COMPARATIVE RELIABILITIES AND VALIDITIES OF TRUE-FALSE AND MULTIPLE CHOICE TESTS

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
DAVID A. FRISBIE
1971

This is to certify that the

thesis entitled

COMPARATIVE RELIABILITIES AND
VALIDITIES OF TRUE-FALSE AND MULTIPLE CHOICE TESTS

presented by

David A. Frisbie

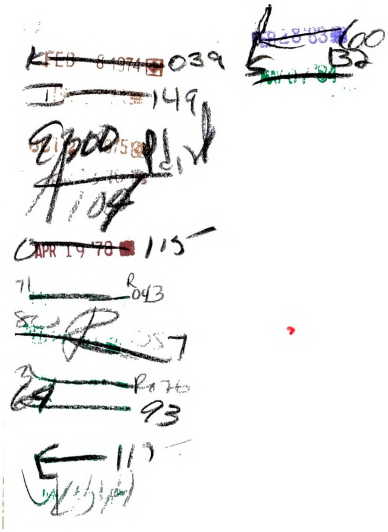
has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Robert L. Ebel
Major professor

Date July 20, 1971

O-7639



ABSTRACT

COMPARATIVE RELIABILITIES AND VALIDITIES OF TRUE-FALSE AND MULTIPLE CHOICE TESTS

By

David A. Frisbie

This study was designed to compare the reliabilities of true-false and multiple choice tests and to determine the concurrent validities of true-false tests. Four major questions were formulated as research hypotheses:

1. Are true-false and multiple choice achievement tests that were designed to measure the same objectives equally reliable?
2. Are true-false tests constructed by the judgmental method as reliable as those developed by the discrimination method?
3. What is the ratio of the number of true-false items attempted to the number of multiple choice items attempted by a group of examinees in a given period of testing?
4. Is there a perfect correlation (+1.00) between true-false and multiple choice test scores when the correlation coefficients are corrected for attenuation?

Two methods were devised for systematically changing multiple choice items to true-false form. The judgmental method involved the use of teachers to choose the multiple choice distractor that would result in the most plausible false statement. The discrimination method relied on item analysis data from a multiple choice testing to identify the distractor that best discriminated between high and low scorers on the test. The first of three phases of testing was needed to gather the item analysis data.

The true-false items generated by the two conversion methods were tried out in the second phase of testing. The revised true-false items were incorporated in the eight test forms used in phase three, the final testing. Each of the 70-item final forms consisted of 35 multiple choice and 35 true-false items.

A sample of 1018 non-urban high school students each responded to one of eight test forms. The three factors that differentiated the forms were:

1. Subject matter (natural science or social studies)
2. Method of conversion (judgmental or discrimination)
3. Subtest order (true-false first or last)

Kuder-Richardson Formula 20 reliability coefficients were calculated for the eight multiple choice and eight true-false subtests. The ratio of the number of true-false to multiple choice items that subjects attempted

in the first eight minutes of testing was also computed. The correlation between multiple choice and true-false subtest scores was calculated and corrected for attenuation for each final test form. Statistical tests were performed to determine if the 16 reliability coefficients were homogeneous and to ascertain if the corrected correlation coefficients significantly departed from unity.

The results associated with each of the major questions of interest were:

1. The reliabilities of the multiple choice tests were significantly greater than the reliabilities of the true-false tests.
2. There was no significant difference between the reliabilities of the true-false tests constructed by the judgmental or discrimination methods.
3. Examinees responded to three true-false items for every pair of multiple choice items attempted.
4. The corrected correlation coefficients for six of the eight final forms were significantly less than unity.

COMPARATIVE RELIABILITIES AND VALIDITIES OF TRUE-FALSE AND MULTIPLE CHOICE TESTS

By

David A. Frisbie

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel
Services, and Educational Psychology

1971

ACKNOWLEDGMENTS

Sincere thanks are extended to Dr. Robert L. Ebel, chairman of my Guidance Committee, for his advice, counsel, and friendship throughout my doctoral program. The contributions of each of my committee members--Dr. Robert C. Craig, Dr. Laurine E. Fitzgerald, Dr. William A. Mehrens, and Dr. Willard G. Warrington--are gratefully acknowledged.

Several individuals made worthy contributions to this research effort. I wish to thank

--the Office of Evaluation Services staff for providing prompt scoring and item analysis services.

--the high school principals and teachers who permitted me to use their classes and who gave me such excellent cooperation in the data collection phase.

--my friends and colleagues who gave their insightful suggestions for improving the study in its early stages.

There is no way I can adequately express my gratitude to my wife, Janet, and my children, Marcy and Scott, for the support they have given me during my graduate studies. Their efforts represent the most important contribution to my successful completion of a doctoral program.

The financial support of the U. S. Office of Education through a Research Director Training Program fellowship enabled the author to complete his doctoral studies at Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

CHAPTER I

THE PROBLEM

Alternative Forms of Test Items

Test construction specialists and teachers who develop instruments to measure educational achievement have a wide variety of item forms at their disposal for accomplishing their specific objectives. Essay, multiple choice, true-false, matching, completion, simple recall or short answer, and novel combinations of these forms are usually mentioned in measurement textbooks as appropriate item forms for teachers to use.

Presently the multiple choice item is "the most highly regarded and widely used form of objective test item" (Ebel, 1965, p. 149). Multiple choice items have been recommended for achievement testing because of their apparent versatility and adaptability to the measurement of outcomes requiring mental processes beyond mere recall (Durost and Prescott, 1962; Brown, 1970; Ahmann and Glock, 1967). Thorndike and Hagen (1969, p. 102) wrote:

> [The multiple choice item] can be used to appraise the achievement of any of the educational objectives that can be measured by a paper-and-pencil test except those relating to skill in written expression and originality.

Advantages and Limitations of True-False Items

Many authors who discuss the use of various item forms suggest that the advantages of true-false items are outweighed by their limitations (Ahmann and Glock, 1967; Gronlund, 1965; Brown, 1970). One of the major shortcomings attributed to true-false items is the difficulty encountered in preparing good statements that measure more than simple factual information. Wesman (1971) noted that many of the important objectives of instruction require generalizations, explanations, evaluations, and inferences. He concluded that since these outcomes often cannot be expressed in statements that are precisely or universally true, they cannot be tested with true-false items.

There is substantial agreement among the authors cited above that true-false items are appropriate only for measuring knowledge of unambiguous factual material, and that multiple choice items can be used to measure a variety of outcomes. However, empirical evidence has not been presented to support this viewpoint. A search of the literature failed to turn up any data that would support or challenge the statements referred to above.

Though true-false items do have some limitations, they do have unique advantages in measuring educational achievement.

True-false items are more efficient than multiple choice items of comparable quality. Examinees can respond to more true-false items than multiple choice items in a given time period. The greater efficiency may lead to a more reliable test.

A true-false test probably can provide a broader sampling of the examinee's knowledge than can a multiple choice test intended to measure the same subject matter. Since more true-false items can be used in a given time period, these items should more thoroughly sample the universe of content than a multiple choice test.

Need for This Study

The arguments for and against either multiple choice or true-false items for measuring achievement can be found in many textbooks and journal articles, but little empirical data is available to substantiate the viewpoints expressed. The most equitable means of comparing the two item forms is through empirical study.

Purpose of This Study

The purpose of this study was to compare the reliabilities and validities of true-false and multiple choice tests that were written to measure understanding of concepts and relationships in the same content areas.

Two systematic procedures for changing multiple choice items to true-false form were used so that corresponding items of the two types would be as equivalent as possible in content. These procedures also made the conversion processes more objective and reproducable. The two conversion methods, judgmental and discrimination, were compared to determine if one yielded a more reliable or valid test than the other. (These conversion methods are defined later in this chapter and are described in detail in Chapter III.)

Data was collected to determine the number of items of each type that examinees attempted in a fixed period of time. This information was used to determine the theoretical length of the true-false tests. The reliabilities of the lengthened tests were compared to the reliabilities of the multiple choice tests with testing time constant. It was necessary to collect these data because the data in the literature were not recent or were not empirically based.

Finally this study was devised to compare items that were written to measure achievement in natural science and in social studies. Two content areas were used so that effects due to particular subject matter could be examined and so that the findings could be generalized somewhat.

Hypotheses

The major hypotheses in this study were:

1. When multiple choice items are converted to true-false form, the reliabilities of the two test forms are not different.
2. When multiple choice items are converted to true-false form using the judgmental and discrimination methods, the reliabilities of the true-false tests are not different.
3. More true-false than multiple choice items are attempted by the same group of examinees in a fixed time period.
4. The simple correlation between individuals' true-false and multiple choice test scores is 1.00 when corrected for attenuation.

