THE EFFECT OF FOUR DISTRACTOR ELIMINATION PROCEDURES ON THE RELIABILITY OF A MULTIPLE CHOICE ACHIEVEMENT TEST

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

COLLEGE OF EDUCATION

BY

WILLIAM D. BOYD, B. M., M. A.

DENTON, TEXAS

MAY, 1984

The Graduate School

Texas Woman's University

Denton, Texas

February 3 19 84

We hereby recommend that the Dissertation prepared under our supervision by William D. Boyd entitled The Effect of Four Distractor Elimination Procedures on the Reliability of a Multiple Choice Achievement Test

be accepted as fulfilling this part of the requirements for the Degree of Doctor of Philosophy.

Committee:

Basil L. Hamilton
Chairman

Department Chairperson

Accepted:

Barbara J. Kramer
Dean of The Graduate School

Dean, College of Education

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Test reliability can be defined as the ratio of true test variance to total test variance. True test variance is the amount of variation in test scores which is due to variations in levels of the trait being measured. Total test variance, on the other hand, is the true test variance plus variance due to error. The higher the ratio is between true and total test variance then the greater the reliability of the test. In the measurement of educational and psychological traits, it is desirable to use tests with high levels of reliability.

There are several techniques of test construction and statistical analysis which were developed to enhance test reliability. Magnusson (1967) stated that increasing the length of a test will increase its reliability. Reliability tends to increase with test length because "when the test increases in length, the true variance represents a greater share of the total variance" (p. 27). Increasing the length of a test, however, will increase the amount of time it takes to administer the test. Longer test periods result in less time for other academic activities.

Serlin and Kaiser (1978) described a technique for

increasing test reliability which utilized information from examinee response patterns. This technique permitted the examinees to choose any, none, or all of the four alternatives from multiple choice test items. Thus, instead of obtaining only one bit of information per item, as is the case in traditional testing procedures, this method resulted in five bits of information per item. This essentially lengthens the test, therefore, increases test reliability.

The item response theory techniques developed by Lord (1980) and others (Hambleton, 1983) appear to be very promising for increasing test effectiveness. Item response theory is based on the fact that examinees with lower levels of the trait being measured will respond differently to particular test items than will examinees with higher levels of the trait being measured. An easy item for one individual may be difficult for another. This variation in item difficulty with respect to ability can be graphed with the probability of making a correct response on the y-axis and a continuum of levels of the trait being measured as the x-axis. The resulting plot is called the item characteristic curve and can be used to predict performance on the item by other individuals possessing different levels of the trait.

The usefulness of item reponse theory for increasing

test reliability is readily apparent. Items can be selected for individuals of particular ability levels so that test discrimination is maximized and error due to guessing is minimized. These techniques are becoming increasingly popular; however, the computer programs required to determine the item characteristics are still very expensive to use.

Williams and Ebel (1957) conjectured that in most test items one good distractor accounts for the largest majority of the wrong responses. The authors believed that elimination of the less discriminating distractors would result in a more reliable test given a limited test period. The four-choice items in a multiple choice vocabulary exam were changed to three-alternative items by dropping the least discriminating distractors. A two-alternative form resulted from eliminating the two least discriminating distractors from each item. The results showed that: (a) the three-choice items took less time to answer than the four-choice items and the two-choice items were answered in less time than the three-choice items; (b) decreases in the number of alternatives were related to decreases in item difficulties and discrimination; (c) the three-alternative and two-alternative tests resulted in reliability coefficients which were nearly equal to that of the four-alternative test; and (d) given an equal amount of

time for each form of the test, the three-choice and two-choice tests both resulted in test reliabilities which were equal to the four-choice test.

In a later study, Ebel (1969) demonstrated that reliability is related to the number of alternatives used in multiple choice items. The author showed that as the number of alternatives increased so did reliability. Thus, tests made up of items with three choices are more reliable than tests made up of two-choice items, tests made up of items with four choices are more reliable than tests made up of three-choice items, and so forth. Tversky (1964), however, mathematically proved that tests made up of items with three alternatives are superior to tests made up of items with four alternatives when each test contains equal numbers of alternatives. This condition of equal numbers of alternatives within each test is called Tversky's condition.

Tversky's (1964) mathematical proof demonstrated that tests utilizing three-alternative items maximized the discrimination capacity, the power, and the information of the test. Discrimination capacity was defined as the total number of response patterns which may be obtained from a given test. Power was defined as one minus the probability of attaining perfect performance by chance alone. Information

was defined as the uncertainty associated with all of the possible response patterns to a given test. Tversky stated that the "three criteria proposed are monotonically related to each other" (p. 387), therefore; maximizing discrimination capacity also maximizes test power and test information. The mathematical proof demonstrated that three-alternative items did, indeed, maximize discrimination capacity given the same total number of alternatives in each test.

In order to empirically demonstrate the superiority of three-choice items over four-choice items, Costin (1970) developed four examinations containing four-choice items. Half of the items in each examination were reduced to three-alternative items by randomly discarding one of the distractors from each item. The results showed that the test containing three alternatives consistently demonstrated greater discrimination, greater power, and higher reliability than the four-alternative test. Costin did not describe how he chose which items were to be reduced to three alternatives. Also, Tversky's condition of maintaining a constant number of alternatives across all the tests was not met. Despite this, the results were significant.

In a later study, Costin (1972) used random selection procedures for selecting 50 of 100 four-choice items for reduction to three alternatives. One distractor from each item selected for reduction was randomly discarded. The

three-choice test and the four-choice test both resulted in similar Kuder-Richardson 20 reliability coefficients. Again, Costin failed to meet Tversky's condition.

Ramos and Stern (1973) found that reduction of the number of alternatives from five to four resulted in a slight decrease in test reliability. The decrease was nonsignificant, however, and Tversky's condition was not met. The decrease in reliability when changing from five to four-choice items is not surprising in light of the findings of Costin (1970, 1972) that three-choice items result in the greatest reliability. Ramos and Stern only considered four and five-alternative items in their study.

In support of the importance of meeting Tversky's condition, Grier (1975) held the number of alternatives constant in a mathematical proof of the superiority of three-choice items. Grier found that reliability is maximized for three-alternative items when the total number of alternatives included in the test is greater than 54.

Lord (1977) applied item response theory techniques to the problem of determining the optimal number of alternatives per item. A comparison of the item characteristic curves generated from three and four-choice tests showed that the three-choice tests were more effective for approximately 80% of the examinees. Three-choice items were less effective than four-choice items, however, for low-ability subjects.

