses, even if that alternative is wrong. This information is lost, however, if all distracters receive the same weight. A weighting system which rewards the choice of plausible distracters and penalizes heavily the choice of implausible ones might be desirable. An empirical weighting technique proposed by Guttman (1941) may accomplish this goal. Each alternative is weighted proportionally to the total score of the examinees who select it. Plausible distracters are usually chosen by high-scoring examinees; these distracters, therefore, receive high weights. Grossly incorrect distracters, on the other hand, are usually chosen by low-scoring examinees; these distracters receive low weights.

Discussion of Guttman's Weighting Scheme and its Relationship to Others in the Literature

Since Guttman's procedure, or estimates of it, are sometimes used without making the connection explicit, it is appropriate to discuss Guttman's approach and variations of it in some detail. Guttman developed his technique for scaling the response categories (i.e., all the options in all the items) in multiple-choice tests for

which there are no a priori correct answers to the items, and thus no clear-cut way of knowing how the categories (i.e., options) should be weighted. Interest or attitude instruments are examples of such tests. Guttman (1941) proposed that the "best" weights be those which maximize the internal consistency of the test. He shows that this problem can be approached from three directions.

First, one can derive a weight for each option in each item such that the weights for options selected by a particular person be as similar as possible among themselves and that these weights, in turn, be as dissimilar as possible from weights of options selected by other people. This aim can be accomplished by maximizing the ratio of variance among people to total variance (i.e., the correlation ratio for weights). Guttman (1941, p. 346) reports that the considerations which gave rise to his correlation ratio for weights were the same as those employed by Horst (1936) and Edgerton and Kolbe (1936) for deriving weights for quantitative variables; the same considerations led Wilks (1938) to his minimum generalized variance solution.

Secondly, one can derive a score for each person such that all persons choosing a particular option have scores as much alike as possible and that these scores, in turn, be as different as possible from the scores of people choosing other options. This aim can be accomplished by maximizing the ratio of category variance to total variance (i.e., the correlation ratio for scores).

Thirdly, one can simultaneously derive a set of weights, one for each category, and a set of scores, one for each person, such that people with similar scores tend to choose categories with similar weights. This aim can be accomplished by maximizing the correlation coefficient between the weight and score associated with each category. (e.g., If there are N individuals in n categories, there are Nn such pairs; the correlation of weights and scores is across these Nn pairs.)

Guttman shows that the square of this correlation coefficient is equal to each of the two squared correlation ratios and that, therefore, maximization of each of these three quantities yields the same solution.

The solution can be expressed in the form of a principal components analysis (Hotelling, 1933). The matrix to be factored is of order $\underline{n}$ x $\underline{n}$, where $\underline{n}$ is the number of possible response categories (e.g., if each of 40 items has 5 options, $\underline{n}$ = 40 x 5). The general element of this matrix is a "certain chi-squared product-moment" (Guttman, 1941, p. 332). Lord (1958, p. 291) has shown that "Guttman's principal components for the weighting system are effectively the same as a certain set of item weights obtained by factoring the matrix of item intercorrelations." Lord (1958) has also shown that Guttman's principal components for the weighting system are the same as the set of weights that will maximize coefficient alpha (Cronbach, 1951).

The solution of any of the three values that Guttman set up to be maximized yields weights for a particular category that are linearly related to the average score on the total test of the people who chose the category in question. (See Guttman, 1941, p. 344, for the exact equation.) This equation is essentially that used in a scaling technique, known as the Method of Reciprocal Averages, which appeared fairly early in the literature (cf. Richardson and Kuder, 1933, and Horst, 1935). However, the full complexity of the underlying mathematical model was not reported until Guttman's (1941) article. For this reason the weighting technique will be attributed to Guttman in this paper.

Guttman's procedure for calculating weights is quite tedious if done without the aid of a modern computer; however, short-cut procedures have

been developed to estimate these weights. Guttman's weights for an option can be estimated by a correlation coefficient between the criterion total score and the dichotomy of marking or not marking the option in question. Estimates of these correlation coefficients can, in turn, be read from a table that is entered with the percent of examinees in the highest and lowest 27% of the criterion-score distribution who mark the given choice. Guttman (1941, p. 341), however, criticizes such procedures. (See Davis, 1959, for a comparison of the estimated weights to those calculated using Guttman's method.)

Option weights estimated in this way have been used in two studies, one by Davis and Fifer (1959) and another by Sabers and White (1969). These two studies differ in the criterion each uses. For Davis and Fifer (1959) the criterion was the total score distribution on a parallel form of the test; their aim was to improve the parallel-forms reliability of the test. For Sabers and White (1969) the criterion variable was an achievement test; their aim was to improve validity. Comparable results would be expected from these two studies, with the exception that in the former study, improvement would be expected in parallel-forms reliability, whereas in the latter study, improvement would be expected in predictive validity. However, although Davis and Fifer (1959) were able to raise the cross-validated parallel-forms reliability of a 45-item test from .68 to .76 without lowering its validity, Sabers and White (1969) were not able to raise either validity or reliability by more than .03. This discrepancy is due at least partly to the fact that in the latter study the cross-validation groups were poorly matched (Sabers and White, 1969, p. 95).

The methodology in these two studies is weak in two respects; the optimum weights were estimated from the upper and lower 27% of the criterion score distribution rather than calculated directly using the entire distribution, and the groups on which these weights were determined were quite small. Davis (1959) demonstrated that the latter point is especially crucial in his discussion of the reliability of the weights. With today's large computers, these methodological weaknesses can be avoided. In the present study the weights were calculated by Guttman's method on quite large samples (2500 each).

The purpose of the present study is to compare the effect of Guttman weighting with the effect of correction-for-guessing weighting. In this study the criterion for the option weights in a particular test is the score distribution on the test itself; thus, the major goal of this study is to improve the internal consistency of a certain multiple-choice objective test by differential option weighting. Another aim of this study was to determine whether the cross-correlation (this term will be explained later) of these tests improved as a result of Guttman weighting.

Design of the Study

Description of the Weighting Scheme

The essence of the weighting procedure suggested by Guttman (1941) is that categories be keyed so that they maximally predict an internal criterion. In this study the categories are the options for each item;

the criterion for each option is the mean standardized total score on the remaining items of the test for all examinees who selected the option in question. The weights were determined by an iterative procedure. Initially the options were weighted with correction-for-guessing weights. The scores for all examinees were calculated from these weights, and new weights were calculated from these scores. However, changing the weight also changes the total scores; therefore, another set of weights can be calculated. These iterations were continued until the internal consistency was sufficiently high and stable. In this study three iterations were deemed adequate, after five were tried. See Appendix A for a detailed description of the way in which the weights were calculated. The weight for "omit" was calculated in exactly the same manner as the weights for the other five options; "omit" will be treated as another option in this paper.