Definition of Terms

Test reliability is defined as a measure of internal consistency estimated by the Kuder-Richardson Formula 20.

The two methods of converting multiple choice items to true-false format are defined as judgmental and discrimination. The judgmental method employs subject matter experts to choose the multiple choice incorrect response that would result in the most plausible false statement when used with the stem. The discrimination

method relies on item analysis data from a multiple choice testing to identify the multiple choice incorrect response that best discriminates between low and high scorers on the test.

Simple correlation is defined by the Pearson product moment correlation formula found in most elementary statistics books (see Chapter III).

Correction for attenuation is defined as the procedure used to estimate the true correlation between two variables represented by unreliable measures. (The formula used for this procedure appears in Chapter III.)

Overview

In Chapter II the literature relevant to the general problem and each of the specific hypotheses is reviewed. The design of the study, the sample, the instrumentation, and the methods of analysis are discussed in Chapter III. Chapter IV, the results of the study, is followed by a final chapter that contains: a summary of the study, a discussion of the findings, the limitations of the study, and suggestions for future research.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

An abundance of research that is concerned with comparing various item forms has been reported in the literature. Many studies were conducted during the late 1920's when objective test items made their initial surge in popularity. The first section of this chapter includes investigations that deal with the comparison of either multiple choice or true-false items and some other item type. Subsequent sections each contain reviews of studies that are related to each of the hypotheses enumerated in the previous chapter.

Section two provides an examination of the research that has focused on the comparative reliability and validity of both true-false and multiple choice items. The third section of this chapter deals with studies that have employed methods of converting items from one form to another. The final section contains a review of research reports that provide information concerning the difference in testing time required for tests composed of different item forms. Brief summaries have been provided at the

close of each section of this chapter in lieu of a general chapter summary.

General Studies Comparing Item Forms

Several studies were reported in which selected test characteristics were compared. Heim and Watts (1967) compared multiple choice and completion vocabulary items and found that the open-ended items were significantly more difficult. A similar study by Andrews and Bird (1938) in psychology yielded the same results. In addition they found higher odd-even reliabilities for completion items. Choppin and Purves (1969) concluded that multiple choice and open-ended literature items measured the same thing in their validity study.

Cronbach (1941) used multiple multiple choice items (more than one correct alternative per item) and multiple true-false items (each multiple choice alternative was marked true or false) based on introductory psychology content. No differences were found between test forms with regard to testing time, difficulty, reliability, and validity.

Validity and Reliability Studies

In what could be considered the pioneer study of this general problem, Toops (1921) compared the

reliabilities of 50-item general information tests, each cast into recall, multiple choice, and true-false form. Each subject took each test but the order of administration was varied so that six groups ranging in size from 39 to 10 were used. The corrected split-halves reliabilities reported were .556 and .507 for multiple choice and true-false tests, respectively. When testing time was held constant these reliabilities were estimated to be .607 and .664.

Completion, multiple choice, and true-false test reliabilities and validities were compared by Rutledge (1926) in his dissertation. Three forms of an elementary psychology midterm exam were constructed so that each form contained 40 test items of each of the three types. The examination formats were arranged so that the first 40 items of form A were true-false, the first 40 items of form B were multiple choice, and the first 40 items of form C were completion. The following items illustrate a typical test item in the three formats:

> A beginning was made in the study of scientific psychology in the _____ century.
>
> The study of scientific psychology began in the 18th century.
>
> A beginning was made in the study of scientific psychology in the (1) 17th, (2) 18th, (3) 19th, (4) 20th, century.

Corrected split-halves reliability coefficients averaged .70 and .89 for multiple choice and true-false items, respectively, across test forms. Nine minutes of testing time were allotted for each of the two forms; therefore, no adjustment of reliability coefficients was made. The average correlation between multiple choice and true-false subtest scores across test forms was .64, corrected for attenuation.

In a similar study by Charles (1926), the reliabilities of five-, three-, and two-response multiple choice tests and a true-false test were compared with a completion test reliability. Fifty factual information items from introductory psychology were administered to each subject in completion form followed by 50 items of one of the other item forms. The results of Charles' study are summarized in Table 2.1. No statistical tests were made but one could conclude that there is little practical difference between the reliabilities of true-false and five-response or three-response multiple choice tests. (Charles offered no good explanation for the unusual performance of the two-response multiple choice tests.)

A study carried out by Ruch and Stoddard (1925) employed a design identical to Charles' and items intended to measure history and social science general information. The reliabilities for the five-, three-, and two-response

multiple choice tests and the true-false test were .886, .748, .849, and .714, respectively, for 100-item tests. The reliabilities were recalculated to equate testing time and the new values were estimated to be .901, .806, .902, and .820.

Table 2.1. Summary of reliabilities and validities from Charles' study.

Item Form	r_{tt}	r_{nn}[a]	r_{cf}[b]
Completion	.603	.752	.603
5-R	.680	.809	.714
3-R	.624	.768	.703
2-R	.477	.646	.639
True-false	.602	.751	.680

[a]Corrected split-halves reliabilities.

[b]Average correlation between the completion test score and the score from each of the other item forms.

Reliability studies reported by Watson and Crawford (1930), Copeland and Gilliland (1943), and Eurich (1931) yielded conflicting results. Watson and Crawford estimated reliabilities favoring multiple choice items on high school physics unit tests. Copeland and Gilliland corrected their reliability coefficients to hold testing time constant and found a higher reliability for the 20-item

true-false child psychology test. Two experiments were reported by Eurich in which educational psychology test items were used. The multiple choice test reliability was higher in one trial and the same as the true-false test reliability in the other trial.

An item conversion method was employed by Burmeister and Olson (1966) to aid in determining whether college-level natural science true-false items could be written that had the same desirable characteristics as the multiple choice form. The authors concluded that true-false items could be constructed that discriminate "almost as well as" multiple choice items, and that true-false items were less difficult because of the guessing effect.

Ebel (1971) used a 90-item natural science multiple choice test as a basis for studying the validity and reliability of true-false items. Two forms, each containing 44 items of each item type were administered to groups of 53 and 50 students in an education course. The mean discrimination indices tended to be higher for the multiple choice tests. The Kuder-Richardson Formula 20 reliabilities for the multiple choice and true-false subtests were .81 and .84, respectively, for form one and .86 and .71 for form two. (The true-false reliabilities were estimated by the Spearman-Brown Formula for a double-length test under the assumption that two items can be attempted for every multiple choice item attempted.) The correlations

between multiple choice and true-false subtest scores on the two forms were 1.20 and .80, corrected for attenuation. In conclusion, there was some support for the conjecture that true-false and multiple choice tests are equally reliable when testing time is equated, and there was no difference between item forms in what was measured. It was recommended that, in future studies of this problem, larger samples should be used in the initial tryout of the true-false items and that more time should be spent revising these items before the final form is administered.

The findings of the research reviewed in this section are far from conclusive. There is no overwhelming evidence to suggest that multiple choice and true-false tests are equally reliable or that one form is superior on this count. Only two of the nine studies reported on the comparative validities of the two item forms. The findings of Charles (1926) and Ebel (1971) lend support to the hypothesis that multiple choice and true-false tests measure the same thing.

Except for the studies by Ebel (1971) and Burmeister and Olson (1966) the research cited here is not recent. There is reason to believe that the nature of objective tests has changed since many of these studies were conducted. Multiple choice tests, in general, probably consist of fewer factual information items than tests constructed in the 1920's. There appears to be a trend

today toward measuring individuals' understandings of concepts and relationships and ability to apply or generalize from learned propositions. This study was devised in an attempt to answer some of the same questions that were asked when objective items first became widely used.

Studies Using Item Conversion Procedures

Only a small number of the studies designed to compare item forms have specified the methods used for constructing items on the same content. Some writers, such as Eurich (1931), gave incomplete details of their procedures. He wrote essay items to cover midterm course material and then developed an acceptable response for each item. The statements in these responses were used to generate stems for completion items. There is no description provided, however, for methods used to formulate true-false or multiple choice items.