Lord stated that "the effect of decreasing the number of choices per item while lengthening the test proportionally is to increase the efficiency of the test for high-level examinees and to decrease its efficiency for low-level examinees" (p. 36). The decrease in efficiency for low-ability subjects was due, in most part, to the greater likelihood of having to use error producing strategies, such as guessing, in order to respond to relatively difficult items. When guessing is a frequently used strategy, as is usually the case for examinees with very low levels of ability, a smaller number of alternatives increases the probability of guessing the correct answer. Thus, the total score may reflect correct guesses rather than actual ability.

Straton and Catts (1980) used four forms of a test in order to study not only the relative reliability of two, three, and four-choice test items, but also the relative reliability in terms of the average time taken to respond to the various item forms as well. The four-choice test consisted of 30 randomly selected items. A three-alternative form was developed by randomly selecting 10 additional items and randomly discarding one distractor from each item. Another three-choice form was developed by having a panel of three teachers discard the one distractor from each item

which they considered as the least attractive to examinees. A two-alternative test was developed by randomly selecting 20 additional items and randomly discarding two distractors from each item. The time at which the test began and ended was noted by the examinees on their tests. The results showed that the reliability was highest in the three-choice condition. Using the Spearman-Brown prophesy formula to determine the relative expected reliabilities of the tests, given the number of items completed during a certain period of time, the authors found that the three-alternative form remained superior. The random elimination of distractors resulted in higher reliabilities than those obtained by using panel selection procedures.

The evidence appears to favor the use of three-alternative multiple choice items in test construction when the necessary conditions are met. Many of the multiple choice tests currently used in school systems use four or five-alternative items. The purposes of the present study were: (a) to confirm the findings of previous studies that tests made with three-choice items are at least as reliable as tests made with four-choice items; and (b) to compare the various distractor elimination procedures in order to determine which one maximizes test reliability.

The hypotheses for the present study were: (1) the

three-alternative multiple choice tests would result in Kuder-Richardson 20 coefficients which were equal to or greater than the four-alternative multiple choice test under the conditions of equal numbers of alternatives per test and equal amounts of time per test; (2) the distractor elimination procedure which eliminates the least discriminating distractor would result in a higher Kuder-Richardson 20 coefficient than any of the other procedures.

CHAPTER II

METHOD

Subjects

The subjects consisted of all the sixth-grade students in two school districts who consented to participate in the study. Thirty-two subjects in the smaller of the two school districts were used to obtain the item statistics for an academic achievement test. One hundred sixty-eight subjects in the larger school district were administered the experimental forms of the achievement test.

Consent forms were sent home with all of the sixth-grade students in both school districts. The child and a parent or legal guardian were requested to sign the form indicating consent for the child to participate. An example of the consent form is included in Appendix A. The forms were then returned to the school by the students. Of 311 forms sent home, 192 were returned. Of the 192 consent forms that were returned, four subjects could not be identified or located during the test administrations. Ten of the tests resulted in a large number of omissions (three or more), a random or systematically invalid response pattern, or incomplete exam. These were not included in the data analysis.

Instruments

The academic achievement test consisted of 52 multiple choice items randomly selected from a pool of 160 items. The items included in the pool were four-alternative multiple choice items provided in a teacher manual for a health text book (Thompson, Althaus, Corbin, Gray, Sroka, & Thompson, 1983). This health text book was used by both school districts as part of their health curriculums. The items concerning chapter two of the text were omitted from the item pool because the topic covered by this chapter was not covered by all of the teachers. The 52 items selected from the pool are in Appendix B.

Five experimental forms of the test were developed. Form A resulted from eliminating one distractor from each item using random procedures. The distractors were eliminated from form B by a selection procedure using a panel of three sixth-grade health teachers. The teachers selected the one distractor from each item which they felt was the least attractive to students. In the event that a consensus was not achieved, the distractor selected by two of the three teachers was eliminated. In the event that the criteria of two out of three could not be achieved then the distractor was randomly selected for elimination. Form C was developed by discarding the least popular alternative. The least

popular alternative was defined as the alternative in each item chosen least frequently by the subjects who were administered the test in the smaller school district. When two distractors achieved equally low levels of popularity then one of the two distractors was eliminated at random. The least discriminating distractor, determined through an analysis of the item alternatives, was eliminated from form D. The procedure for performing the alternative analysis will be described in the procedures section. The distractors eliminated from each test form are indicated in Appendix C.

Form E consisted of 39 four-choice items randomly selected from the 52 items included in the original health test. The items used in form E are specified in Appendix D. The total number of choices per test for each form was 156, as indicated in table 1; therefore, Tversky's condition of an equal number of alternatives per test was met.

Procedures

In order to determine which distractors were the least popular and least discriminating, the 52 four-alternative multiple choice items were administered to 32 sixth-graders in the smaller of the two school districts. The least popular distractor was defined as the distractor in each item which was selected least frequently. The least discriminating distractor was defined as that distractor

in each item which had the lowest value for predicting test performance.

Table 1

Number of Subjects, Number of Items, Number of choices per Item, and Number of Choices per Test Under Each Condition

	Test Form				
	A	B	C	D	E
Subjects	33	30	32	33	30
Items per Test	52	52	52	52	39
Choices per Item	3	3	3	3	4
Choices per Test	156	156	156	156	156

The discrimination value of the distractors was determined by a means similar to one typically used to determine the discrimination value of items themselves. The proportion of examinees who scored in the upper 50% in terms of total test score and selected a particular distractor was subtracted from the proportion of examinees from the lower 50% who selected the particular distractor. For example, the item response data for an item k is given in Table 2. The discrimination value of distractor B is

10/40 - 5/40 which equals .125, while the value of distractor C is 4/40 - 0/40 which equals .1, and the discrimination value of distractor D is 1/40 - 0/40 which is equal to .025; thus, distractor D is the least discriminating distractor for this item. This procedure was repeated in order to eliminate the least discriminating distractor from every item of form D. Appendix E contains the results of the analyses. In the event of two distractors in one item with equal discrimination values, one of the two distractors was randomly selected for elimination.

Table 2

Example of Responses to an Item

	Alternatives			
Examinees	A*	B	C	D
Upper 50%	15	5	0	0
Lower 50%	5	10	4	1

One-hundred fifty-eight subjects of the larger school district were randomly assigned to five groups. Group one had 33 subjects and was administered form A, 30 subjects in group two were given form B, group three had 32 subjects and was administered form C, form D was administered to 33

subjects in group four, and the 30 subjects in group five received form E. The subjects were administered their repective forms of the test during their regularly scheduled health instruction period.