Description of the Test

The test used was form QSA43 of the Scholastic Aptitude Test (SAT), which had been administered to 296,640 examinees (most of them high-school juniors and seniors) at the College Entrance Examination Board's regular testing in November of 1968. The verbal section of the SAT contains a 40-item subtest and a 50-item subtest; these are administered and timed separately. Both tests contain sentence completion items, analogies, antonyms, and reading-comprehension items; however, the proportion of the various types of items is different in the two subtests. The mathematics section of the SAT also contains two subtests, which are administered and timed separately. The first subtest consists of 17

general mathematical problems and 18 data-sufficiency items,[1] while the second subtest consists only of 25 general mathematical problems.

Table 1 shows the composition of form QSA43 of the SAT. Note that only Subtests 1, 2, 4, and 5 were used in this study. Subtest 3 is only used for equating and pretesting purposes and is not part of the scored portion of the SAT. For this reason it was not used in this study.

Description of the Sample

The responses of 5000 men and 5000 women (each group was selected randomly from the large group of examinees retained for item-analysis purposes) to each item of form QSA43 of the SAT were obtained from the Educational Testing Service (ETS). The 5000 examinees of each sex were further divided into two randomized-block groups of 2500 examinees each by blocking on total verbal scores. Blocking in this way makes it extremely likely that the total verbal score distributions of the two groups are approximately the same and therefore that the verbal mean and standard deviation of one group will be almost exactly the same as the verbal mean and standard deviation of the other group.

Weights were calculated separately in each of the four subtests for each of the four groups of 2500 examinees. Therefore, a set of weights was calculated in each of two independent groups in each subtest. Table 2 illustrates the way in which the group of 5000 men were divided. The scores of the 5000 women were divided in an identical manner. The analysis was then conducted separately for each sex.

[1] In a data-sufficiency item, an examinee is presented with a question and facts A and B, pertaining to the question. He may respond in one of five ways, by saying (a) that A alone is sufficient to answer the question, (b) that B alone is sufficient to answer the question, (c) that both A and B together are sufficient, but neither alone is sufficient, (d) that either A or B alone is sufficient, or (e) that A and B together are not sufficient.

Table 1

Composition of the Scored Portion of

form QSA43 of the SAT

Section	Subtest*	Time	Item nos.	Item Types
Verbal	1	30 min	1 - 10	Sentence Completions
			11 - 20	Antonyms
			21 - 30	Analogies
			31 - 35	Reading Comprehension
			36 - 40	Reading Comprehension
Verbal	2	45 min	1 - 5	Reading Comprehension
			6 - 10	Reading Comprehension
			11 - 18	Sentence Completions
			19 - 26	Antonyms
			27 - 35	Analogies
			36 - 40	Reading Comprehension
			41 - 45	Reading Comprehension
			46 - 50	Reading Comprehension
Mathematics	4	45 min	1 - 17	General Problems
			18 - 35	Data-Sufficiency
Mathematics	5	30 min	1 - 25	General Problems

*Subtest 3 was used for equating purposes only. Since it is not part of the scored portion of the SAT, it was not used in this analysis.

Table 2

How the Group of 5000 Men Was Divided

(The 5000 women were divided in the same way.)

RESPONSES OF 5000 MEN							
Randomly divided into two groups							
GROUP 1 RESPONSES OF 2500 MEN to the				GROUP 2 RESPONSES OF 2500 MEN to the			
verbal section		mathematics section		verbal section		mathematics section	
subtest 1	subtest 2	subtest 4	subtest 5	subtest 1	subtest 2	subtest 4	subtest 5

Weights were calculated for each sex in a given subtest for double cross-validation purposes, i.e., the weights from one group of 2500 examinees were applied to the other group of 2500 and vice versa. These groups of 2500 examinees will be referred to as "cross-validation groups." Weights calculated in one group and used in the other will be called "cross-validated weights." All comparisons in this study were carried out using cross-validated weights in order to avoid capitalizing on the idiosyncrasies of the group from which weights were calculated. (See Mosier, 1951, for a discussion of cross-validation.)

Description of the Investigations

The main focus of these investigations was to determine whether the internal-consistency reliability of the four subtests in the SAT improved when Guttman weights were used. The effect of Guttman weighting on the intercorrelation of these four subtests (these will be called "cross-correlation coefficients") was also investigated. Certain of these cross-correlation coefficients might be thought of as quasi-parallel-forms reliability and others as quasi-validity. In the following paragraphs the writer explains what quasi-parallel-forms reliability and quasi-validity coefficients are, and why the qualifier "quasi" must be used.

As already noted, the verbal section of the SAT consists of two subtests. These two subtests contain the same types of items (sentence-completion, antonyms, analogies, and reading-comprehension), but the total

number of items and the number of items of each type is different in the two tests. See Table 1, where the chief difference in item type is seen to be in the proportion of reading-comprehension items--25% in Subtest 1, versus 50% in Subtest 2. Therefore, these two tests can be considered comparable, but only approximately so. The two subtests in the mathematics section can likewise be considered comparable, although less so than the verbal subtests, because more than half of the items in Subtest 4 are data-sufficiency, whereas none in Subtest 5 are. Correlation of the two verbal or two mathematical subtests will therefore be termed "quasi-parallel-forms reliability" because the subtests are not truly parallel with respect to content.

Elementary measurement textbooks usually state that "validity refers to the extent to which the test measures what we actually wish to measure" (see Thorndike and Hagen, 1969, p. 62.) If we wish to measure "general ability" in both the verbal subtests and the mathematics subtest, then intercorrelating these two independent measures of general ability might be thought of as a type of validity (quasi-validity) coefficient, albeit a very poor one.

Four separate investigations were carried out. The first two were designed to compare the effect of using cross-validated Guttman weights with the effect of using correction-for-guessing weights. These comparisons were made in two groups of men and two groups of women on all four subtests of the SAT. In the first study internal-consistency reliability coefficients were compared; in the second study intercorrelations of the subtests were compared. In the third investigation the regression and correlation of Guttman scores and formula scores was determined. In the fourth study the differences in weights for men versus women were exam-

ined to see if the two sexes responded differently.

Effect of Option Weighting on the Internal-Consistency Reliability in Each Subtest

Experimental Procedure

The subjects in this experiment were all 5000 men and 5000 women; all four subtests of the SAT were used. A stratified form of Hoyt's (1941) internal-consistency coefficient was used to calculate the reliability of each subtest (see Rajaratnam, Cronbach, and Gleser, 1965, for a discussion of stratified-parallel tests.) In this study the item types form the "strata" of the subtests. Each subtest of the SAT except subtest 5 contains more than one type of item. For example, both subtests in the verbal section contain four types of items: sentence completion, antonyms, analogies, and reading comprehension. Therefore, there are four "strata" in each subtest in the verbal section. The stratified internal-consistency reliability of each subtest was calculated for each group in each sex.