The item conversion methods used in this study (as defined in Chapter I) were used by Owens, Hanna, and Coppedge (1970) in a study devised to compare completion and multiple choice geometry items. The judgmental (J), frequency (F), and discrimination (D) methods were employed to select multiple choice distractors based on responses to completion items. Form J was constructed by using 13 secondary mathematics teachers who chose the

three most plausible distractors that appeared as errors in responses to the completion form. Distractors for the final form were selected from those most frequently chosen by the 13 judges. Form F was constructed using the most frequently occurring errors from the completion tryout. Examinee errors from the completion form were used to select the distractors that best discriminated between high and low scorers on the test to build form D. The three 17-item final tests utilized 51 unique distractors of which 13 were common to forms J and F, two overlapped in forms J and D, and 21 were identical in forms F and D. The authors concluded that the three methods were equally valid and reliable for choosing multiple choice distractors. They suggested that the study be replicated with test content as an independent variable.

Loree (1948) used the judgmental and frequency methods for selecting distractors in his study of the characteristics of multiple choice items. The validities and reliabilities of the two multiple choice forms were not significantly different.

The frequency method was used by Burmeister and Olson (1966) in a study cited previously. Multiple choice items were converted to true statements if the incorrect options were equally attractive in the tryout. If one distractor was most frequently chosen the item was changed to a false statement.

Ebel (1971) and Williams and Ebel (1957) employed a discrimination procedure in their item conversion processes. Ebel changed each multiple choice item to a pair of true-false items (one true and one false item) and compiled two true-false test forms. These forms were administered to a group of subjects and the most discriminating item of the original pair was retained for inclusion in the final true-false and multiple choice composite test.

Williams and Ebel (1957) studied the effect on internal consistency reliability of varying the number of response alternatives in multiple choice tests. Item analysis data on 150 four-choice items were utilized to form three-choice and two-choice items. The least discriminating distractors were eliminated from the original test items. The findings revealed no significant differences in reliability on the three forms.

The judgmental and discrimination methods have been used as procedures for selecting multiple choice distractors. The two methods have not been used for converting from multiple choice to true-false form in the studies cited in this section. Although Ebel (1971) used a discrimination procedure for selecting true-false items, he did not employ a systematic and replicable procedure for converting multiple choice items to true-false form.

Studies Comparing Amount of Testing Time

Few recent studies dealing with the comparison of item forms have focused on the amount of testing time required for each form. Two of the studies reviewed here do not supply empirical evidence to support their conclusions. Williams and Ebel (1957) stated that subjects finished faster as the number of response alternatives diminished, but they did not indicate how much faster. In another study it was assumed that subjects typically attempt two true-false items for every multiple choice item tried (Ebel, 1971).

More dated studies by Toops (1921), Watson and Crawford (1930), and Copeland and Gilliland (1943) demonstrated agreement in their findings that three true-false items can be tried for every two multiple choice items attempted.

Two other studies (Charles, 1926; Ruch and Stoddard, 1925) reported on testing time for true-false items and multiple choice items that varied in the number of response alternatives available. The ratio of testing time for an equal number of true-false and multiple choice items can be calculated from the data presented in Table 2.2. The ratios in Charles' study for the five-response and three-response forms are 1.4 and 1.2, respectively. The corresponding ratios from the Ruch and Stoddard study

are 1.6 and 1.3. These results coincide with the findings of the previously cited studies.

Table 2.2. Amount of time required to respond to an equal number of true-false and multiple choice items.

	Time in Minutes	
Item Form	Study 1[a]	Study 2[b]
5-R	25.5	8.0
3-R	21.5	6.8
2-R	19.6	5.7
True-false	18.3	5.1

[a]From Charles (1926).

[b]From Ruch and Stoddard (1925).

There is some agreement in the research reported here that 1.5 true-false items can be attempted in the time required to respond to one multiple choice item. Provision was made in this study to collect data on the number of items attempted by examinees in a fixed period of time because recent empirical evidence was lacking.

CHAPTER III

DESIGN AND PROCEDURES

Introduction

This research study was designed to examine the reliabilities, concurrent validity (correlation between true-false and multiple choice subtest scores), and the amount of testing time required for subjects to respond to true-false and multiple choice social studies and natural science achievement tests. Two methods of converting multiple choice items to true-false form, judgmental and discrimination, were compared to determine if one yielded more reliable true-false test scores than the other.

Sample

The subjects that participated in this study were selected from classrooms in six public high schools located in South-Central Michigan. Schools and classrooms were selected on a voluntary basis; there was no random sampling procedure employed for determining the study sample. The goal of the sample selection scheme used in this study was to identify schools in four types of communities and to choose at least one school from each

strata for inclusion in the study. The four community types were defined by the Michigan Department of Education (1970) as city, town, urban fringe, and rural.

The high school students that took part in this study probably represent a crossection of non-urban high school students in science and social studies achievement levels. The schools from which they were drawn are described briefly in Appendix A.

Three phases of testing were required for instrument development and data collection. Phase I involved gathering item analysis data for an item conversion method and phase II was used to try-out the true-false items. The subjects that participated in phase III, the final testing, are described in Table 3.1. A total of 509 students responded to the social studies tests and 509 students responded to the natural science tests. A minimum of 125 students attempted each of the eight test forms that were administered in the final phase of testing.

Instrumentation

The multiple choice items that were employed in this study appeared in a widely used battery of achievement tests.[1] The social studies items were written to measure

[1]Permission to use these items for this research was obtained from the publisher. The publisher requested that the source of the items not be identified. The test items used for illustrative purposes in this thesis are copyrighted and may not be reproduced.

Table 3.1. Description of sample used in phase III.

School	Grade	Form Totals	
		Social Studies	Natural Science
A	9		25
	10	76	
	11	46	19
	12	45	16
B	9		36
	10	34	67
	11	26	
	12	69	
C	9		23
	10	23	74
	11	32	24
	12	45	
D	9		23
	10	12	
	11		19
	12	43	
E	9		
	10		
	11		67
	12		43
F	9		73
	10		
	11		
	12	58	

knowledge and understanding of contemporary social institutions and practices. The following items are typical of those used in the test:

1. When was the United States Constitution written?

   (1) Immediately after the French and Indian War.

   (2) During the early years of the Revolutionary War.

   *(3) Shortly after the Revolutionary War.

   (4) During the Reconstruction period which followed the Civil War.

2. In the absence of government controls, what ordinarily happens to the price of goods if the supply increases and the demand remains unchanged?

   (1) The price increases.

   *(2) The price decreases.

   (3) The price remains about the same.

   (4) The price changes rapidly.

The natural science items were intended to measure general knowledge and understanding of scientific terms and principles. The following items are representative of those used in the test:

3. What is the chief use of the cyclotron?

   (1) To change lead into gold.

   (2) To generate electricity from steam.

   *(3) To get high speed particles for atomic research.

   (4) To mix the essential ingredients of the atomic bomb.

4. Is it more dangerous to prick oneself with a pin than with a needle? Why?

   (1) Yes. Because pins are usually made of brass, which is poisonous to human flesh.

   (2) Yes. Because bacteria are more likely to be present on a pin than on a needle.

   *(3) No. The two are about equally dangerous.

   (4) No. The needle is much more dangerous for it is likely to have traces of rust on it.

The items from the achievement battery were selected for use in this study and were deemed appropriate for the study sample because:

1. The items were expertly written and were tried-out and revised with extreme care by the authors.
2. A classification of the items by subject matter suggested that the items covered objectives reflecting the current high school science and social studies curricula. (This notion was confirmed by the secondary teachers that reviewed the test content during the three phases of testing.)
3. The reported reliabilities of the social studies and natural science tests were in excess of .90. The tests had demonstrated high reliability in the past.
4. The test items were intended by their authors to be used for measuring achievement in grades 9-13. The tests were suited for a broad range of

achievement and concern about a low ceiling effect could be reduced.

Item Conversion Procedures

The judgmental and discrimination methods defined in Chapter I were used to convert multiple choice items in each of the 70-item achievement tests to true-false statements. The two methods will be described below.

Judgmental Method

Five secondary science and social studies teachers were asked to judge the quality of the multiple choice distractors from the test items in their respective areas of expertise. They were directed to select the distractor for each item that appeared to be most plausible for making a false statement with the original stem. The specific directions given to the judges appear in Appendix B.

The responses of the judges were tabulated and a decision was made to use the correct response or one distractor to make a true or false statement. If at least four of the five judges agreed on a best distractor, it was used to make a false statement. If the judges failed to agree on one best distractor, the correct response was used to make a true statement.

The use of this method resulted in 41 false statements and 29 true statements in social studies. There was

consensus among the judges on their choices for 12 false statements and four of the five agreed on their choices for the other 29 false statements. Item one from the examples listed previously in this chapter was converted to:

> The United States Constitution was written during the early years of the Revolutionary War.