In order to determine the amount of time it took subjects to complete the various forms of the test, a digital clock was placed at the front of the room. The subjects were instructed to write down the time at the beginning and completion of their tests. There was, however, no time limit given for completing the test. The average time taken to answer items in the various test forms was determined by dividing the time it took, on the average, to complete the entire test by the number of items in the test.

After the tests were distributed, the following instructions were given to the subjects:

> You are now going to take a health test. This test will not be counted as part of your grade but will be used to see how well your class has learned about health. This test has no time limit, but usually takes from 15 to 30 minutes to complete. Do the best you can to answer every item. Leave no item blank. When you begin the test, write down the time that is showing on this clock (point to clock). Also, write down the time when you finish the test. Are there any questions? Begin the test.

A question asking: "What time is showing on the clock at the front of the room?" and space for a response were included at the beginning and ending of every test form.

The resulting Kuder-Richardson 20 coefficients were analyzed using a statistical procedure described by Marascuilo (1966). This procedure allowed the comparison of all five coefficients simultaneously. The resulting value was comparable to Chi-square with J - 1 degrees of freedom (J = the number of coefficients to be compared). A computed value which is less than the tabled value indicates that the differences among the coefficients could have resulted from chance.

Marascuilo (1966) also demonstrated pairwise and contrast procedures for post hoc analysis of correlation coefficients. These techniques maintain the experiment-wise type I error rate at a predetermined level. Contrasts may be made between two individual coefficients, between an individual coefficient and a group of coefficients, or between two groups of coefficients.

An analysis of variance was performed on the average response time to the test items. The Tukey method for pairwise comparisons was used to determine which test forms resulted in significantly different response times.

CHAPTER III

RESULTS

All of the three-alternative multiple choice tests resulted in Kuder-Richardson 20 coefficients which were equal to or greater than the four-alternative multiple choice test under the conditions of an equal number of alternatives per test and an equal amount of time per test; therefore, the first hypothesis was confirmed. The distractor elimination procedure which eliminated the least discriminating distractor did not result in a significantly higher Kuder-Richardson 20 coefficient than the other distractor elimination procedures; therefore, the second hypothesis was not confirmed.

Table 3 gives the Kuder-Richardson 20 coefficients achieved under each of the testing conditions. The four-alternative test form (form E) achieved a Kuder-Richardson 20 coefficient of .67848, while the three-choice form of the test from which the least discriminating distractor was eliminated (form D) achieved a Kuder-Richardson 20 coefficient of .84787. The three-choice test forms which utilized random, teacher panel selection, and least popular alternative procedures for eliminating distractors achieved Kuder-Richardson 20

coefficients of .75102, .75546, and .73153, respectively. Using the statistical procedures for comparing correlation coefficients developed by Marascuilo (1966), no significant differences were found between any of the Kuder-Richardson 20 coefficients. A value of 2.925 was computed which is comparable to Chi-square with four degrees of freedom. Chi-square with Alpha equal to .05 and four degrees of freedom is 9.488; therefore, no significant differences were determined to exist between any of the coefficients.

Table 3

Method of Distractor Elimination and Kuder-Richardson 20 Coefficients

Test Form	Distractor Elimination Procedure	KR 20
A	Random	.75102
B	Teacher Panel Selection	.75546
C	Least Popular Distractor	.73153
D	Least Discriminating Distractor	.84787
E	None	.67848

The average time taken to compete form A of the test was 742 seconds, form B took an average of 846 seconds, form C averaged 916 seconds, an average of 973 seconds was taken

to complete form D, and form E took an average of 572 seconds. The average time per item was obtained by dividing the average time per test by the number of items in the test, the results are shown in Table 4.

Table 4

The Average Time per Test, the Number of Items per Test, and the Average Time per Item

	Test Form				
	A	B	C	D	E
Time per Test	742	846	916	973	572
Number of Items per Test	52	52	52	52	39
Time per Item	14.27	16.26	17.62	18.71	14.67

Table 5 shows the results of analysis of variance performed on the amount of time required for administration of the various test forms. The results indicated that significance was obtained at an Alpha level of .05. A pairwise comparison of the average time per item for each test form resulted in significance between the time achieved by form D and form E, and between form D and form A. The results are given in Table 6. This outcome was contrary

to the findings of previous studies in which three-choice items were found to take less time than four-choice items.

Table 5

Analysis of Variance of the Average Testing Times

Source of Variation	df	Mean Square	F
Between Groups	4	1142.37	35.662*
Within Groups	152	32.03	

* Significant at Alpha = .05

The average time per item for each test form was divided into the average total amount of time taken to complete form D; this resulted in the number of items under each testing condition which would be necessary to make tests which are equal in testing time to form D. The Kuder-Richardson 20 coefficients which would be expected, given tests containing the number of items necessary to achieve equality in testing time, were determined by using the Spearman-Brown prophesy formula. The results are given in Table 7. Using Marascuilo's (1966) technique to test the differences between coefficients given equal testing times, a value of 2.018 was obtained. This value is comparable to Chi-square with four degrees of freedom. Chi-square with Alpha equal

to .05 and four degrees of Freedom is 9.488; therefore, no significant differences were found.

Table 6

Pairwise Comparisons of Average Testing Times

Forms	Difference		$1- q_{J,J(n-1)}(MS_w/n)$	Confidence Interval
A - B	-1.985	±	3.669	(1.684,-5.654)
A - C	-3.352	±	3.669	(.317,-7.020)
A - D	-4.370	±	3.669	(-.701,-8.039)*
A - E	- .394	±	3.669	(3.275,-4.063)
B - C	-1.367	±	3.669	(2.302,-5.036)
B - D	-2.452	±	3.669	(1.217,-6.121)
B - E	1.591	±	3.669	(5.260,-2.078)
C - D	-1.085	±	3.669	(2.584,-4.754)
C - E	2.958	±	3.669	(6.627,- .711)
D - E	4.043	±	3.669	(7.712, .373)*

* Significant at Alpha = .05

In order to make a test with a Kuder-Richardson 20 coefficient similar to that achieved by Form D, approximately 137 four-choice items, equivalent to those included in form E, would be necessary. By multiplying the average time per item by the number of items needed to obtain a

Kuder-Richardson 20 coefficient equal to that obtained by form D, it was found that approximately 2010 seconds of testing time would be needed for form E, 1370 seconds for form A, 1512 seconds for form B, and 1868 seconds for form C (see Table 8). Despite these differences in length of testing time and number of items necessary to achieve expected equality in terms of Kuder-Richardson 20 coefficients, there were no statistical differences between the four distractor elimination procedures.