Experimental Results and Discussion

Table 3 shows the improvement in internal-consistency that came

Table 3

Reliability of the Verbal Section when Responses Are Weighted Conventionally, Compared with Reliability when Responses Are Weighted with Guttman Weights

RELIABILITY	Subtest 1 (40 items)				Subtest 2 (50 items)			
	Men		Women		Men		Women	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
using correction-for-guessing weights	.8502	.8457	.8428	.8423	.8678	.8711	.8530	.8542
using crossvalidated Guttman weights	.8992	.8963	.8823	.8825	.9191	.9214	.8927	.8955
Equivalent to an increase in test length of	57.18%	57.70%	39.82%	40.62%	73.37%	76.46%	43.38%	46.27%

about when Guttman weights were used to weight differentially the options in two subtests in the verbal section. Table 4 shows the same results for the two subtests in the mathematics section. The first row of Table 3 and Table 4 shows the internal-consistency coefficients obtained when correction-for-guessing weights were used; the second row shows the internal-consistency coefficients obtained when Guttman weights from an independent group were applied to the responses of the group in question. Row 3 shows the percent by which a conventionally scored test would have to be lengthened in order to achieve the gain in reliability that resulted from the use of Guttman weights. The effective increase in test length varied quite a bit from subtest to subtest and between sex groups. It ranged from a high of 78.25% to a low of 19.09%. The average effective increase in test length was 49%.

This is a dramatic increase. To choose an especially striking example of this increase, the 50-item subtest 2 scored with Guttman weights is more internally consistent for the men than the entire 90-item verbal section would be if it were scored using correction-for-guessing weights. Furthermore, the internal consistency was increased without increasing the number of test items or increasing test-taking time. It seems as though Guttman scoring can enable a short test (e.g., 50 verbal items) to be as internally consistent as a longer test (e.g., 90 verbal items). Thus, using Guttman weights could save testing time. Donlon (1963) described some desirable additions to the SAT which cannot presently be incorporated

Table 4

Reliability of the Mathematics Section when Responses Are Weighted Conventionally, Compared with Reliability when Responses Are Weighted with Guttman Weights

RELIABILITY	Subtest 4 (35 items)				Subtest 5 (25 items)			
	Men		Women		Men		Women	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
using correction-for-guessing weights	.8422	.8330	.8045	.8091	.8156	.8132	.7818	.7830
using crossvalidated Guttman weights	.9016	.8989	.8676	.8698	.8412	.8383	.8164	.8119
Equivalent to an increase in test length of	71.68%	78.25%	59.24%	57.62%	19.77%	19.09%	24.11%	19.62%

because the time limit proves to be a major constraint.

Note in Tables 3 and 4 that using Guttman weights, the reliability for men was increased more than for women in subtests 1, 2, and 4. This result is somewhat surprising. It seems that options discriminate more sharply among men than among women in this study.

Another somewhat puzzling result concerns the fact that subtest 2 was effectively lengthened more than subtest 1, and subtest 4 was effectively lengthened more than subtest 5. Subtests 2 and 4 are the longer tests in the verbal and mathematics sections, respectively. The greater length could be responsible for greater reliability, because the longer the test, the more reliable the criterion (i.e., test scores on the $\underline{I} - \underline{1}$ items); and the more reliable the criterion, the less the shrinkage after Guttman weighting. A 49-item test, for example, is likely to be considerably more reliable than a 39-item test. Also, the fewer the number of items in the criterion, the greater the change in reliability of the criterion as one goes from one ($\underline{I} - \underline{1}$) set to another in weighting items.

However, other factors besides length alone could be responsible for causing subtests 2 and 4 to be effectively lengthened more than their counterparts. It is also possible that a difference in the content of the subtests is responsible. The difference in the average effective length for each of the two subtests in the mathematics section is particularly striking. Subtest 4 was effectively lengthened by an average of 67%, whereas subtest 5 was effectively lengthened by an average of only 21%. Although both of these subtests are in the mathematics section, they differ in that subtest 4 consists of both data-sufficiency items and general mathematical problems, while subtest 5 consists of only general

mathematical problems. An hypothesis put forth very tentatively to explain the result is that data-sufficiency items afford more opportunity for making educated guesses than do general mathematical items, and therefore, they allow an examinee to use his partial knowledge about the question.

The success of this weighting scheme depends on the correctness of the assumptions that the quality of the distracters varies considerable and that groups of similar ability tend to endorse the same distracter. It is easy to see that the quality of the distracters varies systematically in data-sufficiency items and that this variation affords the examinee a chance to use his partial knowledge. For example, if the correct response is that both pieces of information are sufficient to answer the question, then the examinee is more correct if he says that one but not the other is sufficient than if he says that neither are sufficient to answer the question.

On the other hand, it is not as easy to see the difference in the quality of distracters in general mathematical problems. Perhaps the assumptions are not correct for general mathematical problems; that is, one cannot say that one algebraic mistake is "more correct" than another or that an algebraic error is "less wrong" than an arithmetic one.

In the verbal section, on the other hand, the difference in the average effective length for each of the two subtests is less striking than it was in the mathematics section. The average effective increase for subtest 1 was 49%; the average effective increase for subtest 2 was 60%. These two tests differ somewhat in content also. The main content difference is in the percent of reading comprehension items that each subtest contains--25% for subtest 1 verses 50% for subtest 2. The higher reliability

of Guttman scores in Subtest 2 could be due to the fact that the alternatives in the reading comprehension items are differentially attractive to examinees of varying ability levels.

Indeed, it seems reasonable to suppose that reading comprehension items allow the examinee more opportunity to make an educated guess than do antonym, analogy, or sentence completion items. For example, if an examinee has no idea what "archipelago" means, he has no basis for selecting this word's antonym from the five alternatives. In verbal omnibus items of this sort, it is likely that one alternative is clearly right and the others clearly wrong.

Often this clear-cut distinction between the right alternative and the wrong ones does not exist in reading comprehension items. Rather, one alternative is "best" in some sense. The examinee is forced to read and weigh carefully all the alternatives before deciding which one is best. For example, examinees are sometimes asked to pick which of the five alternatives best states the main theme of the reading passage.
Often, all the topics mentioned in the five alternatives were discussed in the passage, although only one alternative states the central theme. It seems reasonable to assume that the more thoroughly an examinee understands the passage, the more likely he is to recognize the alternative in which the central theme is stated. Intermediate degrees of understanding will enable the examinee to eliminate certain alternatives, and a high degree of understanding will enable him to choose wisely among the alternatives remaining. Thus, it seems likely that the alternatives in reading-comprehension items are able to differentiate between examinees of varying ability levels better than the alternatives in the analogy, antonym, or

sentence completion items. If that is the case, then when Guttman weights are used, the reliability in the verbal section would indeed be greater in the subtest having the more reading comprehension items.

