THE EFFECTS OF MULTIPLE-CHOICE FORMATS AND COMPLETION FORMATS ON TEST RELIABILITY

By

ROSA TORABI-PARIZI

Bachelor in Guidance and Counseling

University of Teacher Education

Tehran, Iran

1973

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
MASTER OF SCIENCE
May, 1980

THE EFFECTS OF MULTIPLE-CHOICE FORMATS

AND COMPLETION FORMATS ON TEST

RELIABILITY

Thesis Approved:

Nana Jo Campbell
Thesis Adviser

Joseph Pearl

David W. Perrin

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

The author wishes to express her sincere appreciation to Dr. Noma J. Campbell, her thesis advisor, for her constant guidance and assistance through the author's graduate program.

Appreciation is also expressed to the other members of the committee, Dr. David W. Perrin and Dr. Joseph H. Pearl, for their helpfulness in the preparation of this thesis.

Sincerest thanks are extended to my friend Ms. Terry O'Brien for her valuable help and advice.

Sincerest respect and gratitude also is expressed to my brother Mr. Muhammadu A·Daba for his valuable assistance and moral support.

The author also expresses special thanks to her beloved parents for their worthwhile encouragement and making her graduate study possible. To her sister, Farzaneh, and to her brothers, Manuchehr, Hamid and especially Mehdi, for their loving consideration and patience during the time of this study.

TABLE OF CONTENTS

Chapter		Page
I.	INTRODUCTION	1
	Significance of the Study	4
	Statement of the Problem	5
	Definition of Terms	6
	Limitations of the Study	7
II.	REVIEW OF LITERATURE	8
	Introduction	8
	Multiple-Choice Items	8
	Completion Items	16
	Summary	18
III.	METHODOLOGY	20
	Introduction	20
	Sample	20
	The Instruments	21
	Procedure	24
	Statistical Analysis	25
	Problem One	26
	Problem Two	27
	Summary	29
IV.	ANALYSIS OF THE DATA	30
	Introduction	30
	College Level	31
	Section One	31
	Section Two	31
	Section Three	39
	Elementary Level	55
	Section One	55
	Section Two	57
	Summary	62
V.	SUMMARY AND CONCLUSIONS	66
	Introduction	66
	Interpretation of Results: College Level	67
	Multiple-Choice Forms	67

Chapter Page

Completion Forms 68
Interpretation of Results: Elementary Level 70
Multiple-Choice Forms 70
Completion Forms 70
Conclusions 70
College Level: Multiple-Choice Tests 71
College Level: Completion Tests 71
Elementary Level: Multiple-Choice Tests 72
Elementary Level: Completion Tests 72
Recommendations 72

SELECTED BIBLIOGRAPHY 75

APPENDIXES 79

APPENDIX A - COLLEGE AND ELEMENTARY MULTIPLE-CHOICE TESTS AND COMPLETION TESTS 80

APPENDIX B - INSTRUCTIONS FOR ADMINISTERING MULTIPLE-CHOICE TESTS AND COMPLETION TESTS 101

APPENDIX C - KEY RESPONSES OF COLLEGE AND ELEMENTARY MULTIPLE-CHOICE TESTS AND COMPLETION TESTS . . . 103

APPENDIX D - ITEM ANALYSIS OF COLLEGE LEVEL DATA 108

APPENDIX E - ITEM ANALYSIS OF ELEMENTARY LEVEL DATA 113

LIST OF TABLES

Table		Page
I.	Number of Students per Class at the College Level	21
II.	Means, Standard Deviation, and Standard Error of Measurement of Multiple-Choice Tests and Completion Tests--College Level	32
III.	Discrimination, Difficulty, and Internal Reliability Estimates of Multiple-Choice and Completion Tests--College Level	33
IV.	Test-Retest Reliability Estimates of Multiple-Choice and Completion Tests--College Level	36
V.	Internal Reliability Estimates for Two Levels of Vocabulary	41
VI.	Test-Retest Reliability Estimates for Two Levels of Vocabulary--College Level	43
VII.	Internal Reliability Estimates for Two Levels of Comprehension	46
VIII.	Test-Retest Reliability Estimates for Two Levels of Comprehension--College Level	48
IX.	Internal Reliability Estimates for Two Levels of Reading Rate	51
X.	Test-Retest Reliability Estimates for Two Levels of Reading Rate--College Level	53
XI.	Means, Standard Deviations, and Standard Error of Measurement of Multiple-Choice and Completion Tests--Elementary Level	56
XII.	Discrimination, Difficulty, and Internal Reliability Estimates of Multiple-Choice and Completion Tests--Elementary Level	58
XIII.	Item Difficulty and Discrimination Indices of Multiple-Choice Tests--Form A of College Level	109

Table | Page

XIV. Item Difficulty and Discrimination Indices of Multiple-Choice Tests--Form B of College Level 110

XV. Item Difficulty and Discrimination Indices of Completion Tests--Form A of College Level 111

XVI. Item Difficulty and Discrimination Indices of Completion Tests--Form B of College Level 112

XVII. Item Difficulty and Discrimination Indices of Multiple-Choice Tests--Form A of Elementary Level 114

XVIII. Item Difficulty and Discrimination Indices of Multiple-Choice Tests--Form B of Elementary Level 115

XIX. Item Difficulty and Discrimination Indices of Completion Tests--Form A of Elementary Level 116

XX. Item Difficulty and Discrimination Indices of Completion Tests--Form B of Elementary Level 117

CHAPTER I

INTRODUCTION

The assessment of students' achievement is of major concern in educational institutions. It is very important for educational personnel to develop or select precise and reliable measures of student's performance. Because of the importance of accuracy and reliability of such measurements, efforts have been focused on the development of several types of tests. However, it has attracted the attention of few educational researchers to study the reliability of different types of tests in order to help the flow of science toward a more precise measure of students' performance.

Since very little empirical research into the value of many of the item writing rules has been carried out, this study will attempt to determine the effect of varying the item form on the reliability of scores obtained by elementary level and college level students on two types of tests: multiple-choice and completion tests.

In the multiple-choice test, two types of items (three-alternative and four-alternative) will be focused on. It is generally assumed by text experts (Noll & Scannell, 1972; Thorndike & Hagen, 1977) that tests which have multiple-choice items with four alternatives are more reliable than multiple-choice items with three alternatives. This assumption may arise partially from the fact that in multiple-choice items, the examinee

will have less chance for guessing the correct answer in items with four alternatives than in items with three alternatives.

A multiple-choice item is a written item which requires the examinee to select the best answer or the corrrect answer from the alternatives suggested. This type of item has gained considerable popularity among test constructors and standard test users. It is also widely used by the classroom teachers.

Multiple-choice items are applicable to all levels of cognitive domain. Although it is claimed by some experts, such as Sax (1974), that most multiple-choice items only measure factual knowledge, it is believed by other experts, such as Remmers and Gage (1955), that multiple-choice items also measure other levels (i.e., understanding, application of principles, analysis, synthesis, and evaluation levels). Since a large number of items can be used during one period of examination, it is possible to measure attainment of several instructional objectives in one test. The scoring procedure is completely objective. The well-constructed items may be scored rapidly and accurately even by those scorers who are unqualified to teach in a subject area being examined.

As mentioned previously, the other dimension of this study is concerned with the reliability of scores on the completion items. In the completion tests, two types of items--items with the blank at the beginning and items with the blank at the end-- will be considered. It is generally stated by test experts (Nunnally, 1972; Sax, 1974) that tests having completion items with the blank at the end are more reliable than those completion items with the blank at the beginning.

This statement may partially be based on the fact that items with the blank at the beginning require more time to answer because the

examinee may have to re-read the item to comprehend what he/she is expected to do. But, when the blank appears at the end, the examinee can easily discover what the task requires.

Completion items have been widely found in workbook tests when they are accompanied with the textbooks. This type of item is useful in quantitative problem solving in mathematics or science when the results of computational process and complex reasoning can be expressed in a few symbols and simple comprehension (Thorndike & Hagen, 1977). It also can be used to measure recall of facts (Marshall & Hales, 1971).

The preferable grade level of application of this type of test is the intermediate level as explained by Marshall and Hales (1971):

> This type of test can be used at almost all grade levels, but it seems to be especially appropriate at the intermediate level since much of the material taught at this level lends itself to the completion type of examination (p. 66).

Completion items may have several blanks or one blank only. Placement of the blank at the end is more common and it is recommended by Sax (1974) to avoid ambiguity of the items.

Completion items minimize the possibility of guessing according to Marshall and Hales (1971). Construction of completion items is relatively easier than multiple-choice items. However, when measuring the higher levels of mental processes, Marshall and Hales caution that limitations of the completion item will cause some measurement problems in these domains. Because of this difficulty, it seldom is used in measuring higher-order mental processes.

Students' misinterpretation of the item is a problem that one should be aware of. A well constructed completion item will help to avoid such misinterpretation. Completion items lack scoring economy.

Sometimes students are led directly to the correct answer by such clues as the articles or even the length of the blank of an item.

Significance of the Study

The significance of this study lies in its attempt to determine the number of alternatives, three or four, per item in the multiple-choice items and the placement of the blank, beginning or end, in the completion items which will yield the more reliable scores.

Test constructors usually include four or five alternatives per item in multiple-choice test items. It is widely believed that the use of less than four alternatives per item will increase the chance of guessing the correct answer. Many test constructors point out that a different number of wrong answers should be provided to elicit wrong answers from those who have various kinds of misinformation.

However, in a suggestion for data analysis, Ebel (1963) states that if a good distractor which can attract wrong responses is provided in the items of a test, the elimination of the other distractors might not harm the discrimination power of the test very much. There are other facts to be considered. A smaller number of alternatives will facilitate the test constructor's effort in finding a plausible distractor. Also, a smaller number of alternatives will take less of the examinee's time for reading and response.

In terms of completion items, some test constructors (Wick, 1973; Sax, 1974) recommend the placement of the blank at the end of the sentence. Wick and Sax believe that if the blank appears at the end of the sentence the examinee will be able to comprehend the question before

arriving at the blank. It will also save the examinee's time for reading and response.

Because of lack of information relating to effects on the reliability of scores on the above mentioned types of items, there seems to be a need to study the reliabilities of scores obtained on these items. Thus, the purpose of this study is to determine: (1) which of the two types of multiple-choice items (items with "three" alternatives and items with "four" alternatives) yields the most reliable measure of achievement, and (2) which of the two types of completion items (items with the blank at the beginning and items with the blank at the end) yields the more reliable measure of achievement.

Statement of the Problem

This study is addressed to the question of the reliability of the instrument. The present study points to the importance of the number of options or placement of the blank in reliability of the instrument. The problem of this study is: What are the effects of number of options (three or four) in multiple-choice items and placement of the blank (beginning or end) in completion items on the test-retest reliability and internal consistency of the scores?

Also, this study is designed to determine: Are there differences in the reliabilities of the scores on the four college level tests according to the reading levels of the students? To identify the differences between reliabilities, the study will attempt to determine the effect of number of options (three or four) in multiple-choice items and the effect of placement of the blank (beginning or end) in completion

items on the test-retest reliability and internal reliability of the instrument.

Definition of Terms

Completion item is "a written item which requires the examinee to supply the correct word or short phrase in response to an incomplete sentence, a question, or a word association" (Marshall & Hales, 1971, p. 64).

Multiple-choice item is "an item consisting of a main part of a question (stem) and a number of options from which the student is to select the correct response" (Sax, 1974, p. 88).

Stem (the main part of an item) is a question or an incomplete sentence which presents the problem.

Options (alternatives) are the suggested responses to an item. Options are composed of a keyed response and distractors.

Distractors are the suggested responses which are not the correct answers.

Discrimination Index (D) "measures the extent to which an item is capable of measuring individual differences" (Sax, 1974, p. 235).

Difficulty Level (P) is "the proportion of students responding correctly to an item" (Sax, 1974, p. 239).

Reliability is the extent to which a test is consistent in measuring whatever it does measure (Sax, 1974).

Test-retest (stability) reliability describes the consistency of the examinee's score on the two performances of a test as well as the consistency of the operation of the measurements (Sax, 1974).

Internal reliability is the mean correlation among all possible pairs of the items on a single test (Sax, 1974).

Limitations of the Study

The subjects were not randomly selected. The college level subjects who enrolled in Reading and Study Skills (C&IED 1232) have had reading problems. Also, the elementary school subjects include only fifth grade classes. The generalization of the results is limited only to subjects having very specific characteristics. The validity of the four tests was not determined.

CHAPTER II

REVIEW OF LITERATURE

Introduction

The purpose of this chapter is to review research literature relevant to the purpose of this study. The literature reflects comments and studies regarding the reliability of recognition items and recall items. The studies and comments included in this chapter are separated into the two topics of concern.

Multiple-Choice Items

The studies reviewed are mostly concerned with the use of multiple-choice tests and the reliability of different types of multiple-choice items. However, in several studies the investigations center on the reliability of the different types of items having different numbers of options. There are few studies which specifically investigate differences in the reliabilities of the three- and four-option items.

Toops (1921) compared three types of examination methods. He developed a 50-item test of general information. The items were primarily constructed in forms of recall type. Then, they were revised to two different types of items--recognition type with five options and true-false type.

One hundred and twenty-four students at the Teachers College of the Columbia University participated in the study conducted by Toops. The

tests were administered in six different orders in which each order of test administration included different numbers of students. Thus, each order had a range of students from 10 to 39.

A split-half reliability coefficient was computed to determine the reliability of each test. The results revealed a range of reliability coefficient between .448 and .340. However, these coefficients were corrected by Spearman-Brown formula to determine the reliability of the whole test. It was shown that the coefficient ranged from .618 to .507 in which the recall test had the highest and the true-false test had the lowest reliability estimates. However, in conclusion, Toops (1921) stated:

> . . . (1) when equal numbers of information questions are given on the three forms of information test, the recall is always the most reliable, followed in order by recognition and true-false forms, but that (2) where equal amounts of examination time are taken on the three forms of test, the reliabilities do not differ greatly (p. 51).