Table 7

Number of Items Necessary to Make Tests Equal in Testing Time to Form D and the expected Kuder-Richardson 20 Coefficients

	Test Form				
	A	B	C	D	E
Number of Items	68	60	55	52	66
Expected KR 20	.798	.781	.742	.848	.728

Table 8

The Amount of Time Needed for Each Test to Achieve a Kuder-Richardson 20 Coefficient Equal to that of Form D

	Test Form				
	A	B	C	D	E
Number of Items	96	93	106	52	137
Time per Item	14.27	16.26	17.62	18.71	14.67
Time per Test	1370	1512	1868	973	2010

CHAPTER IV

DISCUSSION

No statistical differences were found between the Kuder-Richardson 20 coefficients resulting from administration of the five test forms. The first hypothesis which predicted at least equivalence between the three and four-choice forms, given the conditions of equal testing times and equal numbers of alternatives, was confirmed. The hypothesis predicting superiority of the procedure which eliminated the least discriminating distractor was not confirmed.

The results lend further support to the findings of previous studies that three-alternative multiple choice items are at least as reliable as four-choice items. Although no statistical differences were found between the Kuder-Richardson 20 coefficients of the various test forms, the three-choice tests resulted in increased test efficiency; test efficiency is defined as the greatest reliability with the fewest test items and fewest alternatives per item in the least amount of time. Although form D resulted in the greatest average time per item, it was estimated that 137 four-choice items, equivalent to those included in form E, would be needed to make a test with a Kuder-Richardson 20 coefficient identical to that of form D; thus, form D appears to be more efficient than form E.

The time and resources necessary to perform a pretest of the items in order to do the analysis of the distractors should be taken into consideration in determining the efficiency of a test. Random elimination of distractors resulted in an improvement in test efficiency over the four-choice items and took much less time to prepare than did the panel selection procedure, the least popular alternative procedure, or the least discriminating distractor procedure. The procedure which eliminated the least popular alternative took about the same amount of time as the least discriminating distractor procedure and resulted in a non-significantly lower Kuder-Richardson 20 coefficient. The teacher panel selection procedure also required a great deal of time and effort. The most efficient distractor elimination procedure for a given testing situation will likely depend on the number of individuals to be tested and the number of times a test is to be used. For tests used once with small numbers of people the random procedure may be best; whereas, the least discriminating distractor procedure may result in the greatest test efficiency if a test is to be given repeatedly or to a large number of examinees.

The teacher panel selection procedure of eliminating distractors resulted in a test which achieved a

Kuder-Richardson 20 coefficient practically identical to that achieved by the random selection procedure. There was little agreement among the panel members concerning the alternative which would be least attractive to students. Unanimous consensus of all three teachers was achieved on only 24 of the items. When asked about their reaction to the task, the teachers commented that their selection of alternatives to be eliminated depended upon how they covered the material and what was stressed during instruction; thus, there was sometimes very little agreement about which distractor was the least attractive. The panel selection technique appeared to be a very inefficient method for eliminating distractors.

The elimination of the least popular distractor required practically as much time to accomplish as the distractor analysis procedure and yet resulted in a Kuder-Richardson 20 coefficient similar to the other procedures. The items had to be pre-tested in order to determine the least popular distractor; therefore, this procedure was very time consuming.

The analysis of the alternatives in order to eliminate the least discriminating distractor resulted in several interesting observation. Some of the distractors actually resulted in negatively discriminating items; that is, the subjects scoring within the upper 50% on the entire test

sometimes selected a particular distractor while those scoring in the lower 50% would select the correct answer. The distractor analysis procedure resulted in the elimination of these negatively discriminating distractors; thus, the discrimination value of some of the items and perhaps the entire test was improved. Forms A, B, C, and E all contained seven negatively discriminating items each. Form D, on the other hand, contained only three negatively discriminating items. Form D also resulted in the greatest average time for responding to items. Although the average time per item for form D was not significantly different from those required by forms B and C, it was significantly greater than the time required for forms A and E. Perhaps the elimination of very attractive, but negatively discriminating, distractors resulted in the subjects taking more time to respond.

The lack of statistical differences between Kuder-Richardson 20 coefficients may have resulted from the relatively small number of subjects tested under each condition. The Kuder-Richardson 20 coefficients were measures of internal consistency. What may be more important than the internal consistency of tests may be their efficiency. If a shorter, easier, less time consuming test can achieve a similar Kuder-Richardson 20 coefficient to

that achieved by a longer, harder, more time consuming test than the shorter test may be the preferred choice.

Although there were no statistical differences between the distractor elimination procedures, it may be that administration of the test forms to larger numbers of subjects would have resulted in significance. Forms A, B, and C tended to cluster around a Kuder-Richardson 20 coefficient of .75 while form D resulted in a coefficient of .84. While not significant, this could indicate a trend which would continue given additional subjects. Further study may prove this to be so.

Based upon the results of previous studies and the results of the present study, it can be concluded that three-alternative multiple choice tests are at least as reliable as, and likely more efficient than, four-choice multiple choice tests. Further study is warranted in order to determine how the procedure which eliminated the least discriminating distractor would compare to other procedures when used with larger samples. Given that the random technique and the least popular alternative technique both resulted in Kuder-Richardson 20 coefficients which were practically identical, any one of these procedures may be chosen for comparison with the least discriminating distractor technique. Limiting the study to only two

distractor elimination techniques will allow larger samples for study and will simplify the statistical analyses.

The use of three-choice, rather than four-choice, multiple choice items is recommended. Lack of significance between the Kuder-Richardson 20 coefficients achieved in the present study indicated that it may not be of importance how the distractor is eliminated. In terms of efficiency, however, the three-choice items in form D were found to be superior to the four-choice items.

APPENDIX A

CONSENT FORM

My name is Bill Boyd, I am a doctoral student in educational psychology at the Texas Woman's University. As part of the requirements for earning a doctoral degree, I am required to perform a research project for a dissertation. The research I intend to do involves you child, therefore, your consent is required for his or her participation.

My research project is designed to determine better ways to develop tests. This means that I would like to administer a specially designed test to your child. This test is about an hour long health test which will be given during your child's regularly scheduled health period.

There are always benefits and risks associated with research projects. Although the risks are minimal in this study, I would like to describe them so that you are aware of what they are. First, the test will take about an hour of your child's school time. The test is based on your child's health book, however, and will be given during the regularly scheduled health period. Second, there is a possibility that the test scores could be misused. The test scores will be provided to the teachers in summary form which means that the teacher will know how the class as a whole performed but will not know any of the individual scores; in this way, a low score cannot harm a child. I will do everything I can to make sure that no one sees the tests after they have been scored so that they cannot be misused.