Effect of Option Weighting on Cross-Correlational Reliability and Validity

Experimental Procedure

The subjects in this study were the same 5000 men and 5000 women as were used in the earlier part of the investigation; again, all four subtests of the SAT were used. The product-moment correlation coefficients of the total scores on each of the four subtests with all other subtests were obtained, first using correction-for-guessing weights and then using cross-validated Guttman weights. Six intercorrelations are possible with four subtests; they are r_{12} , r_{45} , r_{14} , r_{15} , r_{24} , and r_{25} . The quasi-parallel-forms reliability coefficients are r_{12} (verbal) and r_{45} (mathematics). The other four r's are quasi-validity coefficients.

Experimental Results and Discussion

Table 5 shows the intercorrelations of the subtests for each group in both sexes. The results obtained when correction-for-guessing weights

Table 5

Comparison of the Intercorrelation of Scores from the Four Subtests

CROSS-CORRELATIONAL RELIABILITY	Men		Women	
	Group 1	Group 2	Group 1	Group 2
r_{12}				
using correction-for-guessing weights	.8491	.8409	.8340	.8316
using cross-validated Guttman weights	.8660	.8587	.8475	.8476
Equivalent to an increase in test length of	14.85%	14.98%	10.61%	12.62%
r_{45}				
using correction-for-guessing weights	.7994	.8132	.7881	.7903
using cross-validated Guttman weights	.7562	.7551	.7463	.7729
Equivalent to a decrease in test length of	22.17%	29.17%	20.91%	9.69%
CROSS-CORRELATIONAL VALIDITY				
r_{14}				
using correction-for-guessing weights	.6386	.6265	.6113	.6188
using cross-validated Guttman weights	.5981	.5871	.5896	.6181
Equivalent to a decrease in test length of	15.78%	15.23%	8.65%	.30%
r_{15}				
using correction-for-guessing weights	.6204	.5947	.5845	.5939
using cross-validated Guttman weights	.5966	.5710	.5748	.5798
Equivalent to a decrease in test length of	9.51%	9.29%	3.90%	5.65%
r_{24}				
using correction-for-guessing weights	.6559	.6566	.6284	.6354
using cross-validated Guttman weights	.6003	.6087	.6035	.6322
Equivalent to a decrease in test length of	21.21%	18.64%	9.99%	1.37%
r_{25}				
using correction-for-guessing weights	.6369	.6227	.6152	.5989
using cross-validated Guttman weights	.5952	.5863	.5834	.5870
Equivalent to a decrease in test length of	17.15%	14.13%	12.41%	4.81%

were used are shown in the first row for each correlation coefficient, and the results obtained when Guttman weights were used are shown in the second row. The third row shows the percent by which the length of a conventionally scored test would have to be changed in order to produce the change in reliability that occurred when cross-validated Guttman weights were used. Note that scoring subtests 1 and 2 with cross-validated Guttman weights produced a gain in the correlation coefficient but that in all other cases the use of Guttman weights produced a decrease in correlation. Note also that using cross-validated Guttman weights to score all subtests caused a decrease in quasi-reliability in the mathematics test (r_{45}) greater in magnitude than the increase in the quasi-reliability in the verbal test (r_{12}) in three of the four cases.

Using cross-validated Guttman weights resulted in an average increase in quasi-reliability in the verbal section equivalent to that which would be expected if the conventionally scored test had been lengthened by 13.3%. In the mathematics section the average effective decrease in test length was 20.5%. The corresponding average decrease in the quasi-validity coefficients, r_{14}, r_{15}, r_{24}, and r_{25}, was equivalent to an effective decrease in test length of 10.0%, 7.1%, 12.8%, and 12.1%, respectively. In every case the effective test length was changed less for women than for men.

These results were somewhat surprising. It was assumed that the increase in internal-consistency reliability would be accompanied by an increase in quasi-validity, as indicated by the usual formula relating reliability and validity:

$$\rho_{t_n c_n} = \rho_{tc}\sqrt{\frac{\rho_{t_n t_n}\ \rho_{c_n c_n}}{\rho_{tt}\ \rho_{cc}}} \quad , \qquad (1)$$

where t is a test,

t_n is a test $\underline{n}$ times as long as test $\underline{t}$,

c is a criterion measure,

and c_n is a criterion measure $\underline{n}$ times as long as $\underline{c}$. See Cronbach (1970, p. 171.) The reliability of the longer test, $\rho_{t_n t_n}$, should be greater than the reliability of the shorter test, ρ_{tt} . Likewise, the reliability of the longer criterion measure, $\rho_{c_n c_n}$, should be greater than the reliability of the shorter criterion measure, ρ_{cc} . Therefore, the correlation of the more reliable test with the more reliable criterion should be greater than the correlation of the less reliable test with the less reliable criterion.

A close look at the derivation of this formula reveals the cause for this reasoning being at least partially erroneous in <u>this</u> case. The derivation begins with the following two correction-for-attenuation formulas:

$$\rho_{T_{t_n} T_{c_n}} = \frac{\rho_{t_n c_n}}{\sqrt{\rho_{t_n t_n}}\sqrt{\rho_{c_n c_n}}} \quad , \text{ and } \rho_{T_t T_c} = \frac{\rho_{tc}}{\sqrt{\rho_{tt}}\sqrt{\rho_{cc}}} \quad .$$

If the longer test and longer criterion were lengthened by adding more items of the same type, then the correlation of the true score of the longer test with the true score of the longer criterion ($\rho_{T_{t_n} T_{c_n}}$) should be equal to the correlation of the true score of the shorter test with the true score of the shorter criterion ($\rho_{T_t T_c}$). Equating the values of $\rho_{T_{t_n} T_{c_n}}$ and $\rho_{T_t T_c}$ and rearranging the terms yields formula (1), the desired

relationship. (The reliability coefficients under the radicals in the above formulas must be suitable estimates of the one-form correlation between true and obtained scores, e.g., $\sqrt{\rho_{tt}} = \sqrt{\rho_{T_t t}}$. See Stanley, 1971.)

However, increasing the reliability of a test by weighting options differentially may not be equivalent to increasing the reliability of a test by adding more items of the same type. If it is not, then $\rho_{T_t T_n}$ does not equal $\rho_{T_t T_c}$, and formula (1) should not be used. Therefore, making a test more internally consistent does not necessarily make the test more valid.

In fact, in this study, there seemed to be an inverse relationship between increased internal consistency and increased cross-correlational validity. That is, the groups for which the use of Guttman weights caused the greatest increase in internal-consistency reliability were often the ones for which there was the greatest decrease in cross-correlation. Take the groups in subtest 2 and subtest 4, for example. Tables 3 and 4 show that for both of these subtests the internal consistency increased more for the men than for the women. However, Table 5 shows that the correlation of subtest 2 and subtest 4 was decreased more for men than for women as a result of using Guttman weights. It seems from these data that an increase in internal-consistency reliability has an adverse effect on the particular type of cross-correlation studied (i.e., verbal with mathematical).