In his explanation of equal time he stated that in a time required to answer one recall item 1.23 recognition item, and 1.92 true-false item may be answered. Therefore, he theoretically examined the added number of items to see whether they have any effect on the reliability of the test. However, his results yielded a range of reliabilities from .607 to .664 in which the true-false test yielded the highest and the recognition test had the lowest reliability coefficients.

In addition to the previous types of items which were examined by Toops, there are several other types of items which have attracted the attention of other experts. Among these experts, Ruch and Stoddard (1925) investigated the reliability of five different types of items: recall, five-option, three-option, two-option, and true-false items.

They selected 100 items which originated from the American history and social science areas. The items primarily were constructed in a recall type and were then converted to five-option, three-option, two-option, and true-false types. Five different types of tests were developed in which each test was broken into two equal halves to make two forms--A and B with 50 items in each.

The tests were administered to 562 twelfth grade high school students in Iowa. The sample was divided into four groups having 137 students each. At first, all 562 subjects took the two forms of the recall test; then, each group took the two forms of each of the remaining tests.

A split-half reliability coefficient which was corrected by Spearman-Brown formula was computed to determine the internal reliability for the whole test with 100 items. The results revealed a range of coefficients from .714 to .896 in which the recall test yielded the highest and the true-false test had the lowest estimate of reliability. It was concluded that the five-option test had a higher reliability estimate (.886) than the three-option test (.849). Also, the three-option test yielded a higher reliability estimate than that of the two-option test (.748).

Similarly, a correlation coefficient between the two forms A and B of each test was calculated to obtain the equivalency of the two forms of each test. The results yielded a range of coefficients from .555 to .811 in which the recall test had the highest value and the true-false test had the lowest value. It was shown that the same order of reduction of the reliability was obtained in which the five-option test had an estimate of .796, the three-option had an estimate of .598 and the two-option had .737.

Ruch, McGregor, Maupin and Murdock (1926), with the assistance of Degraff and Gordon, extended Ruch's investigation to study reliability of six types of tests. Two hundred fifty items were developed from the area of United States history. These items were reviewed by six judges in terms of the items' appropriateness for the test and their difficulty. Finally, 200 items were chosen through this process and were divided into two forms (A and B) having 100 items each.

To construct the other five types of tests, first, the recall items were converted into seven-option items. Second, the seven-option items were changed to five-option, three-option, and two-option items by randomly eliminating the extra distractor(s). The true-false items were developed by using the correct answers of the two-option test for half of the items and by using the incorrect answers for the other half of the items.

A sample of 2,453 students from Minnesota, Illinois, Iowa, Missouri, Oklahoma, Texas, Arizona, and California participated in Ruch's study. Their grades ranged from seven to twelve. The subjects were divided into 10 subgroups in which five groups were instructed to guess and the other five groups were instructed not to guess.

A reliability coefficient between two forms (A and B) of each test was computed. The coefficients ranged from .641 to .950 for those groups whose responses were not corrected for guessing. It was concluded that the recall test had the highest estimate and the true-false test had the lowest estimate. The other tests ranked in order of decreasing reliability as five-option test (.864), three-option (.837), seven-option (.800), and two-option (.745).

In 1928, Ruch, with the cooperation of Charles, re-examined the reliability of five different items. They developed a test with 100 items which originated from the test of Woodworth's psychology. Each test was divided into two equal halves. The first half (50 items) constituted Form A and the second half formed the Form B. The items were constructed primarily in the form of recall type which were later altered to five-option, three-option, two-option, and true-false type items.

A sample of 747 college students from Iowa State Teachers College was involved in the study by Ruch and Charles. At the beginning, all subjects took the recall test. Then, on the second day they were divided into four different groups ranging in size from 182 to 189. Each group was given the two forms of each type of the other four tests.

Reliability coefficients were obtained between the two forms (A and B) of each test. These coefficients ranged from .477 to .680. Also, by combining the two forms, an internal reliability was estimated for each of the tests. These estimates ranged from .646 to .809. However, the data revealed that the tests ranked in the following order of decreasing reliability: five-option test (.809), three-option test (.768), recall test (.752), true-false test (.751), and two-option test (.646).

Tversky (1964) mathematically reviewed the optimal number of alternatives per item. In his mathematical investigations, he stated that "given a fixed total number of alternatives for a multiple-choice type test, the use of three alternatives at each choice point will maximize the discrimination capacity of a test" (p. 386). He also stated that "whenever the amount of time spent on the test is proportional to its total number of alternatives, the use of three alternatives at each choice point will maximize the amount of information

obtained per time unit" (p. 390). Then, he explained that this is especially true when more time is devoted to reading the alternative and selecting the correct answer than to reading the question itself.

Ebel (1969) also theoretically examined the expected reliability of multiple-choice tests. He developed a formula which predicts the reliability of an objective test. This formula was:

$$r = \frac{K}{K - 1}\left[1 - \frac{9(N + 1)}{K(N - 1)}\right] \quad (1)$$

in which K represents the number of items in the test, N the number of options per item, and r the reliability coefficient of the test. He stated that this formula can be derived when:

> 1. A reasonable estimate of the mean score of a good objective test is a value midway on the scores scale between the maximum possible score and the expected chance score, 2. A reasonable estimate of the standard deviation of the scores on a good test is one-sixth of the difference between the maximum possible score and the expected chance score, 3. A reasonable estimate of the reliability coefficient is provided by $K\text{-}R_{21}$ formula (p. 566).

He also stated that for very short tests the estimate of standard deviation is one-third of the difference between maximum possible score and the expected chance score for tests having 10 items or less.

Based on his formula, Ebel developed a test of expected reliability estimates of multiple-choice tests with 100 items. These results revealed that the reliability of an objective test will increase when the number of options are increased from two to three. A smaller increase also is expected when four-option items are used, as well as a smaller increase for more options will be expected. He also proved that when total number of options on the test are fixed at some constant number, the three-option and two-option types of items seem to yield higher reliability than the four-option items.

By inspiring Tversky's work, Costin (1970) presented empirical support for Tversky's study. Costin's study was to gather evidence to find the optimal number of options for multiple-choice items in which he was concerned with only two types of multiple-choice items: items with four options and items with three options. He developed four tests consisting of four-option items which were randomly drawn from four different topic pools. The four tests of different topics with various items were: perception, 50; learning, 60; motivation, 60; and intelligence, 50. The items were measuring general knowledge of each topic. After the items were chosen for each test, the test was divided into half. Then, one half of each was revised to a test consisting of items having three options. The revision was made by randomly discarding a distractor from each item. The other half of each test remained with four-option items.

A sample of 207 students from the University of Illinois participated in Costin's (1970) study. All subjects were given four exams, but for the ease of data analysis seven of the subjects were randomly dropped from the study. For the data analysis, a $K\text{-}R_{20}$ formula was applied to estimate the reliability of each test. Also, difficulty and discrimination indices were obtained for his data. The results revealed that mean discrimination indices for the three-option tests were consistently higher than those for four-option tests. The reliability coefficients which were obtained ranged from .50 to .62 for the tests. It was concluded that the three-option tests had consistently higher reliability estimates than those of the four-option tests.

Grier (1975) mathematically tested the assignment of option to multiple-choice items of a test in order to determine the optimal

reliability estimate of the test. With the use of Ebel's formula, his results yielded a better reliability for the three-option items than for the four-option and five-option items. It was shown that three-option items increase the expected reliability when $n > 18$ or $(n \times a) > 54$ in which n is the number of items and a is the number of options. However, the effect of three-option items on the increase of reliability is true only when the number of test items are increased to compensate for a smaller number of options per item. He stated that when n is small, these findings will be true when the assumption of Ebel's formula can be met. Grier also pointed out that three-option items have some practical advantages in addition to the above technical ones. These include ease of construction and explicity of reading.

Grier (1976) re-examined the optimal number of options per item and the optimal number of items of a test when a fixed total time is considered. As he described, the fixed total time is composed of the time spent to read each option (t) and the time required to travel from one item to another (t'). Therefore, time required to finish the test would be:

$$T' = n't' + \sum_{i=1}^{n} ait \qquad (2)$$

in which n is the number of items and a is the number of options. This formula was mathematically improved when it was proven that $a = 2.718$ as the optimal number of options for a total fixed time. He concluded that the three options will be an optimal number of options for an item in a total fixed time as well as two options as the next best number of options for the item.

In addition to the empirical studies for the multiple-choice items, several recommendations were observed in widely used measurement tests. Among the test experts, Ruch (1929) and Remmers and Gage (1955) recommend the use of four- or five-option items as an appropriate type of item. Thorndike and Hagen (1977) also support the use of four- or five-option items and stated that "an item must have at least three answer choices to be classified as a multiple-choice item and the typical pattern is to have four or five answer choices to reduce the probability of the guessing the answer" (p. 228).

However, Noll and Scannell (1972) recommend the five-option items and suggest that "the number of choices in multiple-choice items should be at least four; the generally preferred number is five" (p. 230). They believe that by reducing the number of distractors in an item, the chance of guessing the correct answer increases. They report that while items with five options are more common in standardized tests, recently there has been some tendency toward using items with four options.

Completion Items

The literature on completion items is limited to comments and recommendations in terms of constructing such items. Investigations of reliabilities of specific kinds of completion items (items with the blank at the beginning and items with the blank at the end) have not been reported. Therefore, the following review of literature will present the comments of the test and measurement experts.

Marshall and Hales (1971), in their recommendation for construction of completion items, briefly state that "if possible when only one piece of information is requested, the item should be constructed so that the

blank occurs at the end of the sentence" (p. 68).

Nunnally (1972) states that if a blank is placed near the end of a sentence, students can read the sentence and understand the task that is required. He also explains that an item with the blank near the end of the sentence would help teachers in grading procedures.

Wick (1973), in describing different types of items, endorses the use of shorter answer questions; but he also points out that "if an item cannot be stated in a question format, and you feel you must leave a word out for completion, at least have the blank near the end of the statement" (p. 108).

Sax (1974) also advises that completion items have the blank at the end. He explains that if the blank appears at the end of the statement, it will help the student to comprehend the task which he is asked to do.

Thorndike and Hagen (1977), in their description of writing completion items, recommend placement of the blank at the end of the sentence. They point out that in such items the student will become familiar with the question before he/she arrives at the blank.

Hopkins and Antes (1978) suggest that "the omissions should be placed at or near the end of the sentence" (p. 126). They clarify this statement by explaining that if the blank is presented at the end of the sentence, the student can easily understand the question before he enters the blank. On the other hand, if the blank appears at the beginning of the sentence, two readings are needed: (1) to comprehend the statement being presented and (2) to get the answer. They also state that by requiring less time, more items can be presented.

Summary

As can be seen by the present review of literature, studies dealing with reliability of different types of items have been limited to few investigations and recommendations.

For multiple-choice items, several studies were conducted to examine the reliability of these types of items. Among them, Toops (1921), Ruch and Stoddard (1925), Ruch, Degraff and Gordon (1926), and Ruch and Charles (1928) concluded that five-option and two-option items yield higher estimates over the three-option and true-false items. Ruch, Degraff and Gordon (1926) found that the seven-option items have a similar reliability to that of five-option items.

With reference to the present study, the findings of Costin (1970) are most relevant. His data identified a superiority of three-option items over four-option items in terms of their reliability estimates.

The theoretical examinations of Tversky (1964) and Ebel (1969) present a higher reliability estimate for the three-option items when compared to the two- and four-option items. Grier (1975, 1976) re-examined Ebel's work and found a higher reliability estimate for the three-option than for the four- and five-option items.

Conversely, recommendations of various test and measurement experts were in favor of four- and five-option items. They emphasized that items having more options would reduce the effect of guessing the correct answer (Ruch, 1929; Remmers and Gage, 1955; Noll and Scannel, 1972; and Thorndike and Hagen, 1977).

The literature review of reliability of completion items was limited only to a few recommendations of test and measurement experts. They

unanimously agreed on the placement of the blank near or at the end of the question. They explained that the placement of the blank at the end would help the examinee to comprehend the question (Marshall & Hales, 1971; Nunnally, 1972; Wick, 1973; Sax, 1974; Thorndike & Hagen, 1977; and Hopkins & Antes, 1978). The studies cited in this chapter are offered as a supportive rationale for the present study. Chapter III will outline the methods and procedure applied in collecting data for this study.

CHAPTER III

METHODOLOGY

Introduction

As indicated by the review of literature, little has been done in the area of comparing the estimates of reliability of scores on multiple-choice items with three options to the estimates of reliability of scores on multiple-choice items with four-options. Also, little has been done in the area of comparing the estimates of reliability of scores on completion items with the blank at the beginning to the estimates of reliability of scores on completion items with the blank at the end. The purpose of this study is to determine the effect of three options and four options on the reliability of scores on multiple-choice items and to investigate the effect of the placement of the blank on the reliability of scores on the completion items.

Sample

The subjects used in this study are composed of two educational levels of students: elementary and college levels. Ninety-six Oklahoma State University students who enrolled in Reading and Study Skills (C&IED 1232) in the academic year 1978-79 were involved in this study. This sample included 31 freshman females, 38 freshman males, 10 sophomore females, 10 sophomore males, one junior female, three junior males, two senior females and one senior male. This sample was selected in order

to investigate the effect of the reading ability of the students on the reliability estimates of the scores on the four tests. Table I presents the characteristics of the college level sample.

TABLE I

NUMBER OF STUDENTS PER CLASS AT THE COLLEGE LEVEL

	Class			
	Freshman	Sophomore	Junior	Senior
Female	31	10	1	2
Male	38	10	3	1

n = 96.

Thirty-five participants at the elementary level were involved in the study. The students were enrolled in the fifth grade at a school located in a town in north central Oklahoma. This sample was composed of 20 females and 15 males.

The Instruments

Two sets of achievement tests were constructed by Dr. N. J. Campbell and the researcher. The construction of the tests was based on specific criteria. The items were designed to measure general knowledge and they were based on the content found in achievement test batteries such as Sequential Tests of Educational Progress (STEP).

