This is to certify that the

dissertation entitled

The Comparative Reliability and Validity of Alternate-Choice and Multiple-Choice Tests

presented by

Timothy J. Van Susteren

has been accepted towards fulfillment of the requirements for

the Ph.D. degree in Measurement, Evaluation, and Research Design

Irwin J. Lehmann
Major professor

Date December 2, 1985

THE COMPARATIVE RELIABILITY AND VALIDITY OF ALTERNATE-CHOICE AND MULTIPLE-CHOICE TESTS

By

Timothy J. Van Susteren

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1986

ABSTRACT

THE COMPARATIVE RELIABILITY AND VALIDITY OF ALTERNATE-CHOICE AND MULTIPLE-CHOICE TESTS

By

Timothy J. Van Susteren

Ebel (1980) proposed a new item format termed alternate-choice which he suggested would compare quite favorably with other conventional test item formats with regard to difficulty, discrimination, reliability and validity, yet have the advantage of being easier to write. While this unique item form proposed by Ebel appeared to show potential as an important addition to the repertoire of test item formats item writers have at their disposal, little empirical research existed to substantiate Ebel's claim.

The purpose of this study was to compare the reliability and validity of alternate-choice and multiple-choice tests that were written to measure understanding of concepts and relationships in educational psychology. The difficulty and discrimination of the two formats was also investigated, and examinees' perceptions of the items was explored.

Timothy J. Van Susteren

In this study a series of examinations composed of alternate-choice and multiple-choice subtests were administered to a group of students enrolled in an introductory course in educational psychology. Students were timed to identify the number of alternate-choice and multiple-choice items to which they were able to respond in a given time period, and a questionnaire was administered to the students to explore their perceptions of the items.

The results of the study indicated that the alternate-choice items compared favorably with the multiple-choice items. While the alternate-choice items were easier than the multiple-choice items, they discriminated as well and were as reliable. Also, the alternate-choice items were more efficient, since students were able to respond to three alternate-choice items to every two multiple-choice items. The concurrent validity of the alternate-choice tests did not equal that of the multiple-choice tests, but the validity of both forms was quite acceptable. In addition, students viewed both forms quite positively and did not express a preference for one form over the other. The use of alternate-choice items to measure educational achievement is recommended.

DEDICATION

This dissertation is dedicated to the memory of Dr. Robert L. Ebel whose pursuit of more precise methods of measuring "useful verbal knowledge" led to his conception of the alternate-choice item format. I will always remember Dr. Ebel's charm, wit, and intellect; and be grateful for his advice, counsel, and friendship throughout my doctoral program.

ACKNOWLEDGMENTS

Sincere thanks are extended to Dr. Irvin Lehmann, Chairman of the dissertation committee for the excellent advice and editorial assistance he provided. I am also grateful to each of the dissertation committee members--Dr. Willam Mehrens, Dr. Leroy Olson, and Dr. Joseph Levine--for their contributions.

Dr. J. Bruce Burke deserves special thanks for his constant encouragement and moral support throughout the research project, as well as for his assistance in writing items and gathering data.

I will never be able to adequately express my gratitude to my wife, Lynn, for the support she gave me during my graduate studies. Her love and understanding represent an important contribution to my successful completion of a doctoral program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

CHAPTER I

THE PROBLEM

Alternate Forms of Objective Test Items

Educators and measurement specialists are constantly seeking more versatile and efficient methods of measuring knowledge. With the objective of achieving more precise and valid measurements, researchers have investigated the psychometric properties of essay, multiple choice, true-false, matching, completion, and various novel combinations of these forms in a variety of testing situations and subject matter disciplines (Charles, 1926; Ebel, 1975; Grosse & Wright, 1985; Meihoff & Mehrens, 1985). While reliability and validity are generally considered the ultimate criteria for judging the overall merits of an item form, researchers have also considered the ease or difficulty of writing or producing the item, the adaptability of an item form to a variety of measurement goals and objectives, and examinees' preferences and perceptions of an item type as important criteria in evaluating the practical usefulness of the form.