Your child is allowed to quit or refuse to participate at any time and will not be in any kind of trouble for doing so. I would like your child to read this, or, have it read to him or her. Signing your name indicates your willingness to have your child participate. Your child signing indicates his or her willingness to participate.

Thank you for your help in this project. If you have any questions call me in the evenings at 641-8342.

_____

Parent's Signature Date

_____

Student's Signature Date

APPENDIX B

HEALTH TEST

NAME_____

TEACHER_____

WHAT TIME IS SHOWING ON THE CLOCK AT THE FRONT OF THE ROOM?___

1. When people experience too much stress,
   a. they usually do their best.
   *b. their health may be harmed.
   c. their relationships improve.
   d. they have fewer problems.

2. People in voluntary agencies are helping solve noise problems by
   a. protecting workers from high noise levels.
   b. developing quieter products.
   *c. making people aware of how to protect their hearing.
   d. setting decibel limits.

3. Which is an example of an infectious disease?
   a. high blood pressure
   b. heart attack
   c. cancer
   *d. influenza

4. If you see someone struggling in the water, do NOT try
   a. reaching out with a towel.
   *b. swimming to the person's rescue.
   c. throwing a floating object to the person.
   d. telephoning the fire department.

5. Bicycle accidents may be caused by
   a. stopping at stop signs.
   b. having reflectors on your bike.
   c. paying attention to road conditions.
   *d. ignoring other drivers and pedestrians.

6. Which is NOT a good way to deal with stress?
   *a. keeping problems to yourself
   b. trying to solve your problems
   c. staying healthy
   d. using physical activity to relax

7. Jim has a rash that looks like one Mike had. Mike offers Jim his prescription drug and Jim uses it. This is an example of
   a. a benefit of medicine.
   b. a side effect.
   c. drug overdose.
   *d. drug misuse.

8. The most important way to care for skin is to
   *a. keep it clean.
   b. use a lot of special skin care products.
   c. see a skin specialist.
   d. give up chocolate.

9. The main reason young people lose permanent teeth is
   a. periodontal disease.
   *b. dental caries.
   c. accidents.
   d. bone disease.

10. When something goes wrong with an artery in the brain, a person suffers
   *a. a stroke.
   b. hypertension.
   c. a heart attack.
   d. diabetes.

11. Which can be a long-term effect of noise pollution?
   a. birth defects
   b. infectious disease
   *c. hearing loss
   d. chronic bronchitis

12. Which is NOT a good way to avoid pathogens?
   a. Take care of your body.
   b. Wash your hands before eating.
   *c. Visit a friend who has an infectios disease.
   d. Make sure that food and water are clean.

13. A long-term effect of marijuana is
   a. hallucination.
   b. relaxation.
   c. drowsiness.
   *d. lung damage.

14. If someone's hair or clothing catches fire, first
    *a. wrap the person in a blanket, rug, or jacket.
    b. throw the person to the ground.
    c. ruch the person to a sink for water.
    d. get a garden hose to put out the fire.

15. The main carbohydrates are sugar, starch, and
    a. vitamins.
    *b. fiber.
    c. amino acids.
    d. fats.

16. One way to protect your health in the future is by
    *a. eating a balanced diet.
    b. avoiding regular exercise.
    c. eating saturated fats.
    d. trying not to stay lean.

17. Which is a government agency that is responsible for helping the environment?
    a. Sierra Club
    b. National Wildlife Federation
    *c. National Park Service
    d. Boy Scouts of America

18. If you want to improve your speed, which sport shoud you choose?
    a. badminton
    b. hiking
    c. modern dance
    *d. soccer

19. A dangerous gas in cigarettes is
    a. alcohol.
    b. oxygen.
    *c. carbon monoxide.
    d. LSD.

20. Jim is on a diet. Which would be the best snack for him?
    a. a dish of ice cream
    *b. an orange
    c. a bottle of cola
    d. several cookies

21. The major causes of death in the United States today are
    a. influenza and cancer.
    b. pneumonia and influenza.
    c. cardiovascular disease and pneumonia.
    *d. cardiovascular disease and cancer.

22. A realistic goal is one that is
    a. impossible to reach.
    b. reached by very few people.
    c. accepted by all your friends.
    *d. possible to reach.

23. Water, land, and fuels taken from the land are
    a. pollution.
    *b. natural resources.
    c. particulates.
    d. sanitary sandfills.

24. Substances that give processed foods extra taste, color, or nutrients are called
    a. freeze-dried foods.
    *b. additives.
    c. saturated fats.
    d. calories.

25. Alcohol is a
    *a. depressant.
    b. stimulant.
    c. volatile substance.
    d. narcotic.

26. In cases of severe bleeding, the first thing to try is
    a. artificial respiration.
    b. patting the wound.
    *c. applying direct pressure.
    d. tying a cloth gently over the wound.

27. A dependable source of health care information is someone who
    *a. is up-to-date.
    b. likes you.
    c. is not qualified in the subject.
    d. has let you down in the past.

28. A mentally healthy person
    a. expects to be the best at everything.
    b. lives only for the present.
    *c. accepts his or her limits.
    d. never has problems to deal with.

29. One way you can help your environment at home is by
    *a. planting bushes and trees.
    b. using pesticides.
    c. throwing out newspapers and magazines.
    d. using electric appliances whenever possible.

30. Which food does NOT belong to one of the the Basic Four Food Groups?
    a. beef
    b. skimmed milk
    *c. honey
    d. whole-wheat bread

31. At which decibel level do sounds begin to be dangerous to hearing?
    a. 55
    b. 70
    *c. 80
    d. 95

32. Which form of waste disposal generally is illegal today?
    a. treatment plants
    b. sanitary landfills
    c. burning
    *d. dumping

33. People did not connect pathogens with disease until the late
    *a. 1700s.
    b. 1400s.
    c. 1900s.
    d. 1500s.

34. Which is a good way to make wise health decisions?
    a. Try not to make any decisions at all.
    b. Act on your first feelings.
    c. Take the advice of a friend.
    *d. Follow a step-by-step plan.

35. Industries that return heated water to a body of water cause
    a. sewage pollution.
    *b. thermal pollution.
    c. smog pollution.
    d. particulate pollution.

36. Your body uses the largest number of calories when you are
    *a. running.
    b. sleeping.
    c. walking.
    d. reading.