One set of the tests was designed for use with elementary level students. The other set was designed for use with college level students. Each set of tests consists of two parallel forms, A and B. The parallel forms of each set differ only in terms of number of options in each multiple-choice item and the placement of the blank in each completion item. Each form of the test has 40 items. Twenty items are multiple-choice items: 10 items have three options and 10 items have four options. The other 20 items are completion items: 10 items with the blank at the beginning of the sentence and 10 items with the blank at the end of the sentence.

In the construction of the parallel items, the multiple-choice items having four options were changed to items having three options by randomly dropping an incorrect alternative from each item. According to Ebel (1972), to avoid the assignment of the correct answer in a specific order, the placement of the correct answer among responses should be varied randomly. Thus, the correct response has been randomly placed in each item.

In the construction of parallel completion items, items with the blank at the end were changed to items with the blank at the beginning. All items in each form were randomly assigned until a maximum of 10 items of each type were assigned to a form and then randomly ordered on each form (see Appendix A). The tests were handscored by the researcher.

In the college level forms, all items are identical except item 10 on each of the two forms (Form A and Form B). These two items differ because of a clerical error in terms of the base cost of a car. These two items are presented as follows:

_____ is the down payment required if the down payment on a car at $365 is set at 20%.

If the down payment on a car priced at $3675 is set at 20%, the amount of down payment required is _____.

Two items in the completion part of the two forms of the elementary level differ. These items are as follows:

1. ____, 2, 4, 6, 8.

2. 2, 4, 6, 8, ____.

The four items mentioned above were not omitted because of similarities in the processes used to answer the items.

Items three and nine in the multiple-choice part of one form of the elementary level tests had a printing error which resulted in presenting all incorrect responses in these items. These items and the parallel items on the other form were disregarded in the scoring process.

In this study, the instrument used for measuring the reading ability of college level students was the Nelson-Denny Reading Test, Form D. The Nelson-Denny Reading Test was constructed by M. J. Nelson and E. C. Denny to "provide a measure of three major elements of reading ability: vocabulary, comprehension, and reading rate" (Brown, 1973, p. 3). This test has been constructed in four forms: A, B, C, and D. Forms C and D follow an identical format to that of Forms A and B. All these forms are designed for use from grade 9 through 16 which are administered in one class period. Each form of the test contains 100 items to measure vocabulary and 36 items to measure reading comprehension.

The raw scores can be converted to percentile ranks, standard scores, or grade equivalents. Percentile ranks are used in this study to divide the students into two levels of low and high reading ability scores.

The reliability of the instrument was computed using split-half reliability coefficients and corrected using the Spearman-Brown formula. The estimates range from .96 to .98 for the vocabulary test and from .80 to .83 for the comprehension test. No attempt was made to develop reliabilities for the total score or reading rate. Also, the test-retest method for the two parallel forms of C and D were utilized to determine the estimate of reliability for grades 9 through 12. The test-retest reliability estimates range from .54 to .91.

Validity estimates of the instrument were determined by determining the relationship between scores on this test and scores on the Scholastic Achievement Test. The results revealed a range of coefficients from .10 to .70 and a median correlation coefficient of .40 was obtained.

Procedure

The college students were each given one set of two parallel achievement tests. Each of the college level students took one form twice in December, 1978. The time period between the two testing sessions was one to two weeks, as recommended by Issac (1972) and Thorndike (1951).

Regarding the elementary students who took the elementary level achievement test, each student took either Form A or Form B in March, 1979. Because of problems beyond the researcher's control, the elementary students were only tested one time.

Each student recorded his responses in the test booklet. Each classroom set of test booklets was accompanied with a set of procedures to guide the instructor in administering the test. This set of test

instructions is presented in Appendix B. There was no time limit for taking the tests.

The tests were handscored by the researcher. Total scores on multiple-choice for the "three-option" items and the "four-option" items were recorded for each student on each of the two tests. Also, scores on the two completion parts were recorded for each student. The keyed responses for college and elementary levels are presented in Appendix C.

The scoring system assigned was +1 for each correct response and 0 for each incorrect response. There was no correction for guessing because the students had ample time to answer the items.

Statistical Analysis

The statistical analysis was performed using facilities at the Oklahoma State University Computer Center. Computation of means, standard deviations, standard error of measurements, percentages, discrimination indices, difficulty indices, correlation coefficients, and $K\text{-}R_{20}$ coefficients yielded the basic information required for the study.

The criterion used to determine the above statistics was total scores. Since within each test performance the four types of items were considered as four separate tests, they were analyzed separately.

The computation of means and standard deviations yielded the basic information in terms of likeness and variability of the groups. The computation of standard errors of measurement yielded the basic information in terms of variability of obtained scores around the hypothesized true value of scores.

Although several techniques have been suggested to measure item discrimination, Flanagan (1939) suggests that one of the most common

techniques used is the point-biserial coefficient which reduces relatively larger amounts of time and effort for its computation. According to Lindquist (1940), this coefficient is applied when the correlation between a continuous variable and a dichotomous variable is concerned. He also defined a dichotomous variable as "one which can be classified in only two categories" (p. 241).

The computation of difficulty indices was used to estimate the extent of the difficulty of the tests. The difficulty indices are also used to measure the difficulty of the items. Sax (1974) defines this difficulty as "the proportion of students responding correctly to an item" (p. 239). Thus, the higher this proportion the easier the item is. The size of difficulty indices ranges from zero to one. The average difficulty index of the test was obtained by dividing the mean score for the whole test by the number of items.

The K-R_{20} reliability coefficient was used to determine the consistency of the instrument in terms of intercorrelations of its items and their measuring of the same trait (internal consistency). The consistency of the examinee's scores on two performances on the instrument (stability) was estimated using test-retest reliability. These computations were carried out to answer the problems stated in Chapter I.

Problem One

What are the effects of number of options in multiple-choice items and placement of the blank in completion items on the internal consistency and the test-retest reliability of the scores? Computation of four correlation coefficients yielded the test-retest reliability of the scores on the following four college level tests: three-option

test, four-option test, the test having items with the blank at the beginning, and the test having items with the blank at the end. As recommended by Wick (1973) and Sax (1974), the Pearson product-moment correlation coefficient was calculated between the two sets of scores for each test in order to obtain the stability reliability coefficients.

To investigate the other aspect of the problem which deals with the internal reliability of the instrument, the Kuder-Richardson formula 20 ($K\text{-}R_{20}$) was used to obtain the internal consistency reliability coefficient of each of the two forms of each of the four tests, or both the college and elementary levels. Even though an underlying assumption of this formula is that each item is highly correlated with every other item (Wick, 1973), Ebel (1972) reports that when the items do not vary widely in difficulty, the $K\text{-}R_{20}$ may be employed. Similarly, Nunnally (1959) states that the $K\text{-}R_{20}$ may be applied when the test is scored dichotomously (answers are either correct or incorrect).

Since the Statistical Package for the Social Sciences (SPSS) program was used to obtain the internal consistency reliability estimates, the Cronbach coefficient alpha was computed for the scores. Nie and Hull (1977) support this procedure by their statement that "if the data are in dichotomous form, alpha is equivalent to the reliability coefficient $K\text{-}R_{20}$" (p. 66). This statistical program was used to identify the items which decrease the reliability of the instrument. The criterion used to determine the internal reliability coefficients was the scores of the first test performance.

Problem Two

Is there any relationship between the reading ability of the subjects

and the reliabilities of the scores obtained using the four item types on the college level achievement tests? Each of the reading scores (vocabulary, comprehension, and reading rate) was classified as a high and low level. The test-retest reliabilities of the four tests were calculated in order to determine the estimates of the stability of the four tests when the level of reading ability is taken into consideration. Each level contained at least 30 subjects--the students making the highest and the lowest 30 scores on each part of the reading scores. For example, the test-retest reliabilities of the scores on the three-option test for students having low or high vocabulary scores were compared to the reliabilities of the four-option test for students having low or high vocabulary to determine if there are any differences between the reliabilities.

To obtain the internal reliability estimates for this problem, the subjects were classified into high and low reading ability categories using the sample's median scores as the cutting point and the internal reliability estimates were calculated for the scores on the four tests, both Form A and Form B.

Four test-retest reliability estimates and eight internal reliability estimates were obtained for the scores on the multiple-choice test using the vocabulary scores as a means of separating the readers into ability groups. This procedure was repeated using scores on the two completion tests. The same approach was carried out for the comprehension scores and the reading rate scores. Therefore, 24 test-retest reliability estimates and 48 internal reliability estimates yielded the information required for problem two.

Since the reliabilities of scores on each form of the four tests were computed using a different sample, z tests were used to determine the significant difference between the internal reliabilities of the scores on the four tests as according to Bruning and Kintz (1977). The test-retest reliability estimates were computed using the scores of the same sample. According to Morrison (1976), a confidence interval should be obtained for each test-retest reliability estimate. Then if the confidence intervals of the reliability estimates overlap, there is no significant difference between these reliability estimates. The following procedure is used to determine the confidence interval of each test-retest reliability estimate: first, an equivalent z score of the reliability estimate is obtained by means of a table of transformation of r to Z_r. Then, the following formula is applied to the obtained Z_r score to determine the confidence interval of each test-retest reliability estimate:

$$Z_r - 1.96 \times \sigma_z \leqslant z \leqslant Z_r + 1.96 \times \sigma_z \qquad (3)$$

when

$$\sigma_z = \frac{1}{\sqrt{n-3}} \qquad (4)$$

Summary

Included in this chapter is a description of the subjects and the instruments used. The achievement tests (Form A and B) were administered to 96 college and 35 elementary students. The Pearson product-moment correlation and the K-R_{20} coefficients were utilized to estimate the reliabilities of the scores on the instruments. Confidence intervals and Z tests were used to identify significant differences between reliability estimates.

CHAPTER IV

ANALYSIS OF DATA

Introduction

The purpose of this chapter is to report the findings of the study. The presentation of results is divided into two parts. The first part deals with the results of college level data and contains three sections. Reported in section one are the means, standard deviations, and standard errors of measurement. Section two deals with a part of the problem concerning the reliabilities of the scores on the four tests and includes the discrimination and difficulty indices of the four instruments. Section three contains a discussion of problem two of the study: the test-retest reliability and internal reliability of the scores obtained using the four item types when the levels of reading ability of the college students are taken into consideration.

The second part of the results presents the elementary level data. Section one of this part presents a descriptive analysis of the data: means, standard deviations, and standard errors of measurement. Section two deals with problem one--the internal reliability of the scores on the four types of items.

College Level

Section One

Mean scores were computed for both test and retest performances of each test and are presented in Table II. There are very small differences in the scores on the two types of multiple-choice test items. Also, as Table II reports, the standard deviation of the scores on multiple-choice tests are very similar. Likewise, the standard errors of measurement of the scores on multiple-choice tests are identical and there is no difference between standard errors of measurement of the scores on multiple-choice tests.

Table II also provides the means, standard deviations, and standard errors of measurement computed for the test scores on the completion tests. These data illustrate that there are very small differences in the scores on the two completion tests. Also, very small differences are observed between standard deviations of the scores on the two completion tests. The standard errors of measurement of the scores on the completion test are very similar.

Section Two

This section deals with problem one. Also indicated are the discrimination and difficulty indices of each of the four tests. These statistics were computed for each form, A and B, of each test and are presented in Table III. Forty-six students completed Form A and 50 students completed Form B.

Problem One. What are the effects of the number of options in the

TABLE II

MEANS, STANDARD DEVIATION, AND STANDARD ERROR OF MEASUREMENT OF MULTIPLE-CHOICE TESTS AND COMPLETION TESTS--COLLEGE LEVEL

Type of Test	Test			Retest		
	$\bar{x}$	SD	SEM	$\bar{x}$	SD	SEM
Multiple-Choice:						
Three-Option	8.28	1.54	.15	8.24	1.58	.16
Four-Option	7.95	1.48	.15	7.94	1.60	.16
Completion:						
Blank at the Beginning	6.10	2.08	.21	6.46	2.13	.21
Blank at the End	5.94	1.95	.19	6.22	2.01	.20

n = 96.

TABLE III

DISCRIMINATION, DIFFICULTY, AND INTERNAL RELIABILITY ESTIMATES OF MULTIPLE-CHOICE AND COMPLETION TESTS--COLLEGE LEVEL

Type of Test	No. of Items	Mean Discrimination Index	Mean Difficulty Index	Internal Reliability K-R 20	
				r	n
Multiple-Choice:					
Three-Option					
A*	10	.41	.82	.51	45
B**		.37	.83	.50	48
Four-Option					
A*	10	.34	.78	.31	45
B**		.40	.81	.52	48
Completion:					
Blank at the Beginning					
A*	10	.42	.67	.36	27
B**		.47	.64	.65	19
Blank at the End					
A*	10	.44	.61	.69	27
B**		.35	.66	.38	19

*n = 46.

**n = 50.

multiple-choice items and the placement of the blank in completion items on the test-retest reliability and internal consistency of the scores?

Multiple-Choice Forms

Internal Reliability. The K-R_{20} formula was used to obtain the internal reliability estimates of scores on the four tests. The reliability estimates of scores on multiple-choice tests range from .31 to .52.

The testing of significant difference between reliability estimate of scores on Form A of the three-option test and those of the scores on Form B of the four-option test yields a z value of .06 which is not significant at the .05 level. Also, the z value calculated between the reliability estimate of scores on Form B of the three-option test and that of the scores on Form A of the four-option test is 1.06 which is not significant at the .05 level.

As was mentioned in Chapter III of this study, the SPSS computer program used to calculate the internal reliability allows one to identify the items which reduce the reliability of the instrument. An examination of Form A of the three-option test reveals that items 1, 15, and 18 reduce the reliability of the instrument. Consequently, if item 1 is deleted from the test, the correlation coefficient would be increased to .54. Similarly, if items 15 or 18 are omitted from the instrument, the correlation coefficient would be increased to .52. The internal reliability estimate of the scores on Form A of the three-option test is .51.

In Form B of the three-option test, items 7, 8, 12, and 17 have an adverse effect on the reliability of the test. Thus, when items 7 or

17 are eliminated from the test, the reliability increased to .52. Elimination of items 8 or 12 also results in a reliability estimate of .51. The internal reliability estimate of the scores on Form B of the three-option test is .50.