The judges unanimously agreed that response alternative two was the most plausible.

There were 45 false statements and 25 true statements written in natural science. Four of the five judges agreed on their choices for 40 false statements and all were in accord on only five false statements. Item three from the examples listed in this chapter was changed to:

> The chief use of the cyclotron is to mix the essential ingredients of the atomic bomb.

Four of five judges thought choice four was the most plausible.

The two true-false tests developed by the judgmental method were labeled form SJ (social studies) and form NJ (natural science).

Discrimination Method

The original 70-item multiple choice tests were labeled form SM (social studies) and form NM (natural science). Forms SM and NM were each administered to a

minimum of 100 subjects in classrooms from schools that appeared in the final sample. Table 3.2 describes the 103 students that responded to form SM and the 101 students that took form NM in phase I of testing. Answer sheets were scored and responses were put on magnetic tape using the OpScan system of the Office of Evaluation Services at Michigan State University. A computer program developed by the Office of Evaluation Services was used to generate item analysis data from phase I of testing.

Table 3.2. Description of the sample used in phase I.

		Form	
School	Grade	SM	NM
A	9		30
	10		
	11	25	
	12	30	
B	9		19
	10		
	11	48	18
	12		7
C	9		
	10		27
	11		
	12		

Kuder-Richardson Formula 20 reliabilities for forms SM and NM were .905 and .918, respectively.

A decision was made to change each item to a true or false statement depending on the value of the discrimination index of each distractor. A form of the Upper-Lower Index, known frequently by D, was used as a discrimination index. In this case, the proportion in the upper group that responded to each distractor was subtracted from the proportion in the lower group that responded to each distractor. The foil with the largest lower-upper difference was used to make a false statement. If the indices for an item did not differ by more than .09 or if the largest index was less than .20, the item was converted to a true statement.

The 70-item true-false social studies test, labeled form SD, contained 33 true statements and 37 false statements. Item two from the examples listed previously was changed to this true statement:

> In the absence of government controls, if the supply of goods increases and the demand remains unchanged, the price of the goods decreases.

The item analysis data that were used to convert this item are given in Table 3.3. The lower-upper indices for distractors 1, 3, and 4 were .18, .14, and .21, respectively. Though distractor four had an index in excess of .19, it did not satisfy the second criterion and was, therefore, not used to make a false statement.

Table 3.3. Item analysis data used in the discrimination method.

	Response Alternatives					
	1	*2	3	4	Omit	Total
upper	0	28	0	0	0	28
27%	0%	100%	0%	0%	0%	100%
middle	1	41	4	2	0	48
46%	2%	85%	8%	4%	0%	99%
lower	5	11	4	6	2	28
27%	18%	39%	14%	21%	7%	99%

The 70-item true-false natural science test, labeled form ND, consisted of 33 true statements and 37 false statements. Item four from the examples was recast as this false statement:

> It is more dangerous to prick oneself with a pin than with a needle because bacteria are more likely to be present on a pin than on a needle.

Four true-false tests (forms SJ, SD, NJ, and ND) were generated with the two item conversion methods. There were 30 common items to the 70-item forms, SJ and SD, and 21 items common to the 70-item forms, NJ and ND.

True-False Try-Out

The four true-false test forms were each administered to a group of 50 students in phase II of testing. The purpose of phase II was to attempt to identify poor or ambiguous items. The rationale for including this step in

the instrument development sequence was that the original multiple choice items sustained extensive study and revision before they were incorporated into the final form.

Participants in phase II are described in Table 3.4. The four schools included in this phase of testing were also involved in the final phase of testing.

Test scoring and item analysis services were furnished by the Office of Evaluation Services. Kuder-Richardson Formula 20 reliabilities for forms SJ, SD, NJ, and ND were .764, .799, .798, and .764, respectively. Item difficulty and discrimination indices were examined for all items in the four true-false tests. One item common to forms NJ and ND was reworded because it was a negative discriminator. Two items from form SJ and three items from form SD were reworded for the same reason. These revised true-false forms were used to compile the final forms for phase III of testing.

Design

This study was designed with five major principles in mind for controlling extraneous factors that had the potential for introducing error.

1. No student responded to more than one test in social studies or natural science across the three phases of testing.
2. The four test forms in each subject matter area were randomly distributed to subjects within

Table 3.4. Description of subjects used in phase II.

School	Grade	Form			
		SJ	SD	NJ	ND
A	9			12	12
	10	21	21		
	11				
	12				
B	9				
	10			16	16
	11				
	12				
C	9				
	10			8	7
	11			16	19
	12	9	8		
D	9				
	10				
	11	21	21		
	12				

classrooms in an attempt to control for differential abilities and achievement levels of classrooms.

3. The final test forms were arranged in two different orders to control effects that could occur due to one item form continuously preceding the other. This arrangement was also conducive to gathering data on the number of items attempted.

4. In the final phase each subject received both a multiple choice and a true-false subtest score so that individual scores could be correlated.

5. One individual, experienced in test administration procedures, administered all tests in the three phases of testing with standardized directions designed for the separate phases.

The final test forms were arranged so that each subject responded to both multiple choice items and true-false items converted by one of the two methods. The eight 70-item final forms are depicted in Table 3.5. Different orderings of the item subtests within forms are designated by A and B, S refers to social studies, N refers to natural science, and J and D represent judgmental and discrimination, the two item conversion methods. Form SJA, for example, consisted of items 1-35 of the original multiple choice form, SM, and items 36-70 of form SJ, social studies items converted by the judgmental method. Form SJB was comprised of items 1-35 of form SJ and items 36-70 of form SM. The other six forms were arranged in a similar manner.

The final test forms were administered in high school social studies and science classrooms. The four

forms within each subject matter area were randomly distributed to subjects within classrooms so that approximately the same number of each form was used in each classroom. Standardized directions and procedures were used by the same test administrator in all phases of testing in this study.

Table 3.5. Arrangement of test forms used in phase III.

Test Form	Subtest Order	
SJA	MC	TF
SJB	TF	MC
SDA	MC	TF
SDB	TF	MC
NJA	MC	TF
NJB	TF	MC
NDA	MC	TF
NDB	TF	MC

Subjects were timed with a stopwatch to supply information regarding the number of items of each item type that were attempted in a fixed period of time. Subjects were asked to stop working after ten minutes and were then asked to circle in their test booklet the number of the item that they were currently working on. A preliminary examination of this data showed that ten minutes enabled many students to respond to more than 35 items. The time period was subsequently reduced to eight minutes and data were collected from 967 subjects.

Hypotheses

The research hypotheses that were examined in this study were:

1. When multiple choice items in social studies and natural science are converted to true-false form, the Kuder-Richardson Formula 20 reliabilities of the two test forms are not different, regardless of the subject matter.
2. When multiple choice items in social studies and natural science are converted to true-false form using the judgmental and discrimination methods, the Kuder-Richardson Formula 20 reliabilities of the true-false tests are not different, regardless of the subject matter.
3. A group of examinees can attempt more true-false than multiple choice items in eight minutes of testing time.
4. The simple correlation between individual's true-false and multiple choice test scores is 1.00, corrected for attenuation.

Analysis

The Kuder-Richardson Formula 20 reliability coefficient was computed for each of the two subtests in each of the eight final forms. The true-false subtest

reliabilities were adjusted for a lengthened test using the data gathered regarding the number of items of each form that subjects responded to in an eight-minute period. The ratio of the number of true-false items attempted to the number of multiple choice items attempted was substituted in the Spearman-Brown Prophecy Formula for this purpose. The formulas for the Kuder-Richardson Formula 20 (Equation 3.1) and the Spearman-Brown formula (Equation 3.2) are given by Ebel (1965)

$$r = \frac{k}{k-1} \left[1 - \frac{\Sigma pq}{\sigma^2}\right] \tag{3.1}$$

where r is reliability, k is the number of items in the test, Σpq is the sum of the item variances, and σ^2 is the test variance.

$$r_n = \frac{nr_s}{1 + (n-1)r_s} \tag{3.2}$$

where r_n is the reliability of the lengthened test, n is the number of times the original test was lengthened, and r_s is the reliability of the shorter test.