37. What causes the most land pollution?
    a. smog
    b. liquid wastes
    *c. solid wastes
    d. oil spills

38. The changes of stress
    a. occur only when something bad happens.
    *b. prepare the body to face demanding situations.
    c. cause less bloood to go to the brain and heart.
    d. are the same for everyone.

39. Which is an example of how to build a good relationship?
    a. Sue borrows Phil's things without asking.
    b. Larry keeps his problems from his friend Mike.
    *c. Pat tells Ina why she disagrees with her.
    d. Ted makes jokes about Jeff's handicap.

40. Billy wants to go to the ball game. Lee wants to stay home. Which is a compromise?
    a. They go to the ball game.
    b. They stay home.
    *c. They watch the ball game at home on TV.
    d. They stop talking to each other.

41. Which is NOT true of body fat?
    a. It gives the body shape.
    *b. It keeps the body cool in hot weather.
    c. It stores energy.
    d. It protects organs such as the heart.

42. To be safe in a boat, do NOT
    a. remind passengers to stay seated.
    b. wear a personal flotation device.
    *c. stay out in bad weather.
    d. watch out for swimmers.

43. Jim is having trouble with his science project. Tina, who is good in science can be a friend by
    *a. making suggestions to improve the project.
    b. doing the project for Jim.
    c. saying nothing so Jim will not be angry with her.
    d. laughing at Jim.

44. Tiny liquid or solid wastes added to the air are called
    *a. particulates.
    b. sewage.
    c. decibels.
    d. natural resources.

45. Which is NOT a cardiovascular disease?
    a. hypertension
    b. atherosclerosis
    *c. diabetes
    d. stroke

46. A drug that speeds up the work of the nervous system is a
    *a. stimulant.
    b. depressant.
    c. hallucinogen.
    d. medicine.

47. At what rate does the liver usually break down alcohol?
    a. one drink in one day
    b. two drinks in one hour
    *c. one drink in one hour
    d. twice as fast as usual if you drink coffee

48. People should avoid
    a. activities they cannot do well.
    *b. situations that cause too much stress.
    c. thinking about problems or discussing them.
    d. compromising with family members.

49. To avoid ingrown toenails, you should
    a. file the toenails in a rounded shape.
    b. remove dirt from under the toenails.
    *c. cut the toenails straight across.
    d. wear heavy socks.

50. Which is a good way to deal with the pressure to use drugs?
    a. Try everything first and then decide.
    *b. Think through your ideas on drugs ahead of time.
    c. Base your decision on what your friends do.
    d. Base your decision on how much pressure you feel.

51. Which is NOT an effect of air pollution on people?
    *a. improved coordination
    b. eye irritation
    c. breathing difficulties
    d. dizziness

52. Knowing signs of disease is important to help you decide
    a. what disease you have.
    *b. whether to call a doctor.
    c. what medicine to take.
    d. whether to take your temperature.

WHAT TIME IS SHOWING ON THE CLOCK AT THE FRONT OF THE ROOM?___

* Correct answer

APPENDIX C

The distractors Eliminated from the Experimental Test Forms

	Test Form			
Item	A	B	C	D
1	c	d	d	d
2	a	b	d	d
3	a	a	b	c
4	c	d	c	a
5	b	a	b	b
6	b	c	c	c
7	a	a	c	b
8	d	d	c	c
9	a	d	d	a
10	d	d	d	c
11	a	a	b	d
12	b	a	b	a
13	b	b	b	c
14	d	d	c	b
15	c	c	c	d
16	c	c	c	b
17	a	a	d	d
18	c	a	a	a
19	b	a	b	d

Item	Test Form A	B	C	D
20	d	a	a	a
21	c	b	c	c
22	c	a	b	b
23	d	a	a	a
24	d	c	a	a
25	d	c	b	c
26	b	a	b	d
27	b	d	d	b
28	d	d	d	d
29	c	d	c	b
30	d	b	a	d
31	b	a	a	b
32	b	a	a	b
33	b	b	b	b
34	c	a	a	a
35	c	d	d	a
36	d	b	c	d
37	a	a	a	a
38	c	d	d	d
39	b	a	a	a
40	d	d	d	d

Item	Test Form A	B	C	D
41	d	a	a	c
42	d	b	d	b
43	d	d	d	d
44	c	c	c	c
45	d	d	a	b
46	c	c	d	b
47	d	d	b	d
48	c	d	d	d
49	d	b	d	a
50	c	a	c	a
51	d	b	c	c
52	c	d	d	a

APPENDIX D

The 39 Items Selected from the Health Test for Form E

Form E Item Number	Health Test Item Number
1.	1.
2.	2.
3.	3.
4.	5.
5.	6.
6.	7.
7.	9.
8.	11.
9.	12.
10.	14.
11.	15.
12.	16.
13.	17.
14.	18.
15.	19.
16.	20.
17.	22.
18.	23.
19.	24.
20.	25.

Form E Item Number	Health Test Item Number
21.	26.
22.	27.
23.	29.
24.	30
25.	32
26.	33.
27.	34.
28.	35.
29.	37.
30.	38.
31.	39.
32.	40.
33.	43.
34.	44.
35.	46.
36.	47.
37.	48.
38.	50.
39.	51.

APPENDIX E

Distractor Analysis of the Health Test Items

Item	Distractor	Number Selecting Alternative From: Upper 50%	Lower 50%	Discrimination Value
1	a	1	3	.0625
	*b	14	8	
	c	0	3	.09375
	d	1	2	.03125
2	a	3	5	.0625
	b	2	3	.03125
	*c	7	7	
	d	4	1	-.09375
3	a	3	5	.0625
	b	1	2	.03125
	c	11	8	-.09375
	*d	1	1	
4	a	3	1	-.0625
	*b	12	5	
	c	0	3	.09375
	d	1	7	.1875

5	a	0	1	.03125
	b	0	0	.0
	c	0	2	.0625
	*d	16	13	
6	*a	13	11	
	b	2	4	.0625
	c	0	0	.0
	d	1	1	.0
7	a	2	7	.15625
	b	2	1	-.03125
	c	0	0	.0
	*d	12	8	
8	*a	16	14	
	b	0	1	.03125
	c	0	0	.0
	d	0	1	.03125
9	a	6	4	-.0625
	*b	5	8	
	c	5	4	-.03125
	d	0	0	.0
10	*a	8	3	
	b	3	8	.15625
	c	4	4	.0
	d	1	1	.0