Scores on Form A of the four-option test have a reliability estimate of .31. Deleting item 2 increases the reliability coefficient to .32, while deleting item 6 results in a reliability estimate of .41.

Scores on Form B of the four-option test have a reliability estimate of .52 when all 10 items are taken into account. Elimination of items 1 or 11 increases the reliability to .53, while the omission of item 13 produces a reliability estimate of .56.

Test-Retest Reliability. The test-retest reliability was calculated using the Pearson product-moment correlation coefficient. To determine this statistic, scores of each item type across both forms were analyzed. Therefore, it was possible to base these estimates on a large number of subjects (n = 96). The reliability of the scores on the three-option test is (.65), while that of the scores on the four-option test is (.72). These correlation coefficients are significant at the .001 level. These estimates are contained in Table IV. The confidence interval calculated for the reliability estimates of the scores on the three-option test is .51 to .75 and that of the four-option test is .60 to .80. No significant difference is found at the .05 level.

Discrimination. For the computation of discrimination index a point-biserial coefficient for each item was calculated. An average discrimination index was computed for each form of each test. This data is presented in Table III. The average discrimination indices of

TABLE IV

TEST-RETEST RELIABILITY ESTIMATES OF MULTIPLE-CHOICE AND COMPLETION TESTS--COLLEGE LEVEL

Type of Test	Coefficient
Multiple-Choice:	
Three-Option	.65*
Four-Option	.72*
Completion:	
Blank at the Beginning	.84*
Blank at the End	.82*

n = 96.

*$p < .001$.

the tests ranged from .34 to .41. Also, the discrimination indices computed for the items on the two forms of multiple-choice tests are presented in Tables XIII and XIV (Appendix D).

Difficulty. An examination of difficulty indices for the multiple-choice tests indicated that the average difficulty indices range from .78 to .83. These values are presented in Table III. Also, a difficulty index was computed for each item on the two forms of the multiple-choice tests. The values are presented in Tables XIII and XIV (Appendix D).

Completion Forms

Internal Reliability. The same procedures were followed to obtain internal reliability estimates, test-retest reliability estimates, discrimination indices, and difficulty indices for the completion test as were detailed in the previous section. An examination of internal reliability indicated that the internal reliability estimates of the scores on the completion tests range from .36 to .69 (see Table III).

The testing of significant difference between reliability estimate of scores on Form A of the test with items having the blank at beginning and that of Form B of the test with items having the blank at the end yields a z value of .07 which is not significant at the .05 level. Similarly, the z value computed between the reliability estimate of scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is .22, which is not significant at .05 level.

The items which reduce the reliability were identified. The items

3, 14, 15, 16, and 19 were identified as items which decrease the reliability of scores on Form A of the test with the items having the blank at the beginning. The reliability computed for scores on this test is .36. The elimination of item 3 results in a slight increase of reliability to .37. The omission of item 19 results in a reliability of .40, while elimination of items 15 or 16 result in a reliability estimate of .41. The deletion of item 14 increases the reliability of the scores to .46. Scores on Form B of the same test have a reliability estimate of .65, but this reliability estimate would be increased to .67 if items 6 or 17 are excluded from the test. The reliability estimates would be increased to .69 if item 11 is deleted from the test.

Scores on Form A of the test with items having the blank at the end have a reliability estimate of .69. The deletion of item 6 results in a reliability estimate of .72. Scores on Form B of the test with items having the blank at the end have a reliability of .38. Elimination of item 13 produces a reliability estimate of .39 and omission of item 14 results in a reliability of .47. Deleting item 15 or 16 increases the reliability to .40.

Test-Retest Reliability. Test-retest reliability was examined to evaluate the stability of the completion tests. Scores on the test with items having the blank at the beginning have a reliability estimate of .84. The scores on the test with items having the blank at the end have a reliability estimate of .82 (see Table IV). These correlation coefficients are significant at the .001 level. A confidence interval found for the reliability estimate of the scores on the test with items having the blank at the beginning is .74 to .89 and that of the test with items having the blank at the end is .75 to .88. No

significant differences are found between the test-retest reliabilities of scores on the two types of completion tests at the .05 level.

Discrimination. Computation of average discrimination indices was carried out for each item of each completion test. As is presented in Table III, these indices range from .35 to .47. In addition to the average discrimination indices of the completion tests, a discrimination index for each item was calculated and are presented in Tables XV and XVI (Appendix D).

Difficulty. An examination of average difficulty of the completion tests reveals a range from .61 to .67. A difficulty index for each item of the completion tests was computed and is presented in Tables XV and XVI (Appendix D).

Section Three

In this section problem two of the study is discussed. A comparison of the reliabilities of the scores obtained using the four item types when the reading levels of the college subjects were taken into consideration is reported.

Problem Two. Is there any relationship between the reading ability of the subjects and the reliabilities of the scores on the college level achievement test?

To investigate this problem, estimates of internal reliability and test-retest reliability were obtained for the scores obtained on each of the four types of tests for the students identified as having high or low scores on the vocabulary, comprehension, and reading rate

subtests on the Nelson-Denny Reading Test.

Effect of the Students' Vocabulary Level on the Reliability of the Scores

The vocabulary part of the reading score was divided into levels of high and low using the median point of the subjects' vocabulary scores. For each of the two levels of vocabulary, an internal reliability estimate was computed for each form of the test. Table V presents the internal reliability estimates for scores on Form A and Form B of the four tests at the levels of how and high vocabulary. The internal reliability estimates of the scores obtained using multiple-choice tests range from .05 to .48.

When low subjects' level of vocabulary is taken into account, an examination of the significant difference between reliability estimate of scores on Form A of the three-option test and that of Form B of the four-option test yields a z value of .71 which is not significant at the .05 level. Also, when low subjects' level of vocabulary is taken into consideration, the z value calculated between the reliability estimate of scores on Form B of the three-option test and that of Form A of the four-option test is .88 which is not significant.

When high subjects' level of vocabulary is taken into account, the z value computed between reliability estimate of scores on Form A of the three-option test and that of Form B of the four-option test is 1.23 which is not significant at the .05 level. The z value calculated between the reliability estimate of scores on Form B of the three-option and that of Form A of four-option is .17 which is not significant at .05 level, when high vocabulary level of subjects is taken into account.

TABLE V

INTERNAL RELIABILITY ESTIMATES FOR TWO LEVELS OF VOCABULARY

Type of Test	Vocabulary Levels			
	Low*	High*	Low**	High**
	Form A		Form B	
Multiple-Choice:				
Three-Option	.37 (n=22)	.41 (n=23)	.48 (n=24)	.24 (n=24)
Four-Option	.24 (n=22)	.29 (n=23)	.16 (n=24)	.05 (n=24)
Completion:				
Blank at the Beginning	.02 (n=14)	.55 (n=13)	.50 (n=9)	.72 (n=10)
Blank at the End	.62 (n=14)	.64 (n=13)	.14 (n=9)	.52 (n=10)

*n = 23.

**n = 25.

The comparisons of the test-retest reliability estimates of the scores of the subjects having different levels of vocabulary were investigated across the two forms of each test. Since the Pearson product-moment correlation coefficient requires 30 subjects (Sax, 1974) to achieve stability, the highest and lowest 33 vocabulary scores were designated to identify the low and high levels of vocabulary.

As indicated in Table VI, the test-retest reliability estimates of the scores on the multiple-choice items range from .51 to .77. When low vocabulary level of subjects is taken into account, the confidence interval found for the reliability estimate of scores on the three-option test is .21 to .72 and that of the four-option test is .39 to .80. Also, when high level of subjects' vocabulary is taken into consideration, the confidence interval found for the reliability estimate of the scores on the three-option is .59 to .88 and that of the four-option is .33 to .78. At the .05 level, there are no significant differences between the reliabilities of the scores on the two types of multiple-choice tests when subjects' level of vocabulary is taken into account.

The internal reliability estimates of the scores on the completion tests range from .02 to .72 and are presented in Table V. When low level of subjects' vocabulary is taken into consideration, the calculated z value between the reliability estimate of scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .23 which is not significant at the .05 level. Also, the computed z value between reliability estimate of scores on Form B on the test with items having the blank at the beginning and that of Form A on the test with items

TABLE VI

TEST-RETEST RELIABILITY ESTIMATES FOR TWO LEVELS OF VOCABULARY--COLLEGE LEVEL

Type of Test	Vocabulary Levels[a]	
	Low	High
Multiple-Choice:		
Three-Option	.51*	.77*
Four-Option	.64*	.60*
Completion:		
Blank at the Beginning	.83*	.73*
Blank at the End	.74*	.86*

[a] $n = 33$.

*$p < .001$.

having the blank at the end is .34 which is not significant at the .05 level when low vocabulary level of subjects is taken into account.

When high level of subjects' vocabulary is taken into consideration, the z value obtained between the reliability estimate of scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .08 which is not significant at .05 level. Similarly, when high level of subjects' vocabulary is taken into account, the estimated z value between reliability estimate of the scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is .30 which is not significant at .05 level.

The test-retest reliability estimates of the scores obtained using the completion tests range from .73 to .86 (see Table VI). The coefficients are significant at .001 level. When low vocabulary level of subjects is taken into consideration, the confidence interval found for the reliability estimate of the scores on the test with items having the blank at the beginning is .68 to .91 and the confidence interval found for the test with items having the blank at the end is .54 to .86. Also, when high level of subjects' vocabulary is taken into account, the confidence interval found for the reliability estimate of the scores on the test with items having the blank at the beginning is .52 to .85 and that of the test with items having the blank at the end is .73 to .93. At the .05 level, no significant differences are found between the test-retest reliabilities of the scores on the two completion tests when the vocabulary level of subjects are taken into consideration.

Effect of the Students' Comprehension Level on the Reliability of the Scores

For Form A of the multiple-choice tests, the comprehension reading scores were divided into two levels of high and low at the median scores of the sample. Since three students scored at the median, these three students were dropped from this part of the study. Thus, 21 students were identified as having a low level of comprehension and 22 students were identified as having a high level of comprehension. The internal reliability estimates of the two forms of the four tests were computed for the scores of the students at each level of comprehension. The comprehension scores of Form B of the multiple-choice tests were cut at the median point in which the middle six scores were omitted from the sample to develop more different levels of low and high. The low level of comprehension consisted of 24 and the high level of comprehension consisted of 20 subjects.

Scores on the multiple-choice tests considering two levels of comprehension have a range of reliability coefficients from -.20 to .54 (see Table VII). When low comprehension level of subjects is taken into account, the z value calculated between the reliability estimate of scores on Form A of the three-option test and that of Form B of the four-option test is .37 which is not significant at the .05 level. The obtained z value between the reliability estimate of scores on Form B of the three-option test and that of Form A of the four-option test is .90 which is not significant at the .05 level, when low level of subjects' comprehension is taken into account.

When high level of subjects' comprehension is taken into consideration, the computed z value between reliability scores on Form A of the

TABLE VII

INTERNAL RELIABILITY ESTIMATES FOR TWO LEVELS OF COMPREHENSION

Type of Test	Comprehension Levels			
	Low[a]	High[b]	Low*	High**
	Form A		Form B	
Multiple-Choice:				
Three-Option	.39 (n=20)	.54 (n=22)	.43 (n=23)	.31 (n=19)
Four-Option	.16 (n=20)	.23 (n=22)	.49 (n=23)	-.20 (n=19)
Completion:				
Blank at the Beginning	.37 (n=12)	.17 (n=13)	.47 (n=9)	.51 (n=9)
Blank at the End	.66 (n=12)	.69 (n=13)	.39 (n=9)	-.07 (n=9)

[a] n = 21.

[b] n = 22.

*n = 24.

**n = 20.

three-option test and that of Form B of the four-option test is 2.38 which is significant at the .05 level. However, the z value between the reliability estimate of scores on Form B of the three-option test and that of Form A of four-option test is .25 which is not significant at .05 level, when high level of subjects' comprehension is taken into account. As it can be observed at the high level of comprehension, the test yielded a negative reliability. It seems that this reducing resulted from the dispersion of test scores at this level at which the four-option test yielded a very low standard deviation estimate.

The test-retest reliability of the scores of the subjects having different levels of comprehension was employed for the data across the two forms of each test to increase the number of subjects (low level had 37 and the high level had 38 subjects). The reliability coefficients of scores on multiple-choice tests range from .52 to .65 (see Table VIII). All coefficients were found to be significant at the .001 level. When low level of subjects' comprehension is taken into account, the confidence interval calculated for the reliability of scores on the three-option is .42 to .80 while that of the four-option test is .39 to .79. When high level of subjects' comprehension is taken into consideration, the confidence interval calculated for the reliability of scores on the three-option test is .24 to .72 and that of the four-option test is .29 to .74. At the .05 level, no significant differences are found between the reliabilities of the scores on the two multiple-choice tests when comprehension levels of subjects are taken into account.

The internal reliabilities for completion tests at the two levels of comprehension were computed for each form of each test (Table VII).

TABLE VIII

TEST-RETEST RELIABILITY ESTIMATES FOR TWO LEVELS OF COMPREHENSION--COLLEGE LEVEL

Type of Test	Comprehension Levels	
	Low	High
Multiple-Choice:		
Three-Option	.65* (n=37)	.52* (n=38)
Four-Option	.63* (n=37)	.56* (n=38)
Completion:		
Blank at the Beginning	.86* (n=37)	.79* (n=38)
Blank at the End	.75* (n=37)	.88* (n=38)

*$p < .001$.

The reliability coefficients of scores on the completion tests range from -.07 to .69. When low level of subjects' comprehension is taken into consideration, the z value obtained between the reliability estimate of scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .04 which is not significant at the .05 level. Similarly, when low level of subjects' comprehension is taken into account, the calculated z value between the reliability estimate of scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is .53 which is not significant at the .05 level.

Also, when high comprehension level of subjects is taken into account, the z value obtained between reliability estimate of scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .47 which is not significant at the .05 level. The calculated z value between reliability estimate of scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is .55 which is not significant at the .05 level, when high level of subjects' comprehension is taken into account.