A test statistic known as the paired t test was employed to test the difference between multiple choice and true-false test reliabilities. The statistic, t, was defined as:

$$\frac{\bar{d}}{s_d / \sqrt{n}} \tag{3.3}$$

where $\bar{d}$ is

$$\frac{\sum_{i=1}^{n} (r_{im} - r_{it})}{n} \tag{3.4}$$

and r_{im} and r_{it} are the reliability coefficients for the ith pair of subtests from the eight final test forms; and S_d, the standard error of the differences, is

$$\sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}} \tag{3.5}$$

and d_i is defined as the difference between the multiple choice and true-false reliabilities for each i pair.

$$d_i = r_{im} - r_{it} \tag{3.6}$$

An hypothesis that the multiple choice and true-false test reliabilities are not different is tested against the alternative hypothesis that the null hypothesis is false. The test statistic is referred to Student's t-distribution with n-1 degrees of freedom. Values of the test statistic large in absolute value cause the null hypothesis to be rejected. The alpha level used in all statistical tests in this study was 0.05.

The use of the test statistic, t, depends on the assumption that the sample statistic being tested is

normally-distributed. Though this assumption was not strictly met by the data in this study, the large sample sizes probably overcome that limitation.

Frequency distributions indicating the number of items subjects responded to in eight minutes were constructed based on the A and B test forms. The ratio of the medians of the two distributions was used as evidence for supporting or rejecting the third research hypothesis.

A Pearson product-moment correlation coefficient was computed between individuals' multiple choice and true-false subtest scores on each of the eight forms. The correlation coefficients were adjusted for unreliability in the measurement of the two variables by the correction for attenuation formula given by Ghiselli (1964, p. 268):

$$r_{\infty\infty} = \frac{r_{xy}}{\sqrt{r_{xx}}\,\sqrt{r_{yy}}} \tag{3.7}$$

where $r_{\infty\infty}$ is the true correlation between scores on X and Y, r_{xy} is the correlation between observed scores on X and Y, r_{xx} is the reliability coefficient for the X scores, and r_{yy} is the reliability coefficient for the Y scores.

The corrected correlation coefficients were tested to determine if their values were different from unity. The test statistic used was given by Lord (1957) as:

$$\chi_1^2 = 2.3026\ (N-1)\ \log_{10} \frac{(1-\tilde{\rho}_{12})\ (1-\tilde{\rho}_{34})}{(1-\dot{\rho}_{12})\ (1-\dot{\rho}_{34})}$$

$$\frac{[(1 + \tilde{\rho}_{12})\ (1 + \tilde{\rho}_{34}) - 4\ \tilde{\rho}_{13}]}{[(1 + \dot{\rho}_{12})\ (1 + \dot{\rho}_{34}) - 4\ \dot{\rho}_{13}]} \qquad (3.8)$$

where N is the sample size

$\tilde{\rho}_{12} = \tilde{\rho}_{34} = \tilde{\rho}_{13} = 1/3\ (\dot{\rho}_{12} + 2\dot{\rho}_{13})$

$\dot{\rho}_{12} = \dot{\rho}_{34}$ is the reliability estimate for the two measurements

$\dot{\rho}_{13}$ is the correlation between observed scores for the two measures

Lord's derivation of the above formula depended on the use of the correlation between parallel test forms as a reliability coefficient and it assumed equivalent estimates of reliability for the two measures. The mean Kuder-Richardson Formula 20 reliability coefficient for the true-false and multiple choice subtests in each test form was used as an approximation of $\dot{\rho}_{12}$.

The calculated value of χ_1^2 was referred to the chi-square distribution with one degree of freedom and alpha was preset at the 5% level of significance. The

decision rule for each test was to reject the null hypothesis ($P_{xy}=1$) if $\chi_1^2 > 3.84$.

Summary

The 1018 subjects that participated in the final testing phase of this study were described as representative of non-urban high school students. Each subject responded to one of eight test forms in either social studies or natural science. The 70-item test forms were composed of half multiple choice and half true-false items. The true-false items were converted from multiple choice form by the judgmental and discrimination methods.

Kuder-Richardson Formula 20 reliability coefficients were calculated for all true-false and multiple choice subtests. The ratio of the number of true-false to multiple choice items that subjects attempted in the first eight minutes of testing was computed. Finally, the correlation between individual true-false and multiple choice subtest scores was calculated and corrected for attenuation for each final test form.

CHAPTER IV

RESULTS

Introduction

This chapter is divided into four major sections. The first section deals with the findings regarding the number of multiple choice and true-false items that subjects responded to in eight minutes of testing time.

The second section contains the results relevant to the reliabilities of the multiple choice and true-false subtests. The outcomes associated with the first two research hypotheses are reported separately in section two.

Results that reflect on the concurrent validity of the true-false and multiple choice tests are reported in the third section. A final section, the chapter summary, follows.

Results Concerning Amount of Testing Time

Frequency distributions indicating the number of items subjects responded to in eight minutes were constructed for each of the eight test forms. The medians and number of examinees for each distribution are shown in

Table 4.1. Students, in general, worked more rapidly on the social studies tests than on the natural science tests. Also, students responded to more true-false than multiple choice items in the eight-minute period.

Table 4.1. Median number of items attempted in the first eight minutes of testing for each final test form.

True-False Items Test Form	SJB	SDB	NJB	NDB
Number of examinees	121	120	122	122
Items attempted (Md)	27.44	26.83	23.00	24.83
Multiple Choice Items Test Form	**SJA**	**SDA**	**NJA**	**NDA**
Number of examinees	118	123	120	119
Items attempted (Md)	17.42	17.34	16.42	17.05

The data from the above distributions were combined to form two frequency distributions, one for form A tests and one for form B tests. (See Table A2 in Appendix D for the complete distributions.) The typical performance of subjects on these forms was represented by medians calculated for the two distributions. The median for the

true-false tests, form B tests, was 25.59 and the median for the multiple choice tests, form A tests, was 17.04. The ratio of these medians that serves as an index of the relative rates of work by subjects on the true-false and multiple choice tests was 1.50. The conclusion drawn from these data was that, in general, students attempted three true-false items for every pair of multiple choice items attempted.

Results Concerning Test Reliability

Kuder-Richardson Formula 20 reliability coefficients were computed for each subtest of the eight final test forms. The 16 coefficients are reported in Table 4.2. The true-false subtest reliabilities were adjusted by the Spearman-Brown formula to estimate the reliabilities of tests 1.5 times as long as the original tests. The adjusted reliabilities also appear in Table 4.2.

Hypothesis One

The first hypothesis of interest that was stated in Chapter III was:

H_o^1: When multiple choice items in social studies and natural science are converted to true-false form, the Kuder-Richardson Formula 20 reliabilities of the two tests are not different, regardless of the subject matter.

A visual inspection of the data reported in Table 4.2 indicated that the multiple choice and true-false

reliabilities were different, and in each case the multiple choice reliabilities were higher. The differences noted were tested statistically to determine if these were significant differences.

Table 4.2. Kuder-Richardson Formula 20 reliability coefficients for final subtest forms.

	Subtest		
Test Form	Multiple Choice	True-False Original	True-False Adjusted[a]
SJA	.796	.708	.785
SJB	.827	.654	.739
SDA	.805	.498	.598
SDB	.851	.641	.728
NJA	.835	.759	.825
NJB	.852	.612	.703
NDA	.854	.704	.781
NDB	.862	.645	.732

[a]The adjusted true-false test reliabilities were estimated by the Spearman-Brown formula for a test 1.5 times the length of the original test.

The test statistic, t, reported as Equation 3.3, was used to test the hypothesis H_o^1: $\rho_m = \rho_t$ against the alternative hypothesis that H_o^1 is false. The pairs of reliability coefficients and their corresponding d_i's are included with the computational data in Table 4.3. The computed value of t was 5.520. Since the decision rule

for this test was to reject H_o^1 if $-2.365 \leq t \leq 2.365$ ($t_{(.05)7} = 2.365$), a decision was made to reject H_o^1.

Table 4.3. Computations for testing hypothesis one.

Test Form[a]	r_{im}	r_{it}	d_i	$(d_i-\bar{d})^2$
SJA	.796	.785	.011	.007709
SJB	.827	.739	.088	.000117
SDA	.805	.598	.207	.000067
SDB	.851	.728	.123	.000586
NJA	.835	.825	.010	.007885
NJB	.852	.703	.149	.002520
NDA	.854	.781	.073	.000666
NDB	.862	.732	.130	.000973

$\bar{d} = .0988$

$S_d = .0506$

$n = 8$

[a]Means and variances for the final test forms can be found in Appendix E.

The conclusion based on these data was that the reliabilities of multiple choice and true-false tests were different, and, by inspection, the multiple choice reliabilities were consistently greater.