11	a	1	2	.03125
	b	0	1	.03125
	*c	14	12	
	d	1	1	.0
12	a	3	2	-.03125
	b	0	2	.0625
	*c	12	10	
	d	1	2	.03125
13	a	4	5	.03125
	b	0	2	.0625
	c	4	4	.0
	*d	8	5	
14	*a	8	9	
	b	8	6	-.0625
	c	0	0	.0
	d	0	1	.03125
15	a	3	6	.09375
	*b	3	1	
	c	3	4	.03125
	d	7	5	-.0625
16	*a	14	10	
	b	1	2	.03125
	c	0	1	.03125
	d	1	3	.0625

17	a	3	3	.0
	b	4	7	.09375
	*c	6	4	
	d	3	2	-.03125
18	a	2	1	-.03125
	b	1	4	.09375
	c	1	3	.0625
	*d	12	8	
19	a	1	2	.03125
	b	0	1	.03125
	*c	6	5	
	d	9	8	-.03125
20	a	0	0	.0
	*b	16	15	
	c	0	1	.03125
	d	0	0	.0
21	a	7	9	.0625
	b	1	2	.03125
	c	2	1	-.03125
	*d	6	4	
22	a	0	4	.125
	b	1	1	.0
	c	2	4	.0625
	*d	13	7	

23	a	0	6	.1875
	*b	15	6	
	c	0	2	.0625
	d	1	2	.03125
24	a	0	0	.0
	*b	10	8	
	c	3	3	.0
	d	3	5	.0625
25	*a	2	1	
	b	1	3	.0625
	c	5	4	-.03125
	d	8	8	.0
26	a	1	1	.0
	b	0	2	.0625
	*c	11	10	
	d	4	3	-.03125
27	*a	11	8	
	b	4	5	.03125
	c	1	2	.03125
	d	0	1	.03125
28	a	1	4	.09375
	b	0	1	.03125
	*c	15	11	
	d	0	0	.0

29	*a	12	11	
	b	3	1	-.0625
	c	0	1	.03125
	d	1	3	.0625
30	a	0	1	.03125
	b	2	3	.03125
	*c	13	11	
	d	1	1	.0
31	a	3	3	.0
	b	6	4	-.0625
	*c	5	2	
	d	2	7	.15625
32	a	1	1	.0
	b	2	1	-.03125
	c	1	7	.1875
	*d	12	7	
33	*a	10	2	
	b	0	0	.0
	c	6	13	.21875
	d	0	1	.03125
34	a	1	0	-.03125
	b	1	6	.15625
	c	3	4	.03125
	*d	11	6	

35	a	5	3	-.0625
	*b	5	5	
	c	3	5	.0625
	d	3	3	.0
36	*a	13	11	
	b	2	5	.09375
	c	0	0	.0
	d	1	0	-.03125
37	a	2	2	.0
	b	1	4	.09375
	*c	13	2	
	d	0	8	.25
38	a	4	6	.0625
	*b	8	3	
	c	4	6	.0625
	d	0	1	.03125
39	a	0	1	.03125
	b	0	2	.0625
	*c	16	11	
	d	0	2	.0625
40	a	0	1	.03125
	b	0	2	.0625
	*c	15	13	
	d	1	0	-.03125

41	a	1	4	.09375
	*b	5	2	
	c	7	4	-.09375
	d	3	6	.09375
42	a	0	2	.0625
	b	3	4	.03125
	*c	13	9	
	d	0	1	.03125
43	*a	16	11	
	b	0	4	.125
	c	0	1	.03125
	d	0	0	.0
44	*a	12	5	
	b	2	5	.09375
	c	2	2	.0
	d	0	4	.125
45	a	4	2	-.0625
	b	6	1	-.15625
	*c	1	3	
	d	5	10	.15625
46	*a	4	2	
	b	4	3	-.03125
	c	5	7	.0625
	d	3	4	.03125

47	a	5	4	-.03125
	b	1	7	.1875
	*c	4	1	
	d	6	4	-.0625
48	a	4	4	.0
	*b	11	10	
	c	0	2	.0625
	d	1	0	-.03125
49	a	6	5	-.03125
	b	5	4	-.03125
	*c	5	6	
	d	0	1	.03125
50	a	3	2	-.03125
	*b	8	7	
	c	1	3	.0625
	d	4	4	.0
51	*a	11	3	
	b	2	4	.0625
	c	2	3	.03125
	d	1	6	.15625
52	a	9	3	-.1875
	*b	5	7	
	c	2	4	.0625
	d	0	2	.0625

APPENDIX F

Raw Data: Responses to Form A

Items

Subjects

1111111111111111011111111111111111111111111101111111
1101111101111111001111110111111111111111111101111111
1111111100111111011101110111111111111111111101111011111
1111111110111001011111110111111101101111111111011110
1100111111111110111111111111111111111111011101101100110
1001111110110011011111111111110101010111111111111110111
0101111111111111111101111111111101100111011001001111
1001111110100101011111111101111111101111111111111010110
1111111110100111111111111111111101010110110111101110100
1100111100111111011111010111111101111111111100010011
1000111111101010111011110011111010101111111111101111111
1111111100011111011110111111001110001111111011011110111
1011111100011101110001110111101111111111111111100011011
1001001101110011011111111011011001111111111111111011110
1001111100111011010111111111101101011011111111110101110
1001111101000111010111111011111010101111111111010011111
1101111111011000111111111100111111001101011101111010110
1110110101110001111111110011101111011001111111011011011
1001110100110011111111111111111111100111111111010010010
1111111111111110100010111011101000111111111111100011010
1000111111110001111111111110110100011110110111010111111
1011111100111011110101100111000111011111001100101111
1110110101110111001111100111001101010111111111101000011
0100111110111001111110100000110111010111110111111010101
1010111100010111111110111001111010111001110110100010111
1001111110110101000111110111110000110111011111001010111
1001111100111111010110100011110101010111011101110110
0010001100111011110111111000011011100111111110111111110
0010111110100011010111100101010111010111111110111011010
1110111100010011001110100010111011010011001100110010
1010000100111101011111110100000100010011010001101101
1001010000101101110110010001010001011111001000010100
1101110011101010000110000010010001000011111010010010