The completion tests also have a range of test-retest reliability from .75 to .88 (Table VIII). The coefficients are significant at .001 level. When low level of subjects' comprehension is taken into account, the confidence interval obtained for the reliability estimate of the scores on the test with items having the blank at the beginning is .74 to .92, while that of the test with items having the blank at

the end is .57 to .86. When high level of subjects' comprehension is taken into consideration, the confidence interval found for the reliability estimate of scores on the test with items having the blank at the beginning is .63 to .89 and that of the test with items having the blank at the end is .78 to .94. At the .05 level, no significant differences are found between the reliability of scores on the two completion tests when the subjects' level of comprehension is taken into account.

Effects of the Students' Reading Rate Level on the Reliability of the Scores

The reading rate part of reading scores also was divided into two levels of high and low by cutting the scores at the median point in Form A of the tests. Because three students had the same scores at this point, those three students were eliminated from the two levels of reading rate scores in order to clearly differentiate between the two levels. Thus, the low level of reading rate contained 21 students and the high level of reading rate contained 22 students. Then for each form of the test at each level of reading rate, an internal reliability coefficient was computed. Form B of the test also was cut at the median point to get the two levels of high and low. Each level consisted of 25 students.

Scores on the multiple-choice tests had a range of reliability coefficients from .26 to .66 (Table IX). When low level of subjects' reading rate is taken into account, the z value computed between the reliability estimate of scores on Form A of the three-option and that of Form B of the four-option is .10 which is not significant at the .05 level. Also,

TABLE IX

INTERNAL RELIABILITY ESTIMATES FOR TWO LEVELS OF READING RATE

Type of Test	Reading Rate Levels			
	Low[a]	High[b]	Low*	High*
	Form A		Form B	
Multiple-Choice:				
Three-Option	.29 (n=21)	.65 (n=21)	.36 (n=24)	.60 (n=24)
Four-Option	.38 (n=21)	.30 (n=21)	.26 (n=24)	.66 (n=24)
Completion:				
Blank at the Beginning	.59 (n=12)	.16 (n=14)	.44 (n=6)	.55 (n=13)
Blank at the End	.50 (n=12)	.79 (n=14)	.53 (n=6)	-.40 (n=13)

[a] n = 21.

[b] n = 22.

*n = 25.

the z value obtained between the reliability estimate of scores on Form B of the three-option test and that of Form A of the four-option test is .07 which also is not significant at the .05 level, when low level of subjects' reading rate is taken into account.

When high level of reading rate is taken into consideration, the z value obtained between the reliability estimate of the scores on Form A of the three-option test and that of Form B of the four-option test is .05 which is not significant at the .05 level. Similarly, the computed z value between the reliability estimate of the scores on Form B of the three-option test and that of Form A of the four-option test is 1.19 which is not significant at the .05 level, when high reading rate level of subjects is taken into account.

The test-retest reliability coefficients were computed for the scores on the four tests when the two reading rate levels of subjects were taken into consideration. The investigation was carried out across the two forms of each test to furnish a larger number of sample. The highest 31 and the lowest 31 scores on the reading rate part were chosen to develop the two levels of high and low level of reading rate.

For the multiple-choice tests, the reliability estimates range from .49 to .84 (Table X). The reliability of scores on the three-option test when low level of reading rate is taken into consideration is significant at the .002 level and all other coefficients are found to be significant at the .001 level (Table X). When low subjects' level of reading rate is taken into consideration, the confidence interval found for the reliability estimate of the scores on the three-option test is .16 to .72, while that of the four-option test is .25 to

TABLE X

TEST-RETEST RELIABILITY ESTIMATES FOR TWO LEVELS OF READING RATE--COLLEGE LEVEL

Type of Test	Reading Rate	
	Low	High
Multiple-Choice:		
Three-Option	.49* (n=31)	.75** (n=31)
Four-Option	.56** (n=31)	.84** (n=31)
Completion:		
Blank at the Beginning	.91** (n=31)	.77** (n=31)
Blank at the End	.75** (n=31)	.83** (n=31)

*$p < .002$.

**$p < .001$.

.76. When the high level of subjects' reading rate is taken into account, the confidence interval found for the reliability estimate of the scores on the three-option test is .54 to .87 and that of the four-option test is .69 to .92 At the .05 level, there is no significant difference between reliabilities of scores on the two multiple-choice tests, when reading rate level of subjects is taken into account.

For the scores on the completion tests, the internal reliability coefficients were computed for either low or high reading rate level of the subjects. As Table IX presents, the reliability coefficients of scores on the completion tests range from -.40 to .79. When low level of subjects' reading rate is taken into account, the z value calculated between the reliability estimate of the scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .13 which is not significant at the .05 level. Also, the z value computed between the reliability estimate of Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is .11 which is not significant at the .05 level, when low level of subjects' reading rate is taken into account.

When high level of subjects' reading rate is taken into account, the obtained z value between reliability estimate of the scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is 1.34 which is not significant at the .05 level. The calculated z value between reliability estimate of the scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is 1.03 which is not significant

at the .05 level. As can be observed, the test with items having the blank at the end at the high level of reading rate yielded a very low estimate of reliability.

The test-retest reliability coefficients computed for the scores on the completion tests have a range of reliability coefficients from .75 to .91 (Table X). The reliability coefficients are found to be significant at the .001 level. When low reading rate level of subjects is taken into account, the confidence interval found for the reliability of scores on the test with items having the blank at the beginning is .82 to .96, while that of the test with items having the blank end is .54 to .87. When high level of subjects' reading rate is taken into account, the confidence interval found for the reliability estimate of the scores on the test with items having the blank at the beginning is .57 to .88 and that of the test with items having the blank at the end is .67 to .91. At the .05 level, no significant differences are found between the reliabilities of the scores on the two completion tests, when reading rate of subjects is taken into consideration.

Elementary Level

Section One

The data collected for the elementary level subjects are presented in this section. Means, standard deviations, and standard errors of measurement were computed for the four tests.

The means obtained for the four tests are contained in Table XI. The means computed for the multiple-choice tests reveal that there are very small differences between scores on the two types of multiple-choice

TABLE XI

MEANS, STANDARD DEVIATIONS, AND STANDARD ERROR OF MEASUREMENT OF MULTIPLE-CHOICE AND COMPLETION TESTS--ELEMENTARY LEVEL

Type of Test	$\bar{x}$	SD	SEM
Multiple-Choice:			
Three-Option	7.28	2.16	.36
Four-Option	6.71	2.20	.37
Completion:			
Blank at the Beginning	6.34	2.55	.43
Blank at the End	6.05	2.28	.38

n = 35.

tests. The standard deviations computed for the multiple-choice tests showed that the standard deviations of the two multiple-choice tests are very similar. The standard errors of measurement was computed for multiple-choice tests and are almost identical.

Means computed for the completion tests indicate that the means of the two completion tests are very similar (Table XI). Also, there are very small differences between the standard deviations of the two completion tests. The standard errors of measurement obtained for the completion tests are almost identical.

Section Two

This section deals with problem one of the instrument reliability. However, it should be noted that because of some administration problems the second test performance was not carried out. Therefore, the test-retest reliability estimates of the instrument were not obtained. This section also reports the discrimination and difficulty indices of the four tests.

Problem One: What are the effects of the number of options in multiple-choice items and the placement of the blank in completion items on the internal consistency of the scores?

Multiple-Choice Tests

Internal Reliability. The internal reliability of each form of each test was computed using the K-R_{20} formula. The internal reliability of scores on multiple-choice tests range from -.24 to .76 (Table XII). The computed z value between the reliability estimate

TABLE XII

DISCRIMINATION, DIFFICULTY, AND INTERNAL RELIABILITY ESTIMATES OF MULTIPLE-CHOICE AND COMPLETION TESTS--ELEMENTARY LEVEL

Type of Test	No. of Items	Discrimination Mean	Difficulty Mean	Internal Reliability K-R 20 r	n
Multiple-Choice:					
Three-Option					
A*	9	.48	.87	-.24	12
B**		.50	.81	.55	18
Four-Option					
A*	9	.48	.76	.61	12
B**		.54	.80	.76	18
Completion:					
Blank at the Beginning					
A*	10	.43	.69	.61	13
B**		.57	.66	.68	14
Blank at the End					
A*	10	.27	.70	.14	13
B**		.45	.63	.71	14

*n = 14.

**n = 21.

of scores on Form A of the three-option test and that of Form B of the four-option is 2.94 which is significant at .01 level. However, the obtained z value between reliability estimate of scores on Form B of the three-option test and that of Form A of the four-option test is .21 which is not significant at .05 level.

For Form A of the three-option test, items 8, 9, 13, and 16 were identified as items which lower the reliability of the test scores. If item 8 is deleted from the test, the test reliability increases to -.23. Deletion of item 9 results in a reliability estimate of -.07, while omitting item 13 or 16 results in a reliability estimate of 0. The reliability of the complete test is -.24.

In Form B of the three-option test, items 2 and 4 have a negative effect on the reliability. Elimination of item 2 results in a reliability estimate of .67, while omission of item 4 yields a reliability estimate of .59. The reliability of the test is .55.

Scores on Form A of the four-option test have a reliability coefficient of .61. Deleting items 5, 7, or 18 increases the reliability coefficient to .62, while deleting item 6 increases the reliability to .71.

Scores on Form B of the four-option test have a reliability estimate of .76. The omission of item 1 increases the test reliability to .78, while elimination of item 11 increases the test reliability to .76. Also, omission of item 16 improves the test reliability to .80.

Discrimination. The discrimination indices for test items were computed by using the point-biserial correlation coefficient. The average discrimination index of each test was determined by computing the average discrimination indices of items of that test. The average

discrimination indices of the multiple-choice tests range from .48 to .54 and are reported in Table XII.

A discrimination index was computed for each item of each test. Items on the three-option test have a range of item discrimination indices from -.03 to .95. The indices are presented in Tables XVII and XVIII of Appendix E.

The item discrimination indices for the four-option items range from -.13 to .83. Tables XVII and XVIII of Appendix E list the values.

Difficulty. The average difficulty indices which were computed for the multiple-choice tests range from .76 to .87 (see Table XII). Also, difficulty index was obtained for each item of each multiple-choice test. These indices are reported in Tables XVII and XVIII of Appendix E.

Completion Forms

Internal Reliability. The same statistical procedure as that used with the scores on the multiple-choice tests was applied to determine the internal reliability coefficients of completion tests. The reliability coefficients of completion tests range from .14 to .71 (Table XII). The z value calculated between reliability estimate of scores on Form A of the test with items having the blank at the beginning and that of Form B of the test with items having the blank at the end is .40 which is not significant at the .05 level. Also, the obtained z value between reliability estimate of scores on Form B of the test with items having the blank at the beginning and that of Form A of the test with items having the blank at the end is 1.57 which is not significant at the .05 level.

Through the application of the procedure to obtain the reliability estimates, the items which lower the internal reliability of the scores were also identified. In Form A of the test with items having the blank at the beginning, items 1, 10, 16, and 18 were identified as items that reduced the test reliability. Deleting item 1 increases the test reliability to .65, while discarding item 10 increases the test reliability to .62. The reliability of the scores on the test is .61.

For Form B of the test with items having the blank at the beginning, elimination of items 2, 6, or 17 increases the test reliability to .68, while omission of item 3 increases the test reliability to .77. The test score reliability is .68 when all items are taken into account.

Scores on Form A of the test with items having the blank at the end yields a reliability estimate of .14. Deleting item 2 results in a reliability estimate of .35, while discarding item 3 improves the reliability estimate to .36. Also, elimination of item 6 or 7 increases the reliability to .15, while omission of item 20 increases the reliability to .23.

The scores on Form B of the test with items having the blank at the end have a reliability estimate of .71. Deleting item 10 increases the reliability to .73, while omitting item 11 improves the reliability to .75. Elimination of item 1 or 13 results in a reliability estimate of .72.

Discrimination. The same procedure was applied to obtain the discrimination indices of multiple-choice tests was carried out to determine the average discrimination indices of the completion tests. The estimated indices of these tests are reported in Table XII. The

discrimination indices of the completion tests range from .27 to .57. A discrimination index was computed for each item of each test. The item discrimination indices of the test with items having the blank at the beginning range from -.01 to .86. Also, the item discrimination indices obtained for the test with items having the blank at the end range from -.38 to .79. Tables XIX and XX of Appendix E provide these indices.

Difficulty. The average difficulty was computed for each form of each completion test by application of the same procedure which was applied for the multiple-choice test and are reported in Table XII. The difficulty indices of the completion tests range from .63 to .70. In addition to these test difficulty indices, the item difficulty indices were computed for each item on each form of the completion tests and are reported in Tables XIX and XX of Appendix E. For the test with items having the blank at the beginning, these indices range from .19 to .95.

For the test with items having the blank at the end, the item difficulty indices range from .30 to 1.00. Tables XIX and XX of Appendix E list these values.

Summary

The findings using the college data in the present study revealed that internal reliability estimates of scores on the multiple-choice tests range from .31 to .52. There are no significant differences between the reliability estimates ($\alpha = .05$). The test-retest reliability estimates range from .65 to .72. No significant differences were found

between the test-retest reliability estimates of scores on the multiple-choice test ($\alpha = .05$).

Scores on the college level completion tests have a range of internal reliability estimates from .36 to .69 in which no significant differences were found between the internal reliability estimates ($\alpha = .05$). The test-retest reliability of scores on the completion tests range from .82 to .84. These test-retest reliability estimates do not differ significantly ($\alpha = .05$).

In the study of relationships between test reliabilities and reading abilities, the internal reliabilities of scores on the college level multiple-choice tests range from .05 to .48 at two subjects' levels of low and high vocabulary. There are no significant differences between the internal reliabilities of the scores on the two college level multiple-choice tests ($\alpha = .05$). The test-retest reliability estimates of scores on the multiple-choice tests range from .51 to .77 when two subjects' levels of vocabulary are taken into consideration. No significant differences were found between these reliability estimates ($\alpha = .05$).

Scores on the college level completion tests have a range of internal reliability from .02 to .72 at the two subjects' level of vocabulary. No significant differences were found between reliability estimates of the completion tests when vocabulary level of subjects is taken into account ($\alpha = .05$). The test-retest reliabilities computed for scores on the completion tests at two vocabulary levels of subjects range from .73 to .86. The reliabilities do not differ significantly ($\alpha = .05$).