A look at the relationship of internal-consistency reliability and validity shows why this might be the case. In order for a test to be valid, the items must be reliable but somewhat heterogeneous. If the internal consistency of a test is increased without adding more items, then

the items of the test must have become more homogeneous. The fact that the items do indeed become more homogeneous can be proven by demonstrating that Hoyt's (1941) internal-consistency coefficient and coefficient alpha (Cronbach, 1951) (these two internal-consistency coefficients are algebraically equivalent) are equal to the intraclass coefficient among items, stepped up by the number of items in the test (Stanley, 1957 and 1971).

For men, the homogenity of Subtests 2 and 4 increased the most as a result of Guttman weighting; however, the cross-correlation, r_{24}, decreased more than any other verbal-mathematical correlation when Guttman weights were used. This result seems to suggest that the more homogeneous a test is made, the more poorly it correlates with something quite different.

The fact that the items in a subtest were more homogeneous if the options were weighted with cross-validated Guttman weights than if they were weighted with correction-for-guessing weights implies that the factor structure of the test is different in the two weighting methods. It may be that tests weighted by the former method consist of fewer factors than tests weighted by the latter method. In other words, perhaps low item intercorrelation in a test means that these items are measuring several aspects of a particular ability (e.g., verbal or mathematical) and that high item intercorrelation means that these items are measuring fewer aspects of this ability. Perhaps also in the case of subtests consisting of more than one item type, a particular item type dominates the subtest as a result of weighting. For example, Subtest 2 might be dominated by reading-comprehension items and Subtest 4 by data-sufficiency items. A comparison of the factor structure of the subtests of the SAT weighted with Guttman weights with the subtests weighted with correction-for-

guessing weights is now in progress.

If the factor structure of weighted and unweighted subtests were known, perhaps something could be said about the correlation of these weighted subtests and college grade point average (or some other ordinary validity coefficient). However, on the basis of these results no prediction about that kind of validity can be made. The verbal and mathematics tests measure quite different abilities. It is likely that r_{12} is more like the correlation of the verbal section of the SAT with grade point average than r_{14} , r_{15} , r_{24} , or r_{25} are. Thus, failure to raise these latter quantities does not necessarily mean failure to raise the more usual validity coefficient.

Regression and Correlation of Guttman Scores and Formula Scores

Experimental Procedure

This trial analysis was performed on the 2500 men in Group 1; the test used was Subtest 1 (40 items) in the verbal section. Subtest 1 was chosen because all verbal item types are represented equally, i.e., there are ten sentence-completion items, ten antonyms, ten analogies, and ten reading-comprehension items. The significance of curvilinearity was tested by the analysis of variance technique outlined in McNemar (1969, pp. 306-317). This test was made first for the regression of Guttman scores on

formula scores and then for the regression of formula scores on Guttman scores. The distribution of Guttman scores and the distribution of formula scores were standardized and then transformed so that each distribution had a mean of 20.0 and a standard deviation of 6.0. (In both cases formula scores were the scores obtained when options were weighted with correction-for-guessing weights, and Guttman scores were those obtained when options chosen by this group were weighted with Guttman weights calculated from the other male group.)

Experimental Results and Discussion

The scatter diagram for this analysis is plotted in Figure 1. The values of the formula scores are shown along the horizontal axis; the values of the Guttman scores are shown along the vertical axis. The correlation of Guttman and formula scores computed from the scatter diagram was .9059.

Table 6 shows the analysis of variance results for the regression of Guttman (Y) on formula (X) scores ($Y = 1.88 + .91X$). The correlation ratio, η_{yx}, computed from the scatterplot in Figure 1 was .9110. The last three rows of Table 6 show the three F ratios. F_1 tests the significance of the correlation ratio; F_2 tests the significance of linear correlation; and F_3 tests the significance of curvilinearity. F_1 and F_2 were

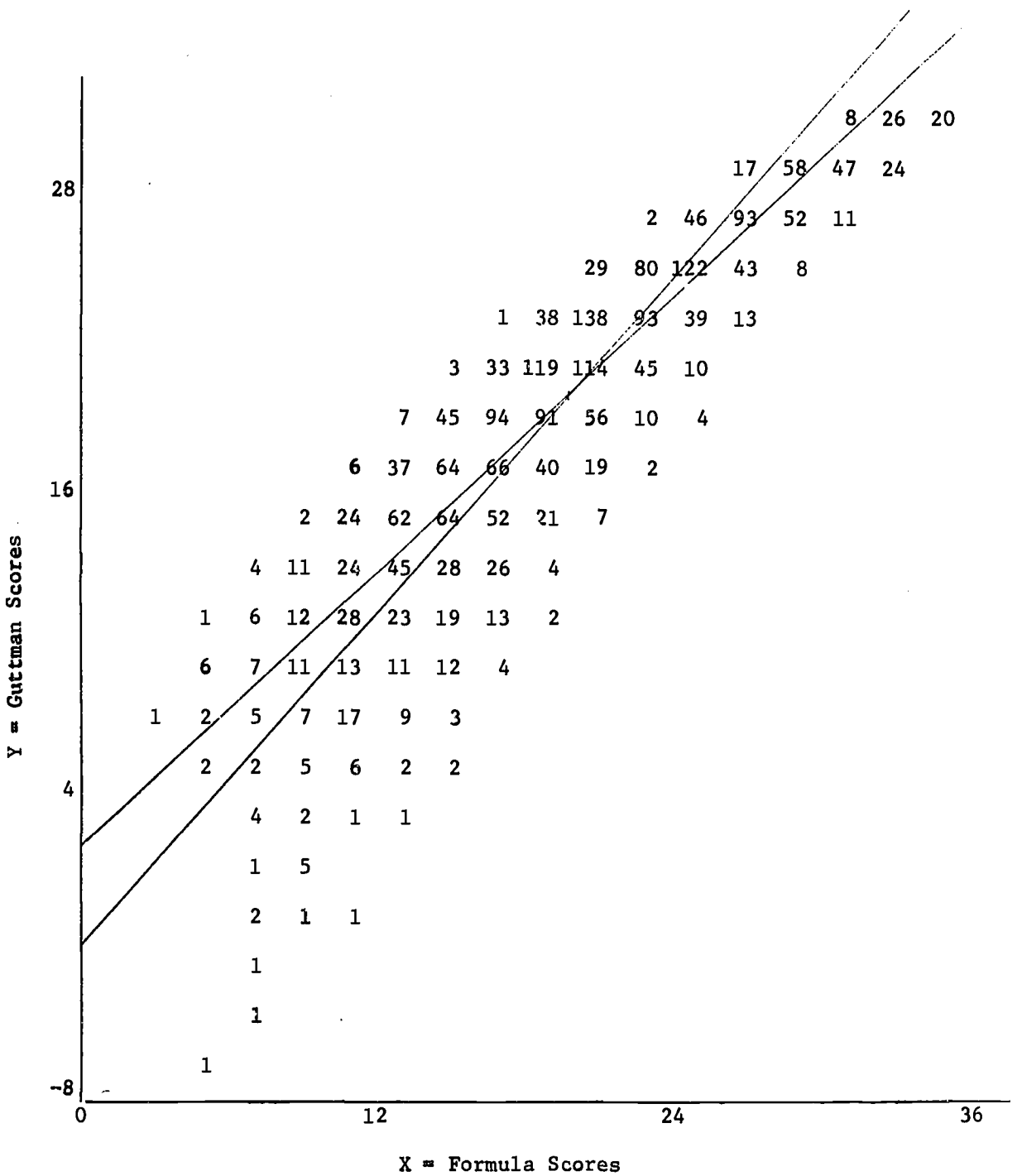

Figure 1. Bivariate Scatterplot for Formula and Guttman Scores of 2500 Men to Subtest 1. (r = .9059)