Hypothesis Two

The second hypothesis of interest that was stated in Chapter III was:

H_o^2: When multiple choice items in social studies and natural science are converted to true-false form using the judgmental and discrimination methods, the Kuder-Richardson Formula 20 reliabilities of the true-false tests are not different, regardless of the subject matter.

An examination of the data reported in Table 4.2 showed that the true-false test reliabilities were relatively homogeneous. The two most extreme values, .825 and .598, favored the judgmental forms.

A paired t test was used to test the hypothesis H_o^2: $\rho_{iJ} = \rho_{iD}$ against the alternative hypothesis that H_o^2 is false. The computations for this test appear in Appendix C. Since the calculated value of the test statistic, 1.307, did not exceed the critical value, $t_{(.05)3} = 3.182$, a decision was made not to reject H_o^2. The conclusion was that the reliabilities of the true-false tests constructed by the judgmental and discrimination methods were not different.

Results Concerning Concurrent Validity

Each subject received a score on the multiple choice and on the true-false subtests of the test form to which he responded. A Pearson product-moment correlation coefficient was calculated between subtest scores on each of the eight final forms. These are presented in Table 4.4. The correlation coefficients were adjusted for

unreliability in the measurement of the two variables by the correction for attenuation formula given as Equation 3.7. These estimates of the correlation between the true scores on the two subtests are also depicted in Table 4.4.

Table 4.4. Correlation coefficients for multiple choice and true-false subtest scores on each final form.

Test Form	r_{mt}	$r_{\infty\infty}$[a]	N
SJA	.578	.769	126
SJB	.697	.947	127
SDA	.564	.891	128
SDB	.430	.582	128
NJA	.661	.831	126
NJB	.728	1.009	129
NDA	.710	.916	125
NDB	.825	1.107	129

[a]Designates r_{mt} corrected for attenuation.

The fourth research hypothesis of interest was stated in Chapter III as:

H_o^4: The simple correlation between individuals' true-false and multiple choice test scores is 1.00, corrected for attenuation.

An inspection of the correlation coefficients in Table 4.4 revealed that two of the corrected correlations exceeded one. The explanation for actual values exceeding the theoretical upper bound of one, according to Lord (1957, p. 208), is sampling fluctuation. Values greater

than unity occur when the correlation between observed scores is larger than the true value or when the observed reliability coefficients are underestimates of their true values.

The test statistic, χ_1^2, reported as Equation 3.8, was used to test the hypothesis H_o^4: $P_{xy} = 1$ against the alternative hypothesis H_1^4: $P_{xy} < 1$ where P_{xy} is the disattenuated correlation coefficient.

Table 4.5 provides the results of the six tests that were carried out, each at the $\alpha = .05$ level. If the alpha level had been reduced initially from .05 to .0001 for each test to favor non-rejection of the null hypothesis, the results would have been the same. Thus, the usual problem of compounding the alpha level for multiple statistical tests did not affect the outcomes in this situation.

Table 4.5. Computations and results of tests for hypothesis four.

Test Form	$\dot{\rho}_{12} = \dot{\rho}_{34}$	$\dot{\rho}_{13}$	$\tilde{\rho}_{12} = \tilde{\rho}_{34} = \tilde{\rho}_{13}$	P_{xy}	χ_1^2
SJA	.791	.578	.649	.769	55.26*
SJB	.783	.697	.726	.947	14.27*
SDA	.702	.564	.610	.891	18.86*
SDB	.789	.430	.550	.582	100.42*
NJA	.830	.661	.717	.831	54.28*
NJB				1.009	
NDA	.818	.710	.746	.916	26.95*
NDB				1.107	

*Significant at $\alpha < .0001$.

The conclusion drawn from the data represented by Table 4.5 was that corrected correlations between individuals' multiple choice and true-false subtest scores were not perfect (equal to 1.00).

Summary

The results of the data analysis for this study were presented in this chapter. The findings concerning the four major research hypotheses were:

1. Students responded to three true-false items for every pair of multiple choice items attempted. In addition, students worked more rapidly on the social studies items than on the natural science items.
2. The Kuder-Richardson Formula 20 reliability coefficients for the multiple choice subtests were greater than those of the true-false subtests.
3. There was no significant difference between the reliabilities of the true-false tests constructed by either the judgmental or discrimination methods.
4. The correlations, corrected for attenuation, between true-false and multiple choice subtest scores were significantly different from unity for six of the eight final test forms.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The purpose of this study was to compare the reliabilities of multiple choice and true-false tests and to determine the concurrent validities of true-false tests that were written to measure understanding of concepts and relationships. The four major questions that were formulated as research hypotheses were:

1. Are multiple choice and true-false achievement tests that were designed to measure the same objectives equally reliable?
2. Are true-false tests that are converted from multiple choice form by the judgmental method as reliable as those converted by the discrimination method?
3. What is the ratio of the number of true-false items attempted to the number of multiple choice items attempted by a group of examinees in a fixed period of time?
4. Is the correlation between individuals' true-false and multiple choice subtest scores perfect (+1.00) when the correlation is corrected for attenuation?

A search of the literature revealed that there were few recent studies concerning the comparison of true-false and multiple choice test reliabilities or validities. The findings of studies completed in the 1920's were incongruous and were based primarily on items that measured factual information. No studies were noted that reported an objective and reproducable procedure for changing items from one form to another. There was agreement in the research cited that 1.5 true-false items could be attempted in the time required to respond to one multiple choice item. No recent empirical evidence was located to substantiate this earlier claim.

A sample of 1018 non-urban high school students in Central Michigan each responded to one of eight test forms constructed to measure social studies or natural science achievement. The original multiple choice items used in this study were selected from a widely used battery of standardized achievement tests.

Two methods were devised for systematically changing multiple choice items to true-false form. The judgmental method involved the use of secondary school teachers to choose the multiple choice distractor that would result in the most plausible false statement. The discrimination method relied on item analysis data from a multiple choice testing to identify the distractor that best discriminated between high and low scorers on the test. The first of

three phases of testing was needed to gather the item analysis data.

The true-false items generated by the two conversion methods were tried out in the second phase of testing. The revised true-false items were incorporated in the eight test forms used in phase III, the final testing. Each of the 70-item final forms consisted of 35 multiple choice and 35 true-false items.

Kuder-Richardson Formula 20 reliability coefficients were calculated for the 16 multiple choice and true-false subtests. The ratio of the number of true-false to multiple choice items that subjects attempted in the first eight minutes of testing was also computed. The correlation between individuals' true-false and multiple choice subtest scores was calculated and corrected for attenuation for each final test form. Statistical tests were performed to determine if the subtest reliabilities were different and to ascertain if the values of the eight corrected correlation coefficients departed significantly from unity.

Conclusions

The reliability and concurrent validity coefficients from the final form subtests are summarized in Table 5.1. The conclusions associated with the four major research hypotheses were:

Table 5.1. Reliability and concurrent validity coefficients for final subtest forms.

Test Form	r_{20}[a]	r_{mt}	$r_{\infty\infty}$	N
SJA_m	.796	.578	.769	126
SJA_t	.785			
SJB_t	.739	.697	.947	127
SJB_m	.827			
SDA_m	.805	.564	.891	128
SDA_t	.598			
SDB_t	.728	.430	.582	128
SDB_m	.851			
NJA_m	.835	.661	.831	126
NJA_t	.825			
NJB_t	.703	.728	1.009	129
NJB_m	.852			
NDA_m	.854	.710	.916	125
NDA_t	.781			
NDB_t	.732	.825	1.107	129
NDB_m	.862			

[a]The true-false test reliabilities were adjusted for a lengthened test by the Spearman-Brown formula.

1. The Kuder-Richardson Formula 20 reliability coefficients were greater for the multiple choice than for the true-false subtests.
2. There was no significant difference between the reliabilities of the true-false tests constructed by either the judgmental or discrimination method.
3. Examinees responded to three true-false items for every pair of multiple choice items attempted. In addition, students worked more rapidly on the social studies items than on the natural science items.
4. The correlations, corrected for attenuation, between true-false and multiple choice subtest scores were significantly different from unity for six of the eight final test forms.

Discussion

The findings of this study are somewhat in agreement with the conclusions drawn by other researchers in recent work. None of the studies previously cited, however, used subject matter as a variable of interest. The reliability coefficients obtained in this study were found to differ depending on item form, and an inspection of the data in Table 5.1 demonstrates that higher reliabilities were observed for the natural science subtests than for the social studies subtests. The corrected

concurrent validity coefficients follow this same trend. The explanation for these observed differences is not readily apparent. It may be true that the content of the high school social studies curriculum is less tightly organized than the subject matter in natural science. Hierarchically-arranged concepts and principles are probably more conducive to measurement with a set of relatively homogeneous items than are more loosely knit units of knowledge. The more heterogeneous social studies test items are likely to produce a lower coefficient of internal consistency than are the natural science items.