At the two subjects' levels of comprehension, the internal reliabilities of scores on the college level multiple-choice tests range from -.20 to .54. The test-retest reliability estimates of scores on multiple-choice tests at the two subjects' levels of comprehension range from .52 to .65. The reliabilities do not differ significantly ($\alpha = .05$).

The internal reliability and test-retest reliability coefficients also were computed for the completion test scores made by the college students identified as low and high levels of comprehension. Scores on the completion tests have a range of internal reliability estimates from -.07 to .69. The test-retest reliability coefficients of completion tests range from .75 to .88. The reliabilities to not differ significantly ($\alpha = .05$).

When the college level subjects are stratified using the two levels of reading rate, scores on the multiple-choice tests have a range of internal reliability coefficients from .26 to .66. The test-retest reliability coefficients range from .49 to .84. There are no significant differences between the reliabilities ($\alpha = .05$).

Also, internal reliability and test-retest reliability coefficients were computed for scores on the college level completion tests at the two subjects' levels of high and low reading rate. Scores on the completion tests have a range of internal reliability coefficients from -.40 to .79. The test-retest reliability coefficients range from .75 to .91. No significant differences were found between the reliabilites ($\alpha = .05$).

The results of the analysis of the elementary school data reveal a range of internal reliability coefficients from -.24 to .76 for the

scores on the multiple-choice tests. The internal reliability coefficients computed for the completion tests range from .14 to .71.

CHAPTER V

SUMMARY AND CONCLUSIONS

Introduction

This study examines the effects of number of options in multiple-choice items and the placement of the blank in completion items on the reliabilities of college level and elementary level instruments. Estimates of internal reliability were obtained to provide the data concerning the reliabilities of both educational levels of tests. Test-retest reliability estimates were also calculated for the college level tests.

This investigation also investigates the relationship between reading ability of the examinee and reliability of the scores on the college level achievement test. The subjects were designated as high or low in ability in three areas of reading according to their vocabulary, comprehension, and reading rate scores on the Nelson-Denny Reading Test. Measures of internal reliability and test-retest reliability were calculated using the scores on each of the multiple-choice and completion tests for the designated reading level groups in order to examine the relationship between instrument reliability and reading ability scores.

Interpretation of Results: College Level

Multiple-Choice Forms

Problem One

What are the effects of the number of options in multiple-choice items on the internal consistency and test-retest reliability of the scores?

An examination of reliability coefficients reveals a range of internal reliability estimates from .31 to .52 on the multiple-choice tests. There are no significant differences between the reliability of scores on the multiple-choice tests ($\alpha = .05$).

The test-retest reliability estimates reveal a range of reliability estimates from .65 to .72 for the scores on the multiple-choice tests. No significant differences are found between the test-retest reliability estimates of the scores on the multiple-choice tests ($\alpha = .05$).

Problem Two

Is there any relationship between the reading ability of the subjects and the reliabilities of the scores on the college level achievement test?

This problem was investigated by classifying each subject's reading scores (vocabulary, comprehension, and reading rate) as low or high. Then for each multiple-choice test, internal reliability and test-retest reliability coefficients were computed for the scores of the subjects at each level of reading.

The internal reliability coefficients of the scores on the multiple-choice tests range from -.20 to .65. No significant differences are

found between the reliabilities of scores on the four-option tests and those of the three-option tests when the reading level of subjects are taken into consideration ($\alpha = .05$).

The test-retest reliability estimates of scores on the multiple-choice items range from .48 to .84 for the subjects classified as high or low in vocabulary, comprehension and reading rate. The coefficients were obtained by combining the scores across the two forms of each test.

No significant differences are found between the reliabilities of scores on the four-option tests and those of the three-option tests when reading level of subjects are taken into account ($\alpha = .05$).

Completion Forms

Problem One

What are the effects of the placement of the blank in completion items on the internal consistency and test-retest reliability of the scores?

An examination of internal reliability of the scores on the completion tests reveals a range of reliability estimates from .36 to .69. No significant differences are found between the reliabilities of the scores on the two completion tests ($\alpha = .05$).

The test-retest reliability estimates of completion tests are .82 and .84. These estimates were computed by combining scores across the two forms of each test. Reliability of the scores on the test with items having the blank at the beginning do not differ significantly from that of the test with items having the blank at the end ($\alpha = .05$).

Problem Two

Is there any relationship between the reading ability of the subjects and the reliabilities of the scores on the college level achievement test?

To investigate this problem, each of the subjects three reading scores (vocabulary, comprehension, and reading rate) was classified as high or low. Then, the internal reliability and test-retest reliability coefficients of the scores on each completion test were calculated for the subjects at each level of reading scores.

The internal reliability coefficients computed for scores on the completion tests range from -.40 to .79 when the reading ability of the students is taken into consideration. No significant differences between the reliabilities of the scores on the two types of completion tests are found when the reading level of subjects are taken into account ($\alpha = .05$).

The test-retest reliability of scores on the completion tests range from .73 to .91. The reliability coefficients were computed after combining the scores across the two forms of each test. No significant differences are found between the reliability of scores on the two completion tests when reading level of subjects is taken into account ($\alpha = .05$).

Interpretation of Results: Elementary Level

Multiple-Choice Forms

Problem One

What are the effects of the number of options in multiple-choice items on the internal consistency of the scores?

The internal reliability estimates range from -.24 to .76 on the multiple-choice tests. The scores on Form A of the three-option tests have a negative reliability estimate. There is a significant difference between reliability of scores on Form A of the three-option test and that of Form B of the four-option test ($\alpha = .01$). However, there is no significant difference between reliability of scores on Form B of the three-option test and that of Form A of the four-option test ($\alpha = .05$).

Completion Forms

Problem One

What are the effects of the placement of the blank in completion items on the internal consistency of the scores?

An examination of internal reliability of completion tests reveals a range of reliability coefficients from .14 to .71. No significant differences are found between the reliabilities of scores on the two types of completion tests ($\alpha = .05$).

Conclusions

Within the limitations of the present study, the results have led

to the following conclusions concerning the reliability of college level achievement tests and of the elementary level achievement tests.

College Level: Multiple-Choice Tests

1. The results to not demonstrate that the scores on the three-option test are more reliable than the scores on the four-option test nor is the reverse demonstrated.
2. No significant differences are found between the test-retest reliability of scores on the four-option test and those of the three-option test ($\alpha = .05$).
3. When high level of subjects' comprehension is taken into account, there is a significant difference between internal reliability of scores on Form A of the three-option test and that of Form B of the four-option test ($\alpha = .05$). However, no significant differences are found between internal reliabilities of scores on the two multiple-choice tests when the two levels of vocabulary and reading rate and low level of comprehension are taken into consideration ($\alpha = .05$).
4. The test-retest reliability estimates of scores on the four-option tests did not differ significantly from those of the three-option tests when reading level of subjects is taken into consideration ($\alpha = .05$).

College Level: Completion Tests

1. No significant differences between the internal reliabilities of scores on the two completion tests are found ($\alpha = .05$).
2. The test-retest reliability coefficients do not indicate that

scores on the test with items having the blank at the beginning differ significantly from those of the test with items having the blank at the end ($\alpha = .05$).

3. No significant differences are found between the internal reliabilities of the scores on the two types of completion tests when reading level of subjects is taken into consideration ($\alpha = .05$).
4. The test-retest reliability estimates of the scores on the test with items having the blank at the beginning do not differ significantly from those of the test with items having the blank at the end ($\alpha = .05$).

Elementary Level: Multiple-Choice Tests

The internal reliability estimates do not present a consistent pattern in which the scores on the four-option test are significantly more reliable than the three-option test ($\alpha = .05$).

Elementary Level: Completion Tests

The findings do not demonstrate any consistent effect of placement of the blank on the internal reliability of scores on the elementary level of completion.

Recommendations

Based on the present investigation, it appears that the results of the multiple-choice tests neither agree with Costin's (1970) findings which support the superiority of the three-option tests over the four-option tests in terms of their internal reliabilities, nor do they

support the statements of Thorndike and Hagen (1977) and Noll and Scannell (1972) that four-option tests are more reliable than the three-option tests. However, the discrepant results do not present conclusive evidence of effect of specific number of options (three or four) on the test reliability.

Also, the reliability of completion tests do not agree conclusively with the recommendations cited in the literature. However, there is no strong evidence of superiority of items having the blank at the end over the items having the blank at the beginning. It should be stressed that all of the items which were used in this study were not analyzed previously in terms of their difficulty and discrimination powers. Therefore, omission or revision of "bad" items could improve the reliability of scores on the tests. Also, the items were constructed to measure only knowledge level of achievement. Construction of items which measure other levels of the cognitive domain may yield different results. However, the tests were taken by subjects with specific characteristics. Thus, generalization of the results will only be limited to students similar to the subjects in this study.

It should be pointed out that this study investigated only two types of multiple-choice items; therefore, further investigation of other types of multiple-choice items, items with two or more options rather than three or four options, may provide different results than the present study. Furthermore, investigation of the two types of multiple-choice items did not consider the sequence of a fixed total number of alternatives in a test as it is suggested by Ebel. Thus, further investigation of multiple-choice items considering a fixed total

number of alternatives may furnish different findings than the present study.

SELECTED BIBLIOGRAPHY

Adkinswood, D. Test Construction. Columbus: Charles E. Merrill Books, Inc., 1960.

Aiken, L. R., Jr. "Another Look at Weighting Test Items." Journal of Educational Measurements, 1966, 3, 183-185.

Bartz, A. E. Basic Statistical Concepts in Education and Behavioral Science. Minneapolis: Burgess Publishing Company, 1976.

Beyar, I. I. & Weiss, D. J. "Comparison of Empirical Differential Option Weighting Scoring Procedures as a Function of Inter-Item Correlation." Educational and Psychological Measurement, 1977, 37, 335-340.

Brown, J. I., Nelson, M. J. & Denny, E. C. Examiner's Manual, the Nelson-Denny Reading Test. Boston: Houghton-Mifflin Company, 1973.

Bruning, J. L. & Knitz, B. L. Computational Handbook of Statistics. Glenview: Scott, Foresman and Company, 1977.

Costin, F. "The Optimal Number of Alternatives in Multiple-Choice Achievement Tests: Some Empirical Evidence for a Mathematical Proof." Educational and Psychological Measurement, 1970, 30, 353-358.

Cronback, L. J. "Coefficient Alpha and the Internal Structure of Tests." Psychometrica, 1951, 16, 297-334.

Davis, F. B. Item-Analysis Data. Cambridge: Graduate School of Education, Harvard University, 1946.

Drew, C. J. Introduction to Designing Research and Evaluation. St. Louis: The C. V. Mosby Company, 1976.

Ebel, R. L. Measuring Educational Achievement. Englewood Cliffs: Prentice-Hall, Inc., 1965.

Ebel, R. L. "Expected Reliability as a Function of Choices per Item." Educational and Psychological Measurement, 1969, 29, 565-570.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs: Prentice-Hall, Inc., 1972.

Engelhart, M. D. "A Comparison of Several Item Discrimination Indices." Journal of Educational Measurement, 1965, 2, 69-76.

Findley, W. G. "A Rationale for Evaluation of Item Discrimination Statistics." Educational and Psychological Measurement, 1956, 16, 175-180.

Flanagan, J. C. "General Considerations in Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient From Data at the Tails of Distribution." Journal of Educational Psychology, 1939, 30, 674-680.

Frisbie, D. A. "Multiple-Choice Versus True-False--A Comparison of Reliabilities and Concurrent Validities." Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D. A. "The Effect of Item Format on Reliability and Validity--A Study of Multiple-Choice and True-False Achievement Tests." Educational and Psychological Measurement, 1974, 34, 885-892.

Gilmore, H. L. Testing. New York: Wiley-Interscience, 1969.

Grier, J. B. "The Number of Alternatives for Optimum Test Reliability." Journal of Educational Measurement, 1975, 12, 109-113.

Grier, J. B. "The Optimal Number of Alternatives at a Choice Point With Travel Time Considered." Journal of Mathematical Psychology, 1976, 14, 91-97.

Hopkins, C. & Antes, R. Classroom Measurement and Evaluation. Itasca: F. E. Peacock Publishers, Inc., 1978.

Huck, S. W. & Bowers, N. D. "Item Difficulty Level and Sequence Effects in Multiple-Choice Achievement Tests." Journal of Educational Measurement, 1972, 9, 105-111.

Issac, S. & Michael, W. B. Handbook in Research and Evaluation. San Diego: Edits Publishers, 1976.

Kuder, G. F. & Richardson, M. W. "The Theory of the Estimation of Test Reliability." Psychometrica, 1937, 2, 151-160.

Lyman, M. B. Test Scores and What They Mean. Englewood Cliffs: Prentice-Hall, Inc., 1963.

Lindquist, E. F. Statistical Analysis in Educational Research. Boston: Houghton-Mifflin Company, 1940.

Lord, F. M. "Optimal Number of Choices per Item--A Comparison of Four Approaches." Journal of Educational Measurement, 1977, 14, 33-38.

Marshall, J. C. & Hales, L. W. Classroom Test Construction. New York: Addison-Wesley Publishing Company, 1971.

Morrison, D. G. Multivariate Statistical Methods. New York: McGraw-Hill Book Company, Inc., 1976.

Nie, N. H. & Hull, C. H. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, Inc., 1977.

Noll, V. H. Introduction to Educational Measurement. Boston: Houghton-Mifflin Company, 1957.

Noll, V. H. & Scannell, D. P. Introduction to Educational Measurement. Boston: Houghton-Mifflin Company, 1972.

Nunnally, J. C. Tests and Measurements: Assessment and Prediction. New York: McGraw-Hill Book Company, Inc., 1959.

Nunnally, J. C. Education Measurement and Evaluation. New York: McGraw-Hill Book Company, Inc., 1972.

Ramos, R. A. & Stern, J. "Item Behavior Associated With Changes in the Number of Alternatives in Multiple-Choice Tests." Journal of Educational Measurement, 1973, 10, 305-310.