Table 6

ANOVA Table for the Regression of 2500 Guttman (Y) on Formula (X) Scores for Data in Figure 1

Source of Variation	Sum of Squares	df	Mean Squares
Linear Regression	73714	1	73714.0 = s_p^2
Deviation of Means from Line	838	15	55.9 = s_d^2
Between-array Means	74553	16	4659.6 = s_b^2
Within Arrays	15275	2483	6.2 = s_w^2
Residual from Line	16114	2498	6.5 = s_r^2
Total (corrected)	89828*	2499	

*Since the sources of variation are not independent, they do not add together to form the total sum of squares.

Significance of Correlation Ratio: $F_1 = s_b^2/s_w^2 = 757.4$; $p << .001$

Significance of Linear Correlation: $F_2 = s_p^2/s_r^2 = 11430.0$; $p << .001$

Significance of Curvilinearity: $F_3 = s_d^2/s_w^2 = 9.1$; $p < .001$

highly significant ($p << .001$). F_3 was also significant ($p < .001$); however, inspection of the formula from which F_3 was calculated (see McNemar, 1969, p. 314) reveals that if $\underline{N}$ is very large, a small difference between η^2 and r^2 will cause F_3 to be large. In this case $\eta^2_{yx} - r^2$ was .0094, demonstrating that non-linear variance accounted for only 0.94% of the total variance. Therefore, although F_3 was significant, because of high power for this statistical test, the significance is not of practical importance.

Table 7 shows the analysis of variance results for the regression of formula (X) on Guttman (Y) scores ($X = 1.88 + .91Y$). The correlation ratio, η_{xy}, computed from the scatterplot in Figure 1 was .9273. The last three rows of Table 7 show the three F ratios. As before the correlation ratio was highly significant ($p << .001$). Although F_3 indicated a significant amount of curvilinearity ($p < .001$), because of the high power for this test, the significance was not important. The difference between η^2_{xy} and r^2 was .0392, demonstrating that non-linear variance accounted for only 3.92% of the total variance.

These results reveal that the formula and Guttman score distributions are related linearly, for the most part. In both regressions (Y on X and X on Y) about 82% of the total variance was accounted for by the straight line. Furthermore, not much of the remaining 18% non-linear variance was due to curvilinearity (.94% in one case and 3.92% in the other).

One point deserves to be mentioned at this time. Wilks (1938, p. 27)

Table 7

ANOVA Table for the Regression of Formula (X) on Guttman (Y) Scores for the Data in Figure 1

Source of Variation	Sum of Squares	df	Mean Squares
Linear Regression	74148	1	74148.0 = s_p^2
Deviation of Means from Line	3540	18	196.7 = s_d^2
Between-array Means	77688	19	4088.8 = s_b^2
Within Arrays	12668	2480	5.1 = s_w^2
Residual from Line	16208	2498	6.5 = s_r^2
Total (corrected)	90356*	2499	

*Since the sources of variation are not independent, they do not add together to form the total sum of squares.

Significance of Correlation Ratio: $F_1 = s_b^2/s_w^2 = 800.5$; $p << .001$

Significance of Linear Correlation: $F_2 = s_p^2/s_r^2 = 11430.0$; $p << .001$

Significance of Curvilinearity: $F_3 = s_d^2/s_w^2 = 38.5$; $p < .001$

demonstrated algebraically that "in a long test of intercorrelated items, it matters very little how the individual items are weighted, thus showing that the relative order of scores . . . tends to be stable, or invariant for different methods of obtaining linear scores." Even though in the present study options rather than items were weighted, the intercorrelation of formula and Guttman scores might be expected to be rather high (as it was--r = .9059). However, although option weighting did not change the score distribution appreciably, it did radically alter the internal consistency (i.e., homogeneity) of the test.

The distinctly fan-shaped nature of the plot is due to greater dispersion of Y-scores within X-arrays at lower values of X than at higher, demonstrating that Guttman weighting has more effect on low-scoring examinees than high-scoring ones. Nedelsky (1954) and Lord (1965, 1968) also found that differential weighting affected least-able examinees most strongly. However, in these two studies the correct answers were not weighted differentially. Although in the present study the correct option, the distracters, and "omit" were weighted differentially, it appears as though the weights of distracters and omitted options have more effect on the scores of the examinee. (An interesting finding of this study was that the weight of "omit" was almost always lower than any of the other distracters in an item, demonstrating that students who omit items scored lower on the test as a whole than students who mark incorrect options.) Thus, low-scoring examinees are the best candidates for Guttman weighting because they have marked many distracters or omitted many items. If partial information is taken into effect via Guttman weighting, some of them improve their position, whereas others score far lower.

Differences in Weights for Men and for Women

Experimental Procedure

The subjects in this experiment were all 5000 men and all 5000 women. The only test used was Subtest 1 (40 items) in the verbal section. The responses of the women in Group 1 were scored with cross-validated Guttman weights, and the total scores were calculated. These responses were then scored with weights derived from a group of men, and the total scores were calculated. The correlation coefficient (r) and the correlation ratio (η_{yx}) were calculated from a scatterplot of these two score distributions. In this experiment only the regression of scores obtained using women's weights (Y) on scores obtained using men's weights (X) was dealt with.

Next, the same thing was done to the responses of the men in Group 1; i.e., their responses were scored first with cross-validated Guttman weights and then with weights derived from a group of women. Next both sets of total scores were calculated as before, and r and η_{yx} were calculated from a scatterplot of the former on the latter.

A two-way analysis of variance was performed on the weights of the options in Subtest 1. A set of weights was derived for both groups of men and for both groups of women. There were four factors in this design; sex (2 levels), items (40 levels), groups (2 levels) nested in sex, and options (6 levels) nested in items. There were 960 cells in the design ($2 \times 40 \times 2 \times 6 = 960$), and each cell contained the weight of a particular option derived in one of the groups. Sex and options were considered

fixed factors; items and groups were considered random factors. The linear model for this analysis is:

$$Y_{sigo} = \mu + \alpha_s + b_i + c_{g:s} + \delta_{o:i} + (\alpha b)_{si} + (\alpha\delta)_{so:i} + (bc)_{ig:s} + (c\delta)_{go:is} + \varepsilon_{sigo} .$$

The symbols α , b , c , and δ represent sex, items, groups, and options, respectively. The letters s , i , g , and o represent the levels of these respective factors. (Greek letters represent fixed factors, and Roman letters represent random factors.)