The results of the statistical tests employed in the data analysis of this study are probably not of paramount importance. The same conclusions could be drawn from an examination of the data in Table 5.1. The multiple choice and true-false subtest reliabilities within each of the test forms are not extremely different for practical group achievement testing purposes. Only two of the corrected correlation coefficients can be interpreted as perfect correlations. Five of the remaining six are probably too low to consider them near perfect. Sampling fluctuation may be an explanation for the low corrected coefficients. Another possibility, however, is the conjecture that true-false and multiple choice test items do not measure the same thing.

The abilities required of examinees to respond to true-false items are probably different from those needed to obtain the same score on a comparable four-response multiple choice test. For example, an individual may mark a statement true because he could not think of a counter-example, a situation or occurrence that would make the proposition false. His search for a counterexample may have been bounded by time limits or the length to which he could stretch his mind or the depth of his retrieval system that he could penetrate. The multiple choice item, however, limits the universe of comparisons that the individual must make. He can decide which alternative makes a true statement with the item stem and then review the remaining alternatives to determine if any of them is a counterexample for the true statement. Though individuals probably differ in the responding schemes they use, the manners of responding to true-false and multiple choice items probably depend on somewhat different abilities. These differences in abilities required may be reflected in the test scores and, therefore, in the correlation coefficient.

The data from this study indicated that examinees' rates of work varied with item form and content. Students worked more rapidly on the true-false tests than on the multiple choice tests and they responded more rapidly to social studies items than natural science items. One

practical application of these findings is that teachers could maximize their classroom testing time by conducting small experiments to find out the rates of work of their classes based on the item forms the teacher typically uses. Those individuals that construct achievement tests should be aware that rate of work may vary with item form, content, and difficulty.

The findings of this study suggested that test quality may be somewhat sacrificed by using true-false rather than multiple choice items. A practical consequence of this finding is that a longer, though perhaps a bit less reliable, test may be used for a given period of testing if the items are in true-false rather than multiple choice form. A true-false test may be a feasible alternative if the examiner is primarily concerned about the adequacy with which his sample of items represents the universe of content. A longer test can probably effect a more thorough sampling of the universe. If, however, the examiner was not willing to sacrifice reliability, a 52-item true-false test with a reliability of .739 would theoretically need to be lengthened to 89 items to obtain a reliability coefficient of .827 that was achieved with a 35-item multiple choice test. (The assumptions required by the Spearman-Brown formula (Equation 3.2) would have to be considered in judging how much confidence to place in such an inference.)

Some teachers express the notion that unambiguous true-false items are more difficult to prepare than multiple choice items. Good items of both types are not actually easy to construct. The individual that finds more difficulty writing true-false items might utilize one of the two item conversion methods employed in this study to make true-false statements. The judgmental method might be an attractive procedure when examinations are prepared as a departmental effort. Only one good distractor is necessary to write a false statement; at least one good distractor is required for an adequate two-response multiple choice item. Also, good four-response multiple choice items can be converted to one true and three false statements. The converted items can be used to build a sizeable item bank.

The results of this study depart from the findings of Storey (1966, p. 285) who concluded his article by writing that "only a trifler and the uninformed pretend to measure anything with the relatively invalid, unreliable, and subject-to-set true-false item." Storey's remark is unwarranted, however, because he did not compare true-false items with any other item form in his study.

Limitations of the Study

The schools and classrooms that participated in this study did so on a voluntary basis. The

generalizability of the findings to other classrooms is left to the reader's discretion since a population was not clearly specified.

Also natural science and social studies tests were arbitrarily selected for this study. The findings should not be indiscriminately generalized to tests covering vastly different subject matter.

The motivation of the subjects to reflect their true achievement levels on the tests is questionable. Students were directed not to guess blindly on any test item but there is no way to determine the extent to which those directions were followed. Fewer than five examinees were observed randomly recording their responses throughout the three phases of testing. Less obvious cases of blind guessing are not easily detected.

The sample sizes for each of the final test forms were probably too small to control the sampling fluctuation that plagues the interpretation of correlation coefficients. Samples approaching 300 subjects would undoubtedly have been preferable to groups of 125.

Suggestions for Future Research

The following suggestions are offered for further investigation into the comparative effectiveness of item forms.

1. The range of true-false subtest scores was restricted compared to the range for multiple choice scores because the chance scores were different for the two. It would be appropriate for future studies to use a lengthened true-false test to estimate reliability instead of estimating the reliability for a hypothetically-lengthened test. This procedure should increase the variability of the true-false scores and, perhaps, produce a better estimate of the relationships between multiple choice and true-false scores.

2. The use of conversion methods to construct content-equivalent items could be extended to include a frequency method and a random method. The frequency method would entail using the multiple choice distractor most frequently chosen by examinees to make a false statement. The random method would involve selecting a distractor at random to make a false statement.

3. The amount of testing time required for a given number of items could be investigated for various item forms in several subject matter areas with item difficulty used as a control variable.

4. The results of this study suggested differences in reliability and concurrent validity associated with the subject matter content of the test. Future investigations might profitably be designed to determine the causes of these differences.

5. Research is needed to investigate the sampling fluctuation that sometimes interferes with the interpretation of the disattenuated correlation coefficient. Perhaps a monte carlo study in which sample size is a variable would shed some light on this matter.

APPENDICES

APPENDIX A

DESCRIPTION OF SCHOOLS PARTICIPATING IN THIS STUDY

Table A1. Description of schools participating in this study.

School	Grade Levels	Number of Students	Number of Teachers in School	Number of Students in the School District
A	9-12[a]	960	40	3489
B	9-12	452	24	1621
C	7-12	885	40	1914
D	9-12	1120	62	3320
E	9-12	1251	54	4112
F	9-12	1300	67	3763

[a]This data was compiled from the Michigan Education Directory and Buyer's Guide (1970).

APPENDIX B

DISTRACTOR JUDGMENT TASK

Distractor Judgment Task

Directions:

The following pages contain 70 multiple choice items with the correct response or best answer circled for each item. For each item you are to choose from the remaining alternatives that response which, in your judgment, would appear most attractive to a student that does not possess sufficient knowledge to respond to the item correctly.

You might go about this task by thinking: "If I were going to make a false statement out of this item, which of the incorrect responses would provide me with the most plausible statement. Or, which statement would an uninformed student be most willing to accept as true."

Please respond to each of the 70 items in this fashion. If two or more incorrect responses seem to be equally plausible, select one of them for your final choice on a random basis.

Mark your choice on the enclosed answer sheet as if you were taking this test. (Of course, none of your responses will be correct when you have finished.) You need not fill in any of the identification blanks on the answer sheet.

Thank you for your cooperation. Your promptness in completing this task will be appreciated.

APPENDIX C

COMPUTATIONS FOR TESTING HYPOTHESIS TWO

Table A2. Computations for testing hypothesis two.

Test Form	r_J	r_D	d_i	$(d_i-\bar{d})^2$
SA	.785	.598	.187	.017902
SB	.739	.728	.011	.001781
NA	.825	.781	.044	.000085
NB	.703	.732	-.029	.006757

$\bar{d}$ = .0532

S_d = .0814

n = 4

APPENDIX D

NUMBER OF ITEMS ATTEMPTED IN THE FIRST EIGHT MINUTES OF TESTING

Table A3. Number of items attempted in the first eight minutes of testing.

Number of Items Attempted	Multiple Choice N=480	True-False N=487
35 or more	0	56
34	0	8
33	1	16
32	0	16
31	1	11
30	1	25
29	1	22
28	3	29
27	3	21
26	8	38
25	7	19
24	8	20
23	16	26
22	17	24
21	32	34
20	45	25
19	33	15
18	36	21
17	57	19
16	51	8
15	47	13
14	35	2
13	20	3
12	15	2
11	21	0
10	8	0
9 or less	10	4

APPENDIX E

MEANS AND VARIANCES FOR THE SUBTESTS OF THE FINAL TEST FORMS

Table A4. Means and variances for the subtests of the final test forms.