Remmers, H. H. & Gage, N. L. Educational Measurement and Evaluation. New York: Harper and Brothers Publishers, 1955.

Remmers, H. H., Gage, N. L. & Rummuel, J. F. A Practical Introduction to Measurement and Evaluation. New York: Harper and Row Publishers, 1965.

Ruch, G. M. The Improvement of the Written Examination. Glenview: Scott, Foresman and Company, 1924.

Ruch, G. M. The Objective or New-Type Examination. Glenview: Scott, Foresman and Company, 1929.

Ruch, G. M. & Charles, J. W. "A Comparison of Five Types of Objective Tests in Elementary Psychology." Journal of Applied Psychology, 1928, 12, 398-403.

Ruch, G. M., Degraff, M. H., Gordon, W. E., McGregor, J. B., Maupin, N. & Murdock, J. R. Objective Examination Methods in the Social Studies. Glenview: Scott, Foresman and Company, 1926.

Ruch, G. M. & Stoddard, G. D. "Comparative Reliabilities of Five Types of Objective Examinations." Journal of Educational Psychology, 1925, 16, 89-103.

Ruch, G. M. & Stoddard, G. D. Tests and Measurements in High School Instruction. Glenview: World Book Company, 1927.

Sax, G. Principles of Educational Measurement and Evaluation. Belmont: Wadsworth Publishing Company, Inc., 1974.

Solomon, H. Item Analysis and Prediction. Stanford: Stanford University Press, 1961.

Thorndike, R. L. "Reliability." In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951, pp. 560-620.

Thorndike, R. L. & Hagen, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley and Sons, Inc., 1977.

Toops, H. A. "Trade Tests in Education." Teacher College Contributions to Education. New York: Columbia University, 1921.

Truab, R. E. & Hambleton, R. K. "Note of Correction on the Article Entitled: The Effect of Scoring Instructions and Degree of Speediness on the Validity and Reliability of Multiple-Choice Tests." Educational and Psychological Measurement, 1973, 33, 877-878.

Tversky, A. "On the Optimal Number of Alternatives of a Choice Point." Journal of Mathematical Psychology, 1964, 1, 386-391.

Wick, J. W. Educational Measurement: Where Are We Going and How Will We Know When We Get There? Columbus: Charles E. Merrill Publishing Company, 1973.

Wiley, D. E. & Glass, G. V. "Formula Scoring and Test Reliability." Journal of Educational Measurement, 1966, 1, 43-49.

APPENDIXES

APPENDIX A

COLLEGE AND ELEMENTARY MULTIPLE-CHOICE TESTS AND COMPLETION TESTS

COLLEGE LEVEL

Form A

This is not a test of your knowledge. It is a research instrument designed to give us more information about types of test items.

Please try to answer every item. Do not mark more than one answer per item. Thank you very much for completing this test.

Student's Name_____

Please write your answers on the test booklet. If you do not know the answer to an item, please make the best guess you can as to the correct answer. Please don't skip any questions. Mark only one answer for each question. Circle the correct answer for each item. Thank you.

1. Which one of the following is not spelled correctly?
   a. tiresome
   b. messanger
   c. plague

2. A centimeter equals
   a. 1/100 of a meter
   b. 100 meters
   c. 1/1000 of a meter
   d. 1/10 of a meter

3. Which of the following names does not belong with the other two?
   a. Benjamin Franklin
   b. Woodrow Wilson
   c. Lyndon Johnson

4. A scavenger consumes mainly
   a. ants
   b. dead organisms
   c. green plants
   d. grains

5. An animal which is a carnivore eats mainly
   a. grass
   b. animals
   c. shrubs
   d. corn.

6. Disclose means the same, or almost the same, as
   a. propose
   b. reveal
   c. dismiss
   d. undress

7. A drought occurs when there is a lack of
   a. sunshine
   b. vegetation
   c. rainfall
   d. animals

8. Which of the following books would you use to find a map of New Zealand?
   a. dictionary
   b. almanac
   c. atlas
   d. thesaurus

PLEASE GO TO THE NEXT PAGE

9. Which one of the following is spelled correctly?
   a. chemistry
   b. equiptment
   c. diarey
   d. acrilic

10. Excluded means the same, or about the same, as
   a. protested
   b. barred
   c. resisted

11. For a fire to burn, the three needed components are fuel, a high temperature, and
   a. oxygen
   b. matches
   c. wood

12. Which of the following animals is a vertebrate?
   a. starfish
   b. goldfish
   c. jellyfish
   d. earthworm

13. Table salt is an example of
   a. an element
   b. a chemical mixture
   c. a chemical compound

14. Menacing means the same, or about the same, as
   a. changing
   b. threatening
   c. unvarying

15. Water is formed when oxygen combines with
   a. air
   b. carbon
   c. hydrogen

16. Laborious means the same, or about the same, as
   a. difficult
   b. amortized
   c. precise

17. If R < S and S < T, then
   a. R = T
   b. R > T
   c. R < T
   d. S < R

18. Solitude means the same, or about the same, as
   a. despondent
   b. seclusion
   c. confusion

PLEASE GO TO THE NEXT PAGE

19. Botany is a science dealing with
    a. plants
    b. animals
    c. insects

20. The perimeter formula for a rectangle is
    a. $P = 2\ell + 2\omega$
    b. $P = \ell + \omega$
    c. $P = \ell\omega$
    d. $P = 4s$

Complete each of the following items by filling in the blank with the correct answer.

1. Rust forms on some metal objects when oxygen combines with_____.
2. The vice president of the United States is _____.
3. _____ is the capital of the state of Oklahoma.
4. 1 kilometer = _____ meter(s).
5. The freezing point of water at sea level is _____°C.
6. The southern boundary of Oklahoma is formed partially by the _____ River.
7. $\sqrt{169}$ = _____.
8. _____% = 2/3.
9. _____ wrote the Gettysburg Address.
10. _____ is the down payment required if the down payment on a car at $365 is set at 20%.
11. A chemical often added to table salt to help prevent the formation of goiter is _____.
12. The common denominator for the fractions 2/3, 3/4, and 5/6 is _____.
13. _____ is the list price of an item costing $5.60 that is advertised at 30% off the list price.
14. The _____ is the plant part through which plants take up water.
15. _____ are the major topics of study in zoology.

PLEASE GO TO THE NEXT PAGE

16. The _____ is the plant part in which plants make most of their food.

17. Oklahoma share the major part of its northern boundary with the state of _____.

18. _____ is used as the punctuation mark at the end of interrogative sentences.

19. x = _____ if 17x + 21 = 72.

20. 363 ÷ 3.3 = _____.

PLEASE RECHECK YOUR ANSWERS

THANK YOU FOR ANSWERING THESE QUESTIONS

COLLEGE LEVEL

Form B

This is not a test of your knowledge. It is a research instrument designed to give us more information about types of test items.

Please try to answer every item. Do not mark more than one answer per item. Thank you very much for completing this test.

Student's Name_____

Please write your answers on the test booklet. If you do not know the correct answer to an item, please make the best guess you can as to the correct answer. Please don't skip any questions. Mark only one answer for each question. Thank you.

1. Which one of the following is not spelled correctly?
   a. tiresome
   b. messanger
   c. ratify
   d. plague

2. A centimeter equals
   a. 1/100 of a meter
   b. 100 meters
   c. 1/1000 of a meter

3. Which of the following names does not belong with the other three?
   a. Jimmy Carter
   b. Benjamin Franklin
   c. Woodrow Wilson
   d. Lyndon Johnson

4. A scavenger consumes mainly
   a. ants
   b. dead organisms
   c. green plants

5. An animal which is a carnivore eats mainly
   a. grass
   b. animals
   c. shrubs

6. Disclose means the same, or about the same, as
   a. propose
   b. reveal
   c. dismiss

7. A drought occurs when there is a lack of
   a. sunshine
   b. vegetation
   c. rainfall

8. Which of the following books would you use to find a map of New Zealand?
   a. dictionary
   b. almanac
   c. atlas

9. Which one of the following is spelled correctly?
   a. chemistry
   b. equiptment
   c. diarey

PLASE GO TO THE NEXT PAGE

10. Excluded means the same, or about the same, as
    a. hurried
    b. protested
    c. barred
    d. resisted

11. For a fire to burn, the three needed components are fuel, a high temperature and
    a. carbon
    b. oxygen
    c. matches
    d. wood

12. Which of the following animals is a vertebrate?
    a. starfish
    b. goldfish
    c. jellyfish

13. Table salt is an example of
    a. an element
    b. a chemical mixture
    c. a chemical compound
    d. an atom

14. Menacing means the same, or about the same, as
    a. changing
    b. threatening
    c. unvarying
    d. increasing

15. Water is formed when oxygen combines with
    a. air
    b. carbon
    c. fire
    d. hydrogen

16. Laborious means the same, or about the same, as
    a. difficult
    b. amortized
    c. miserly
    d. precise

17. If $R < S$ and $S < T$, then
    a. $R = T$
    b. $R > T$
    c. $R < T$

18. Solitude means the same, or about the same, as
    a. despondent
    b. curious
    c. seclusion
    d. confusion

PLEASE GO TO THE NEXT PAGE

19. Botany is a science dealing with
    a. plants
    b. animals
    c. rocks
    d. insects

20. The perimeter formula for a rectangle is
    a. $P = 2\ell + 2\omega$
    b. $P = \ell + \omega$
    c. $P = \ell\omega$

Complete each of the following items by filling in the blank with the correct answer.

1. _____ combines with oxygen to form rust on some metal objects.

2. _____ is the vice president of the United States.

3. The capital of the state of Oklahoma is _____.

4. _____ meter(s) = 1 kilometer.

5. _____ °C is the freezing point of water at sea level.

6. The _____ River forms part of the southern boundary of Oklahoma.

7. _____ = $\sqrt{169}$

8. 2/3 = _____%.

9. The name of the author of the Gettysburg Address is _____.

10. If the down payment on a car priced at $3675 is set at 20%, the amount of down payment required is _____.

11. _____ is the chemical often added to table salt to help prevent the formation of goiters.

12. _____ is the common denominator for the fractions 2/3, 3/4, and 5/6.

13. If an item that is advertised at 30% off the list price costs $5.60, the list price is _____.

14. Plants take up water through the plant part known as the _____.

15. Zoology is a branch of science dealing with the study of _____.

PLEASE GO TO THE NEXT PAGE

16. Plants make most of their food in the plant part known as the _____.

17. _____ is the state Oklahoma shares the major part of its northern boundary with.

18. Interrogative sentences end with a puncutation mark called a(n) _____.

19. If $17x + 21 = 72$ then $x =$ _____.

20. _____ = $363 \div 3.3$.

PLEASE RECHECK YOUR ANSWERS

THANK YOU FOR ANSWERING THESE QUESTIONS

ELEMENTARY LEVEL

Form A

This test will not be used to give you a grade. The teacher will tear off your name before it is turned in to me. Thank you very much for taking this test for me.

My name is_____

Please write your answers on the test booklet. If you do not know the answer to an item, please make the best guess you can as to the correct answer. Please don't skip any questions. Mark only one answer for each question. Circle the letter of the correct answer to each question. Thank you.

1. Select the word that means the same or about the same as logical.
   a. typical
   b. reasonable
   c. unexpected

2. If you wanted to find out the meaning of a word in your social studies book, you should look in the
   a. table of contents
   b. glossery
   c. index
   d. summaries

3. What time will it be 8 hours after 4:45 a.m.?
   a. 1:45 a.m.
   b. 1:45 p.m.
   c. 8:45 p.m.
   d. 1:15 p.m.

4. What is the name of the county in which you attend school?
   a. Stillwater
   b. Payne
   c. Oklahoma
   d. Logan

5. Square inches are used to measure
   a. length
   b. volume
   c. area
   d. width

6. Tom got 22 questions correct on the test. His score was 50%. How many questions were there in all?
   a. 11
   b. 50
   c. 44
   d. 22

7. What product of great value to the United States do the Arab countries produce?
   a. steel
   b. gold
   c. oil
   d. camels

GO TO THE NEXT PAGE

8. Which of the following words would be nearest to the middle of a dictionary?
   a. able
   b. won
   c. merry
   d. value

9. What is often the cause of erosion?
   a. forest fires
   b. new seedlings
   c. too many trees

10. Select the word that means the same or about the same as deserve.
    a. treat
    b. earn
    c. expect

11. Select the correct answer to the following subtraction problem:

    5,681
    - 796

    a. 4,895
    b. 3,885
    c. 4,885

12. Which of the following makes this number sentence true?
    8 + 4 = 19 - ☐
    a. 7
    b. 0
    c. 1
    d. -7

13. Select the correct answer to the following problem:
    5/552
    a. 100 R2
    b. 110 R2
    c. 112

14. Most factories are found in
    a. cities
    b. small towns
    c. mountain areas

15. Which one of the following words is misspelled?
    a. dollars
    b. recieved
    c. candle

16. Which name does not belong in the list below?
    a. Abraham Lincoln
    b. George Washington
    c. Benjamin Franklin

GO TO THE NEXT PAGE

17. What type of book would you use to find a map of France?
    a. dictionary
    b. atlas
    c. bibliography
    d. almanac

18. Which of the following is a large city?
    a. Arizona
    b. New Mexico
    c. Chicago

19. What is the missing number?
    1, 3, 5, 7, __, 11
    a. 8
    b. 9
    c. 10

20. Which of the following words would be nearest to the end of a dictionary?
    a. island
    b. youth
    c. under
    d. olive

Complete each of the following sentences by filling in the blank with the correct word or number.

1. _____, 2, 4, 6, 8.

2. The president of the United States is _____.

3. The largest state in the United States is _____.

4. 3 pints of liquid equal _____ cups.

5. The color purple can be made by combining blue and _____.

6. The capital of the United States is _____.

7. 72 inches = _____ yards.

8. 2 feet, 3 inches = _____ inches.

9. _____ X 8 = 10 X 4.

10. _____ = 36.2 - 3.6.

11. The moon has no _____ and so a candle will not burn on the moon.

GO TO THE NEXT PAGE

12. The temperature at which water freezes is _____ °F.

13. _____ is the month in which Labor Day occurs.