Experimental Results and Discussion

The correlation coefficient (r) and the correlation ratio (η_{yx}) calculated from the bivariate scatterplot for scores obtained by applying first women's weights and then men's weights to the responses of the women in Group 1 are .9767 and .9774, respectively. Non-linear variance accounted for only .13% ($\eta^2_{yx} - r^2 = .0013$). The correlation coefficient (r) and the correlation ratio (η_{yx})/from the bivariate scatterplot for scores obtained by applying first men's weights and then women's weights to the responses of the men in Group 1 are .9822 and .9827, respectively. The difference between η^2_{yx} and r^2 was .0009, and therefore the amount of non-linear variance accounted for was only .09%.

Table 8 shows the results of the analysis of variance. The sources of variation appear in the far left column and the corresponding F for each source of variation in the far right column. Four values of F were significant. Significant main effects were those for sex ($p < .05$), items

Table 8

ANOVA Table for Weights

Source of Variation	df	Mean Squares	Expected Mean Squares	F
Sex (s)	1	1.725×10^{-3}	$\sigma^2 + 6\,\sigma^2_{gxi:s} + 12\,\sigma^2_{sxi} + 240\,\sigma^2_{g:s} + 480\,\theta^2_s$	$F = MS_s/(MS_{g:s} + MS_{sxi} - MS_{gxi:s}) = 30.36^*$
Items (i)	39	1.977×10^{-2}	$\sigma^2 + 6\,\sigma^2_{gxi:s} + 24\,\theta^2_i$	$F = MS_i/MS_{gxi:s} = 132.7^{***}$
Group:s (g:s)	2	2.841×10^{-7}	$\sigma^2 + 6\,\sigma^2_{gxi:s} + 240\,\sigma^2_{g:s}$	$F = MS_{g:s}/MS_{gxi:s} = .0019$
Option:i (o:i)	200	8.528×10^{-3}	$\sigma^2 + \sigma^2_{gxo:si} + 4\,\theta^2_{o:i}$	$F = MS_{o:i}/MS_{gxo:si} = 46.93^{***}$
s x i	39	2.056×10^{-4}	$\sigma^2 + 6\,\sigma^2_{gxi:s} + 12\,\sigma^2_{sxi}$	$F = MS_{sxi}/MS_{gxi:s} = 1.38$
s x o:i	200	3.537×10^{-4}	$\sigma^2 + \sigma^2_{gxo:si} + 2\,\theta^2_{sxo:i}$	$F = MS_{sxo:i}/MS_{gxo:si} = 1.94^{***}$
g x i:s	78	1.490×10^{-4}	$\sigma^2 + 6\,\sigma^2_{gxi:s}$	$F \geq MS_{gxi:s}/MS_{gxo:si} = .820$
g x o:si	400	1.871×10^{-4}	$\sigma^2 + \sigma^2_{gxo:si}$	

*p < .05

**p < .01

***p < .001

($p < .001$), and options nested in items ($p < .001$); the significant interaction was that of sex and options ($p < .001$). Note that the error term for the F ratio testing the effect of sex was made up by combining mean squares. There were 2.36 degrees of freedom for this denominator (Walker and Lev, 1953, p. 373).

The mean square of groups nested in sex was extremely small in this analysis because a deliberate attempt was made to make the groups as much alike as possible. The groups were formed by blocking on the total verbal scores of the examinees. Thus, one would expect the weights of the groups in one sex to be similar. The inverted F ratio (Walker and Lev, 1953, p. 205) for groups nested in sex is significant ($p < .01$).

The correlational analysis showed that interchanging the weights of women and men does not change the distribution of total scores on Subtest 1 very much. However, as was shown in Table 8 the sexes do respond differently. Of particular interest in Table 8 is the fact that despite little statistical power for testing it, the main effect of sex was significant as was the interaction of sex with options, but the interaction of sex with items was not significant at the $p = .05$ level.

Not much concerning the sex differences in the items of the SAT has been published. Coffman's work (1961) is a notable exception. His findings show that although some rough hypotheses can be made about which items will be more difficult for one of the sexes, these hypotheses are not very accurate. Coffman (1961) was concerned about the influence of sex on items. However, the results reported in this study show that the interaction of sex with options is significant, whereas the interaction of sex with items is not. Great care is taken by test specialists at ETS to

choose items so that no bias in favor of either men or women will exist in the test as a whole. Perhaps the greatest source of bias is not in the stimulus word or words of the items but in the cues in the options of the items. It is possible that differences between sexes in responding to options is a neglected source of bias and that a closer look at the options in the test which Coffman used might help explain the reason that some items were more difficult for men than women and vice versa.

Summary and Conclusions

In this study cross-validated Guttman weights were used to score the options of all 150 SAT items. The effect of using Guttman weights was compared with the effect of using the conventional correction-for-guessing weights (1, 0, -1/4 for the five-option SAT). The following conclusions can be drawn on the basis of the findings of this study:

1. Differentially weighting the options of the SAT using Guttman's weighting technique dramatically improved the internal consistency of both verbal and mathematics subtests.

2. Differential weighting also improved the correlation between the two verbal subtests; however, the correlation between the two mathematics subtests decreased in value when Guttman weights were used, as did all four correlations of verbal subtests with mathematics subtests.

3. The correlation between total scores obtained by scoring options in a 40-item verbal subtest with correction-for-guessing weights and

total scores obtained by scoring options in the same subtest with cross-validated Guttman weights was .9059.

4. The ability level of women who choose a particular option is often different from that of men who choose the same option as can be seen from the interaction of sex with options.

44

REFERENCES

BAKER, F.B. Quantifying qualitative variables by the method of reciprocal averages. Occasional paper no. 7, Laboratory of Experimental Design, University of Wisconsin, 1969.

COFFMAN, W.E. Sex differences in responses to items of an aptitude test. NCME Yearbook, 1961, 117-124.

CRONBACH, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 12, 671-684.

CRONBACH, L.J. Essentials of psychological testing, 3rd ed. New York: Harper, 1970.

DAVIS, F.B. Estimation and use of scoring weights for each choice in multiple-choice test items. Educational and Psychological Measurement, 1959, 19, 291-298.

DAVIS, F.B., & FIFER, G. The errect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

DONLON, T.F. Prospectus for a program of research on the differential weighting of responses to objective tests. Educational Testing Service, April 1963.

EDGERTON, H.A., & KOLBE, L.E. The method of minimum variation for the combination of criteria. Psychometrika, 1936, 1, 183-137.

GUTTMAN, L. The quantification of a class of attributes: a theory and method of scale construction. In Paul Horst, The prediction of personal adjustment. N.Y.: Social Science Research Council, 1941.

HENDRICKSON, G.F. An assessment of the effect of differentially weighting options of a multiple-choice objective test using a Guttman weighting scheme. Baltimore. Working Paper No. 6, Center for Social Organization of Schools, The Johns Hopkins University, 1971.