Test Form	Mean	Variance
SJA_m	24.31	30.67
SJA_t	19.55	25.86
SJB_t	24.48	18.09
SJB_m	16.92	43.02
SDA_m	24.21	31.65
SDA_t	19.61	15.42
SDB_t	22.57	18.17
SDB_m	17.44	48.83
NJA_m	21.62	40.88
NJA_t	19.71	30.61
NJB_t	20.78	18.05
NJB_m	17.83	47.32
NDA_m	22.19	44.35
NDA_t	21.00	24.69
NDB_t	19.36	20.97
NDB_m	16.79	50.63

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmann, J. S. and Glock, M. D. Evaluating Pupil Growth. 3d ed. revised. Boston: Allyn and Bacon, Inc., 1967.

Andrews, D. M. and Bird, C. "Comparison of Two New-Type Questions: Recall and Recognition," Journal of Educational Psychology, XXIX (March, 1938), pp. 175-193.

Brown, F. G. Principles of Educational and Psychological Testing. Hinsdale: The Dryden Press, Inc., 1970.

Burmeister, M. A. and Olson, L. A. "Comparison of Item Statistics for Items in Multiple Choice and in Alternate-Response Form," Science Education, L (December, 1966), pp. 467-470.

Charles, J. W. "A Comparison of Five Types of Objective Tests in Elementary Psychology." Ph.D. Thesis, State University of Iowa, 1926.

Choppin, B. H. and Purves, A. C. "Comparison of Open-Ended and Multiple Choice Items Dealing With Literary Understanding," Research in the Teaching of English, III (Spring, 1969), pp. 15-24.

Copeland, J. S. and Gilliland, A. R. "Comparison of the Validity and Reliability of Three Types of Objective Examinations," Journal of Educational Psychology, XXXIV (April, 1943), pp. 242-246.

Cronbach, L. J. "Experimental Comparison of the Multiple True-False and Multiple Multiple Choice Tests," Journal of Educational Psychology, XXXII (October, 1941), pp. 533-543.

Durost, W. N. and Prescott, G. A. Essentials of Measurement for Teachers. New York: Harcourt, Brace and World, Inc., 1962.

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

Ebel, R. L. "Case For True-False Test Items," School Review, LXXVIII (May, 1970), pp. 373-389.

_____. "The Comparative Effectiveness of True-False and Multiple Choice Achievement Test Items." Paper presented at the American Educational Research Association Annual Meeting, New York City, February, 1971.

Eurich, A. C. "Four Types of Examinations Compared," Journal of Educational Psychology, XXII (1931), pp. 268-278.

Feldt, L. S. "A Test of the Hypothesis That Cronbach's Alpha or Kuder-Richardson Coefficient Twenty Is the Same for Two Tests," Psychometrika, XXXIV (September, 1969), pp. 363-373.

Ghiselli, E. C. Theory of Psychological Measurement. New York: McGraw-Hill Book Company, 1964.

Gronlund, N. E. Measurement and Evaluation in Teaching. New York: The MacMillan Company, 1965.

Heim, A. W. and Watts, K. P. "Experiment on Multiple Choice Versus Open-Ended Answering in a Vocabulary Test," British Journal of Educational Psychology, XXXVII (November, 1967), pp. 339-346.

Lord, F. M. "A Significance Test for the Hypothesis That Two Variables Measure the Same Thing Except for Errors of Measurement," Psychometrika, XXII (September, 1957), pp. 207-220.

Loree, M. R. "A Study of a Technique for Improving Tests." Ph.D. Thesis, University of Chicago, 1948.

Marascuilo, L. A. "Large Sample Multiple Comparisons," Psychological Bulletin, LXV (1966), pp. 280-290.

Michigan Department of Education. Levels of Educational Performance and Related Factors in Michigan. Lansing: Assessment Report No. 4, 1970, pp. 19-26.

Michigan Education Directory and Buyer's Guide. Lansing: Michigan Education Directory, 1970.

Mood, A. M. and Graybill, F. A. Introduction to the Theory of Statistics. 2d ed. New York: McGraw-Hill Book Company, 1963.

Owens, R. E.; Hanna, G. S.; and Coppedge, F. L. "Comparison of Multiple Choice Tests Using Different Types of Distractor Selection Techniques," Journal of Educational Measurement, VII (Summer, 1970), pp. 87-90.

Ruch, G. M. and Stoddard, G. D. "The Comparative Reliabilities of Five Types of Objective Examinations," Journal of Educational Psychology, XVI (1925), pp. 89-103.

Rutledge, R. E. "The True-False Examination in Elementary Psychology With Suggestions for Its Improvement." Ph.D. Thesis, University of California, 1926.

Storey, A. G. "Review of Evidence or the Case Against the True-False Item," Journal of Educational Research," LIX (February, 1966), pp. 282-285.

Thorndike, R. L. and Hagen, E. Measurement and Evaluation in Psychology and Education. 3d ed. revised. New York: John Wiley and Sons, Inc., 1969.

Toops, H. A. "Trade Tests in Education," Teachers College Contribution to Education. New York: Teachers College, Columbia University, No. 115, 1921.

Watson, D. R. and Crawford, C. C. "Four Types of Tests," High School Teacher, VI (September, 1930), pp. 282-283.

Wesman, A. G. "Writing the Test Item." Chapter 4 in Thorndike, R. L. (ed.). Educational Measurement. 2d ed. Washington: American Council on Education, 1971.

Williams, B. J. and Ebel, R. L. "The Effect of Varying the Number of Alternatives Per Item on Multiple Choice Vocabulary Test Items," Fourteenth Yearbook of the National Council on Measurements Used In Education, Princeton, 1957, pp. 122-125.

General References

Allen, D. W. "Quick Scoring, Less Guessing on True-False Tests," Clearinghouse, XXXIII (October, 1958), pp. 74-76.

Bayless, E. E. and Bedell, R. C. "A Study of Comparative Validity as Shown By a Group of Objective Tests," Journal of Educational Research, XXIII (1931), pp. 8-16.

Burkheimer, G. J.; Zimmerman, D. W.; and Williams, R. H. "Maximum Reliability of a Multiple Choice Test as a Function of Number of Items, Number of Choices, and Group Heterogeneity," Journal of Experimental Education, XXXV (Summer, 1967), pp. 89-94.

Carter, H. D. and Crone, A. P. "The Reliability of New-Type or Objective Tests in a Normal Classroom Situation," Journal of Applied Psychology, XXIV (1940), pp. 353-368.

Feldt, L. S. "The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty," Psychometrika, XXX (September, 1965), pp. 357-370.

Hurd, A. W. "Comparison of Short Answer and Multiple Choice Tests Covering Identical Subject Content," Journal of Educational Research, XXVI (September, 1932), pp. 28-30.

Karraker, R. J. "Knowledge of Results and Incorrect Recall of Plausible Multiple Choice Alternatives," Journal of Educational Psychology, LVIII (February, 1967), pp. 11-14.

Kinney, L. B. and Eurich, A. C. "Summary of Investigations Comparing Different Types of Tests," School and Society, XXXVI (October 22, 1932), pp. 540-544.

Magill, W. "The Influence of the Form of Item on the Validity of Achievement Tests," Journal of Educational Psychology, XXV (1934), pp. 21-28.

Miklich, D. R. and Gordon, G. P. "Test-Taking Carefulness vs. Acquiescence Response Set on True-False Examinations," Educational and Psychological Measurement, XXVIII (Summer, 1968), pp. 545-548.

Millman, J. and Setijadi. "Comparison of the Performance of American and Indonesian Students on Three Types of Test Items," Journal of Educational Research, LIX (February, 1966), pp. 273-275.

Payne, W. H. and Anderson, D. E. "Significance Levels for the K-R_{20}: An Automated Sampling Experiment Approach," Educational and Psychological Measurement, XXVIII (Spring, 1968), pp. 23-39.

Preston, R. C. "Multiple Choice Test as an Instrument in Perpetuating False Concepts," Educational and Psychological Measurement, XXV (Spring, 1965), pp. 111-116.

Remmers, H. H. and Remmers, E. M. "The Negative Suggestion Effect on True-False Examination Questions," Journal of Educational Psychology, XVII (1926), pp. 52-56.

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

Shulson, V. and Crawford, C. C. "An Experimental Comparison of True-False and Completion Items," Journal of Educational Psychology, XIX (1928), pp. 580-583.

Storey, A. G. "Versatile Multiple Choice Item," Journal of Educational Research, LXII (December, 1968), pp. 169-172.

Wood, B. D. "Studies in Achievement Tests," Journal of Educational Psychology, XVII (1926), pp. 1-22.