Raw Data: Responses to Form B

Items

Subjects

```
110111111111110111111111111111111011100111111011101111
101111110111110111111111100111111111111011111111011111
110011110011100111111111111111111011111111111011111111
111111111011101111111111111111011111011111111100110100
100111111011010101111111011111111011111111111101111101
110111110111111101111111011111100110111011111110110111
111111111111111011011111110101110101111101110011011O
101111111111111110111111101111010011110110011001101111
110111111110111101111111011111111010100110111010110111
101111110011111011111011111110110001111111111101110110
100111111111110111111011101110110111110011111110100101
110011110111011101111011011111110110111111111110010011
100111111111101011101111110001101010111111110101111101
100111110011110101111111100101100111100111111101111110
110111110011010101110010011111111110110111111111010110
100111110010010111111111101011101010110011111101010111
111111110011111101110011011111101000100111111110010000
000111010001001110011110101101111111001111110101010111
100010010001011101111111011111001111101111001011100
100111111001110101111110110101010101001111110100100000
100010111111000111010111011111001001001111110010101111
100110111011000101110111001110100100011101010001001O
110010011011110101111011011101100101101101010001101O
100011111101010111111111001100011001110010111001010000
111110000111110100110011110000010101001111111110100O1
110011110100110101110011111101010100101111110000010101
110110010010001110101011110001101011011001101110010101
100111111111000101010010011000011111011101101100000O1
110011011100011100010011001101000101101111110000011111
100011011100010101110110011101000100111100110001000O
```

Raw Data: Responses to Form C

Items

Subjects

```
11111111011011110111111111110101111110111101111111111
11111111011111110111101111111111111011011111111110110
11011101111111111011111110111011111111111111100110111
10111111011111111011100111111110011110011111111001111
11011111111111011110101110111111100111011101101011110
11011101001110010111111111111101001111111011111000111
11111101101101111101010100110111011111011111110011110
11011111011100010001111101111101011111111011100011111
10011011010110111111011101111110001110110110000111111
11111111011100110111111111100110001010011111110110010
00011111111100101011111111111111101101010100110100101 1
10111111011100010101011101110100011110110111010101 11
11111111011110010111101101010000111110011010010101 11
11011011111110010111011101011111100110110110100110 00
11001101101010010111111001110101101110110110111001 10110
11001111011001010101001101111111011000011011100110111
11011111000101110111010010111111110000110111111000000
10111111111001110001001101111111010100110011010001 10
10011111010000010111111101111100000100111011101101 1
10011111111000101010101100111111010101001110110010 1101
10101111011100010011101011100000010110011111101110111
10111110100010101111000100111011010001111011110000 11
10101111000010110101011001111111010001111001101010110
11001111010110011111011101100100000110100101001101 11
10011111101101010111011111100100100101110101010001 00
01111101100011110111101010110100010100101110001011 00
11011111000110110001010000011111000010111011010100 11
10101011100110110101111100111101100000101111100000 00
11001011000001111001001101110111011010010000000 10001
01100011111111011110100101101110001000001100000000100
11001011000001110101011100001100011110011000001010 00
10011001110111100111100000100001110010110000000000 11
```

Raw Data: Responses to Form D

Items

Subjects

```
10111111111111010111111111101110111011111111111111111
11111111111101110111111111111110100011111111011011111
10011111111111111111111110111111101011011101101111111
11111111111101111111011101111101101100111111101111111
11111101111111111011111110111010101110111001111011111
11011111111110111001111110111110101011111001111011111
11001111011101110111111110111111101011111011101110111
11011111111110101001111110111111111111011011001111110
10111111011111111111111111100100110100111111111000111
11011111101111111111101101111110010010110111010111111
11011111011111111010111111011111011111101110110010001
11101111111110101001101100111110111111010011111110111
11011011111110111011110110111010111011111101010011111
10011111100111010011111101111101011011111111111010110
11011111011001110011111101101110111101111111110010110
10001111011101111111110110111111011100001111111110101111
10111111111110111001111110111011000111011011101011110
10101110001010101111111101111101010110110111111011111
10101111011111110111101011110101010110111111110111000
11011111001101111111111110101110101011111111000011001
10111011111001010111111101010100111111111101000011011
10111101001111111011101110110110001011011011110011110
11011101000111100111001000000100010000110110100111001
10011011001111111111001001001101110111010110011011101
10001111101001110011010101111101111110110111010000011
10001111011001110111001101101101010100111111110010010
10001111001011110011001000110101110110110011000101111
11010011110010111111001011100001110101100010100000000
11000110001111010011111110101100000100110110000100000
10111011101011010011001001100100001100100010000000001
11011001100110100001000001001100111001111100000010100
10110001001000000000011000111100110110001010010011110
10001110010011100110100000010011000110000100101010000
```

Raw Data: Responses to Form E

Items

Subjects

1001110110110111111111101111110111111111
1101111111010011111011001010111111111111
1001110111010101111111111001111111010111
1001110110011011111011110111111111100111
1101110110110101111011111010101111110111
1111111110010011010011110111101111100111
1011110111010111111011011010001111101111
1101111111111101111011110100000111101110
1001111011010101110010011111101111011111
1111110111011101111001011101001110001111
1101110110010111111011111001011111010101
1001110111111010111101111001001111100111
1101110111111010111001110001010111001101
1101111010010111011010110011001111100111
1011110111011101111001100010011110001111
1001100111010101101011011110001110101111
1101111111011010101101010101100111000101
1010101111111010111001001011000111100110
0101000111111011100111010110111010010010
1001101011100011110011111010000111010101
0001110011000101110011111010000110101111
1111010110110101110010011000101111100000
1101110111010011010010100010011110001000
1101110111010101000010110010001110001100
0101101101010001010011010011001111110010
1001111010110101010010100010001101000011
0001001001010101110000111011000110101010
0010100001110011000000111010101110001000
1001101101010001110011001001000110000000
1011100101010001000010011010001110010000

References

Costin, F. (1970). The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 30, 353-358.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 32, 1035-1038.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. Journal of Educational Measurement, 12, 109-113.

Hambleton, R. K. (Ed.). (1983). Applications of item response theory. Vancouver, B. C..: Educational Research Institute of British Columbia.

Lord, F. M. (1977). Optimal number of choices per item - a comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Magnusson, D. (1967). Test theory. Reading, Mass.: Addison-Wesley Publishing Company.

Marascuilo, L. A. (1966). Large-sample multiple comparisons. Psychological Bulletin, 65, 280-290.

Ramos, R. A., & Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple choice items. Journal of Educational Measurement, 10, 305-310.

Serlin, R. C., & Kaiser, H. F. (1978). A method for increasing the reliability of a short multiple-choice test. Educational and Psychological Measurement, 38, 337-340.

Straton, R. G., & Catts, R. M. (1980). A comparison of two, three, and four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 40, 357-365.

Thompson, M. L., Althaus, R. A., Corbin, C. B., Gray, G. E., Sroka, S. R., & Thompson, K. G. (1983). Choosing Good Health, Glenview, Ill.: Scott, Foresman and Company.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 1, 386-391.

Williams, B. J., & Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. The Fourteenth Yearbook. National Council on Measurements Used in Education, 63-65.