HORST, P. Measuring complex attitudes. Journal of Social Psychology, 1935, 6, 369-374.

HORST, P. Obtaining a composite measure from different measures of the same attitude. Psychometrika, 1936, 1, 53-60.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 1933, 24, 417-441.

HOYT, C.J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

LORD, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika, 1958, 23, 291-296.

LORD, F.M. Worst distractor study. memo to W. Turnbull, Educational Testing Service, November 1965.

LORD, F.M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

McNEMAR, Q. Psychological Statistics. (4th ed.) N.Y.: John Wiley and Sons, Inc., 1969.

MOSIER, C.I. Machine methods of scaling by reciprocal averages. Proceedings, Research Forum. N.Y.: International Business Machines Corporation, 1946, 35-39.

MOSIER, C.I. Problems and designs of crossvalidation. Educational and Psychological Measurement, 1951, 11, 5-11.

NEDELSKY, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954, 14, 459-472.

RAJARATNAM, N., CRONBACH, L., & GLESER, G. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-55.

RICHARDSON, M. & KUDER, G.F. Making a rating scale that measures. Personel Journal, 1933, 12, 36-40.

SABERS, D.L., & WHITE, G.W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 1969,6, 93-96.

STANLEY, J.C. K-R 20 as the stepped-up mean item intercorrelation. Pp. 78-92 in the 14th Yearbook of the NCME, 1957.

STANLEY, J.C., & WANG, M.D. Differential weighting: a survey of methods and empirical studies. College Entrance Examination Board. New York, N.Y., November, 1968.

STANLEY, J.C. Reliability, Ch. 13 in R.L. Thorndike (ed.), Educational

Measurement (2nd ed.). Washington: American Council on Education, 1971. (In press.)

THORNDIKE, R.L.& HAGEN, E. Measurement and Evaluation in Psychology and Education. (3rd ed.)N.Y.: John Wiley & Sons, Inc., 1969.

WALKER, H.M., & LEV, J. Statistical Inference, N.Y.: Holt, Rinehart and Winston, 1953.

WILKS, S.S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

48

Appendix A

Discussion of the Weighting Technique Used

In this section the weighting procedure used in the present study to calculate option weights will be explained in detail. This scheme can be used to calculate the weights for any multiple-choice test (the number of alternatives associated with each option need not even be the same); however, the discussion here will be restricted to the test used in this study, the SAT. The data needed to calculate the weights are the options marked for a particular set of multiple-choice items by a particular group of people. These data can be represented in a matrix of ones and zeros like the following.

Data Matrix for the Responses of N People to I Items

Item	Option	Individual 1	2	3	. . .	N
1	1	1	0	1		
	2	0	0	0		
	3	0	1	0	. . .	
	4	0	0	0		
	5	0	0	0		
	omit	0	0	0		
2	1	0	0	0		
	2	1	0	1		
	3	0	0	0	. . .	
	4	0	0	0		
	5	0	0	0		
	omit	0	1	0		
.	.		.			
.	.		.			
.	.		.			
I	1	0	0	0		
	2	0	0	1		
	3	1	0	0		
	4	0	0	0		
	5	0	0	0		
	omit	0	1	0		

There are five alternatives associated with each item of the SAT. Each person must mark one of these alternatives or omit the item; thus, each individual can be placed in one of six mutually exclusive categories in each item. The ones in the matrix indicate which categories have been chosen.

The task at hand is to calculate a set of weights--one for each option in each item. The weighting scheme used is a modification of one devised by Guttman which maximizes the internal consistency of the test. The solution of the maximization process yields equations identical to those employed in a scaling technique known as the Method of Reciprocal Averages. The technique of calculating weights by this scaling method has been discussed in detail by Mosier (1946) and Baker (1969). The technique used in the present study is a modification of the Method of Reciprocal Averages. The modifications will be pointed out as they occur and reasons for incorporating them will be discussed. The following notation will be used:

i	an index for the ith individual
j,k	alternative indices for categories
N	total number of individuals
I	total number of items (also the total number of responses made by an individual)
n	total number of categories ($n = 6 \times I$, in this study)
w_k	weight assigned to category k
ε_{ij}	$\begin{cases} 1, & \text{if individual } i \text{ chooses category } j \\ 0, & \text{if individual } i \text{ does not choose category } j. \end{cases}$

The rights-only score for an individual is obtained by summing down the proper column; the number of individuals who marked a particular option is found by summing across the row associated with that option. Thus, the total score for individual i is

$$T_i = \sum_{k=1}^{n} \epsilon_{ik} w_k$$

and the number of people who choose category j is

$$n_j = \sum_{i=1}^{N} \epsilon_{ij} \quad .$$

The first step in this procedure is to choose an a priori set of weights. In this case the a priori weights used were the ones used presently to score the SAT: 1, for the correct choice, -1/4, for the four distracters, and 0, for "omit." The weights for any category (category j, for example) are calculated iteratively as indicated in the following steps.

Step 1

The scores are calculated on the other (i.e., the $I - 1$) items of the test. This score for individual i choosing category j is

$$S_{ij} = \sum_{k=1}^{n} \epsilon_{ij} \epsilon_{ik} w_k - \epsilon_{ij} w_j \quad .$$

These scores are then standardized. (In order to avoid introducing a new symbol, S_{ij} will now represent the standardized score for individual i.) The new weights will be based on these scores.

Guttman bases his weights on the total score without a correction for overlap and did not standardize these scores. In the present study total scores were calculated on the $I - 1$ items to remove the effect of the item in question on the criterion. However, removing an item has certain ramifications. If the item removed is easy, the scores on the $I - 1$ items will be lower than if the item removed is difficult. Thus, the options of easy items would have lower weights than the options of more difficult ones. To prevent this from occurring and to eliminate the influence of unequal standard deviations (see Stanley and Wang, 1968, p. 27),the scores were standardized in the present study.

Step 2

The average score for all individuals choosing category j ($\bar{S}_j$) is calculated for all categories.

$$\bar{S}_j = \frac{\sum_{i=1}^{N} S_{ij}}{n_j} = \frac{\sum_{i=1}^{N} \left(\sum_{k=1}^{n} \epsilon_{ij}\, \epsilon_{ik}\, w_k - \epsilon_{ij}\, w_j \right)}{\sum_{i=1}^{N} \epsilon_{ij}}$$

This average score is divided by a constant to keep it from becoming unmanageably large. In this case the constant was $I - 1$. Thus, the new weight for category j (w_j), based on the scores of people who chose it is

$$w_j = \frac{\bar{S}_j}{I - 1} .$$

Some researchers then scale these weights (cf. Baker, 1969

and Mosier, 1946). However, in this study the total score distribution was standardized rather than the weights.

Step 3

Repeat steps 1 and 2 until the weights and the internal-consistency reliability are stable. In this case three iterations were deemed adequate, after five were tried.

53