14. _____ is the capital city of Oklahoma.

15. _____, soil, air, and water are needed by green plants in order to live.

16. _____ is the shortest month of the year.

17. One year equals _____ months.

18. _____ = $25.00 - $1.75.

19. _____ is written as XVIII in the Roman numeral system.

20. 5 X (4 + 3) = _____ .

PLEASE RECHECK YOUR ANSWERS

THANK YOU FOR TAKING THIS TEST

How old are you? _____

What grade are you in school? _____

Are you a boy or a girl? _____

ELEMENTARY LEVEL

Form B

This test will not be used to give you a grade. The teacher will tear off your name before it is turned in to me. Thank you very much for taking this test for me.

My name is _____

Please write your answers on the test booklet. If you do not know the answer to an item, please make the best guess you can as to the correct answer. Please don't skip any questions. Mark only one answer for each question. Circle the letter of the correct answer to each question. Thank you.

1. Select the word that means the same or about the same as logical.
   a. typical
   b. skill
   c. reasonable
   d. unexpected

2. If you wanted to find out the meaning of a word in your social studies book, you should look in the
   a. table of contents
   b. glossery
   c. index

3. What time will it be 8 hours after 5:45 a.m.?
   a. 1:45 a.m.
   b. 1;45 p.m.
   c. 8:45 p.m.

4. What is the name of the county in which you attend school?
   a. Stillwater
   b. Payne
   c. Oklahoma

5. Square inches are used to measure
   a. length
   b. volume
   c. area

6. Tom got 22 questions correct on the test. His score was 50%. How many questions were there in all?
   a. 11
   b. 50
   c. 44

7. What product of great value to the United States do the Arab countries produce?
   a. steel
   b. gold
   c. oil

8. Which of the following words would be nearest to the middle of a dictionary?
   a. able
   b. won
   c. merry

GO TO THE NEXT PAGE

9. What is often the cause of erosion?
   a. increased fishing
   b. forest fires
   c. new seedlings
   d. too many trees

10. Select the word that means the same or about the same as deserve.
   a. treat
   b. earn
   c. begin
   d. expect

11. Select the correct answer to the following subtraction problem:

5,681	a. 4,895
- 796	b. 5,115
	c. 3,885
	d. 4,885

12. Which of the following makes this number sentence true?
   8 + 4 = 19 - ☐
   a. 7
   b. 0
   c. 1

13. Select the correct answer to the following problem:
   5/552
   a. 100 R2
   b. 110 R2
   c. 112
   d. 140

14. Most factories are found in
   a. cities
   b. small towns
   c. farming areas
   d. mountain areas

15. Which one of the following words is misspelled?
   a. dollars
   b. daily
   c. recieved
   d. candle

16. Which name does not belong in the list below?
   a. Abraham Lincoln
   b. George Washington
   c. Benjamin Franklin
   d. Jimmy Carter

17. What type of book would you use to find a map of France?
   a. dictionary
   b. atlas
   c. bibliography

GO TO THE NEXT PAGE

18. Which of the following is a large city?
    a. Arizona
    b. New Mexico
    c. Japan
    d. Chicago

19. What is the missing number?
    1, 3, 5, 7, __, 11
    a. 8
    b. 9
    c. 10
    d. 11

20. Which of the following words would be nearest to the end of a dictionary?
    a. island
    b. youth
    c. under

Complete each of the following sentences by filling in the blank with the correct word or number.

1. 2, 4, 6, 8, _____.
2. _____ is the president of the United States.
3. _____ is the largest state in the United States.
4. _____ cups equal 3 pints of liquid.
5. _____ and blue combine to give the color purple.
6. _____ is the capital of the United States.
7. _____ yards = 72 inches.
8. _____ in. = 2 feet, 3 inches.
9. 10 X 4 = 8 X _____.
10. 36.2 - 3.6 = _____.
11. A candle will not burn on the moon because the moon has no _____.
12. _____ °F is the temperature at which water freezes.
13. Labor Day is a holiday in the month of _____.
14. The capital city of Oklahoma is _____.

GO TO THE NEXT PAGE

15. In order to live, green plants need soil, water, air, and _____.

16. The shortest month of the year is _____.

17. _____ months equal 1 year.

18. $25.00 - $1.75 = _____.

19. XVIII is the Roman numeral for _____.

20. _____ = 5 X (4 + 3)

PLEASE RECHECK YOUR ANSWERS

THANK YOU FOR TAKING THIS TEST

How old are you? _____

What grade are you in school? _____

Are you a boy or a girl? _____

APPENDIX B

INSTRUCTIONS FOR ADMINISTERING MULTIPLE-CHOICE TESTS AND COMPLETION TESTS

Procedures

1. Write each student's name on both copies of each pair of tests. Separate each pair of tests.
2. Administer one test to each individual allowing ample time for the student to complete the test. Please do not give any help to the students or discuss the test with the students.
3. During a class period from one to two weeks after the first testing period, administer the second test to each student. Please do wait at least one week but not more than two weeks to administer the second test.
4. Record the student's scores on the Nelson-Denny Reading Test on the last page of the second test. Please record them in the following order: Vocabulary, Comprehension, Rate of Reading.
5. Clip together each pair of completed tests for each student.
6. Tear off the student's name on each test.

Thank you very much for helping me with this research project.

N. Jo Campbell
ABSED

APPENDIX C

KEY RESPONSES OF COLLEGE AND ELEMENTARY MULTIPLE-CHOICE TESTS AND COMPLETION TESTS

College Level Key

Answers in parentheses are acceptable.

Form A	Form B
1. b	1. b
2. a	2. a
3. a	3. b
4. b	4. b
5. b	5. b
6. b	6. b
7. c	7. c
8. c	8. c
9. a	9. a
10. b	10. c
11. a	11. b
12. b	12. b
13. c	13. c
14. b	14. b
15. c	15. d
16. a	16. a
17. c	17. c
18. b	18. c
19. a	19. a
20. a	20. a

---

1. Iron	1. Iron
2. Walter Mondale (Mondale)	2. Walter Mondale (Mondale)
3. Oklahoma City	3. Oklahoma City

4. 1000	4. 1000
5. 0 (zero)	5. 0 (zero)
6. Red	6. Red
7. 13	7. 13
8. 66 2/3 (67, 66.7, 66.67, 66.667, . . .)	8. 66 2/3 (67, 66.7, 66.67, 66.667, . . .)
9. Lincoln (Abraham Lincoln)	9. Lincoln (Abraham Lincoln)
10. $73	10. $735
11. Iodine (Iodide)	11. Iodine (Iodide)
12. 12	12. 12
13. $8.00	13. $8.00
14. Root (roots, root hair)	14. Root (roots, root hair)
15. Animals	15. Animals
16. Leaf (leaves)	16. Leaf (leaves)
17. Kansas	17. Kansas
18. ? (question mark)	18. ? (question mark)
19. 3 (51/17)	19. 3 (51/17)
20. 110	20. 110

Elementary Level Key

Answers in parentheses are acceptable.

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

---

1. 0	1. 10
2. Jimmy Carter (Carter)	2. Jimmy Carter (Carter)
3. Texas	3. Texas
4. 6	4. 6
5. red (pink)	5. red (pink)

6. Washington, D. C.	6. Washington, D. C.
7. 2	7. 2
8. 27	8. 27
9. 5	9. 5
10. 32.6	10. 32.6
11. Oxygen	11. Oxygen
12. 32	12. 32
13. September	13. September
14. Oklahoma City	14. Oklahoma City
15. Sunlight (sun, light, sunshine)	15. Sunlight (sun, light, sunshine)
16. February	16. February
17. 12 (Twelve)	17. 12 (Twelve)
18. $23.25	18. $23.25
19. 18 (eighteen)	19. 18 (eighteen)
20. 35 (7 x 5)	20. 35 (7 x 5)

APPENDIX D

ITEM ANALYSIS OF COLLEGE LEVEL DATA

TABLE XIII

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF MULTIPLE-CHOICE TESTS--FORM A OF COLLEGE LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Three-Option	1	.826	.20925
	3	.826	.42801
	10	.870	.42284
	11	.891	.40346
	13	.630	.64980
	14	.804	.48620
	15	.935	.20442
	16	.644	.60013
	18	.935	.20442
	19	.848	.48177
Four-Option	2	.674	.41034
	4	.957	.37117
	5	.804	.39562
	6	.696	.24902
	7	1.000	*
	8	.978	.09524
	9	.913	.31053
	12	.522	.54773
	17	.913	.48058
	20	.435	.59956

n = 46.

*Coefficient cannot be computed.

TABLE XIV

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF MULTIPLE-CHOICE TESTS--FORM B OF COLLEGE LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Three-Option	2	.720	.49611
	4	.920	.45855
	5	.800	.67783
	6	.740	.55814
	7	.980	.03228
	8	1.000	*
	9	.960	.24959
	12	.660	.44274
	17	.940	.16901
	20	.600	.64569
Four-Option	1	.840	.35762
	3	.840	.39228
	10	.900	.43194
	11	.920	.19480
	13	.510	.38455
	14	.837	.53215
	15	.900	.47429
	16	.600	.48752
	18	.860	.52283
	19	.878	.49756

n = 50.

*Coefficient cannot be computed.

TABLE XV

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF COMPLETION TESTS--FORM A OF COLLEGE LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Blank at the Beginning	3	1.000	*
	8	.442	.62244
	9	.844	.44485
	10	.658	.69200
	13	.128	.50059
	14	.791	.31313
	15	.822	.39661
	16	.683	.39274
	18	.477	.53225
	19	.860	.12170
Blank at the End	1	.067	.40632
	2	.744	.39206
	4	.477	.60906
	5	.500	.47179
	6	.778	.25830
	7	.737	.57409
	11	.526	.38528
	12	.822	.47016
	17	.867	.34975
	20	.610	.51265

n = 46.

*Coefficient cannot be computed.

TABLE XVI

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF COMPLETION TESTS--FORM B OF COLLEGE LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Blank at the Beginning	1	.085	.39289
	2	.889	.50937
	4	.532	.59226
	5	.667	.58333
	6	.667	.43214
	7	.780	.66504
	11	.500	.37046
	12	.816	.32100
	17	.840	.33395
	20	.636	.57993
Blank at the End	3	1.000	*
	8	.638	.48985
	9	.830	.41778
	10	.500	.57471
	13	.154	.32955
	14	.800	.14618
	15	.915	.25895
	16	.317	.38566
	18	.636	.41228
	19	.830	.50491

n = 50.

*Coefficient cannot be computed.

APPENDIX E

ITEM ANALYSIS OF ELEMENTARY LEVEL DATA

TABLE XVII

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF MULTIPLE-CHOICE TESTS--FORM A OF ELEMENTARY LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Three-Option	1	.857	.55552
	8	.769	.56344
	9	.692	.49553
	11	.923	-.03259
	12	.923	.95607
	13	.917	.16116
	14	.923	.95607
	16	.846	.67767
	17	1.000	*
Four-Option	2	.714	.33472
	3	.714	.83173
	4	.429	.61111
	5	.714	.61872
	6	.786	-.13401
	7	.923	.60486
	10	.769	.73321
	15	.846	.79665
	18	1.000	*

n = 14.

*Coefficient cannot be computed.

TABLE XVIII

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF MULTIPLE-CHOICE TESTS--FORM B OF ELEMENTARY LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Three-Option	2	.857	.19702
	3	.810	.57908
	4	.619	.57037
	5	.667	.68250
	6	.714	.75503
	7	.952	.38621
	10	.900	.29159
	15	.900	.59324
	18	.905	.47385
Four-Option	1	.667	.35049
	8	.800	.78588
	9	.789	.77911
	11	.850	.44728
	12	.900	.77743
	13	.737	.68086
	14	.600	.50713
	16	.905	.06213
	17	.952	.51891

n = 21.

TABLE XIX

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF COMPLETION TESTS--FORM A OF ELEMENTARY LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Blank at the Beginning	1	.923	-.01145
	9	.929	.28738
	10	.692	.23131
	11	.538	.59962
	13	.385	.52038
	14	.923	.43498
	15	.615	.78370
	16	.769	.35474
	18	.769	.35474
	19	.385	.76489
Blank at the End	2	.929	-.38243
	3	.308	.04042
	4	.462	.42102
	5	.538	.43038
	6	1.000	*
	7	.769	.68635
	8	.769	.54244
	12	.538	.79526
	17	1.000	*
	20	.692	.22232

n = 14.

*Coefficient cannot be computed.

TABLE XX

ITEM DIFFICULTY AND DISCRIMINATION INDICES OF COMPLETION TESTS--FORM B OF ELEMENTARY LEVEL

Type of Test	Item	Difficulty Index	Discrimination Index
Blank at the Beginning	2	.952	.45394
	3	.190	.04924
	4	.526	.82831
	5	.850	.56467
	6	.950	.53560
	7	.526	.51582
	8	.737	.86284
	12	.450	.66125
	17	.947	.62238
	20	.556	.64027
Blank at the End	1	1.000	*
	9	.632	.76471
	10	.632	.37596
	11	.421	.18225
	13	.300	.39533
	14	.895	.60306
	15	.632	.37596
	16	.850	.60211
	18	.684	.68741
	19	.316	.57329

n = 21.

*Coefficient cannot be computed.

VITA

Rosa Torabi-Parizi

Candidate for the Degree of

Master of Science

Thesis: THE EFFECTS OF MULTIPLE-CHOICE FORMATS AND COMPLETION FORMATS ON TEST RELIABILITY

Major Field: Educational Psychology

Biographical:

Personal Data: Born in Zarand, Iran, December, 1952, the daughter of Mr. and Mrs. Ahmad Torabi-Parizi.

Education: Graduated from Kharazmi High School, Tehran, Iran, in 1968; received Bachelor in Guidance and Counseling degree from University of Teacher Education in 1973; completed requirements for the Master of Science degree at Oklahoma State University in May, 1980.

Professional Experience: School Counselor, Kerman, Iran, 1973-1974; School Counselor, Kashan, Iran, 1974-1976; Research Assistant, Institute for Educational Research, Tehran, Iran, 1976-1977.