Of the various item forms mentioned above, multiple-choice items are by far the most popular (Wesman, 1971). It has been demonstrated that tests

composed of multiple-choice items can be reliable and valid, and that multiple-choice items can be readily employed to test almost any subject matter (Ebel, 1980). It has also been demonstrated that multiple-choice items can be conceived to measure complex mental processes (Bloom, 1958).

While there are many advantages associated with the use of multiple-choice items, there are also difficulties involved. One of the problems most often cited is the difficulty of producing a sufficient number of plausible distractors. Mehrens and Lehmann (1984) point out that:

> Although any test item is good or bad depending on the clarity of expression, the multiple-choice item must have, in addition to a stem that is clear and free from ambiguity, a correct answer and a set of plausible responses. The value of a multiple-choice item depends to a large degree on the skill with which the various distractors are written. (p. 279)

It is apparent that a large portion of the difficulty involved in writing good multiple-choice items is associated with the item writer being able to conceive of a sufficient number of alternatives that are plausible enough to attract the less knowledgeable examinee, yet not so close to the correct answer as to make the item ambiguous and confusing to even the most knowledgeable examinee (Plake & Huntley, 1984).

The related issue of the optimal number of alternatives that should be provided to maximize

efficiency (efficiency is commonly defined as the number of items that examinees are able to answer in a fixed time period) and reliability has been debated from the time multiple-choice items were first introduced to the present (Ruch & Stoddard, 1925; Budescu & Nevo, 1985). Most of the recent empirical studies seem to provide evidence favoring the use of three-choice items over items offering four or five alternatives (Wilson, 1982), with some studies suggesting that tests composed of two-choice items can provide satisfactory reliability (Williams & Ebel, 1957). Grier (1975) summarizes the issue of the optimal number of alternatives and concludes that items with shorter stems and fewer alternatives are frequently found to be more reliable than longer items with more alternatives (four and five choice items). He also notes that there are at least two other reasons that shorter items with fewer distractors are preferable. First, it is usually easier to produce one or two plausible distractors than three or four. Second, there may be a ". . . gain in efficiency, since students might not get lost reading many alternatives and have to return and re-read the question and early alternatives" (1975, p. 112). Employing the same logic that Grier (1975) provides in favor of three-choice items, Ebel (1980) has recently proposed a two-choice item form, termed alternate-choice items.

Alternate-Choice Items

Ebel (1980) explains that alternate-choice items are based on a single proposition rather than complex situations, and that they offer only two alternatives instead of the conventional three, four, or five. Alternate-choice items also differ from multiple-choice items, including the conventional format for two-choice items, in that they include the responses as segments of a continuous sentence rather than listing them in a column under the stem. For example,

> The items teachers write for their classroom tests are likely to be too *a) variable b) uniform in difficulty.
>
> Indices of item difficulty tend to vary *a) less b) more from one group of students to another than do indices of discrimination.
>
> Adrian has finally learned to take turns with classroom toys. In order to maintain this appropriate behavior, his teacher should praise him a) often *b) occasionally.
>
> The concepts of overlearning and satiation are a) very similar *b) distinctly different.

Ebel (1980) notes that in many situations there are good reasons for favoring the use of alternate-choice items. He states that:

> Alternate-choice items have some important advantages over the more familiar multiple-choice item form that offers four answer choices. They are more efficient in that they yield more scorable responses per unit of testing time. They are easiery to write, because they only require two alternate answers.

> Often the important questions an item writer would like to ask have only two plausible alternative answers. A problem is either major or minor, simple or complex. Action in response to it is reasonable or unreasonable. The President either supports or opposes a restriction on the import of foreign automobiles. The birth rate in Russia is higher or lower than that in the United States, increasing the homogeneity of the items in a test either increases or decreases the reliability of the test scores. (p. 115)

According to Ebel, polar alternatives, such as those mentioned above, are very common in real life. The unique format of alternate-choice items allows the item writer to pose these realistic alternative questions in test items free of the constraints involved when he/she must conceive of two or three additional plausible distractors.

In a preliminary study, Ebel (1982) found that alternate-choice items compare quite favorably with true-false items. The results of that study indicated that tests composed of alternate-choice items tend to be (a) easier, (b) more highly discriminating, and (c) demonstrate higher reliability than true-false tests. He also notes that students seemed to prefer alternate-choice items and perceive them to be less ambiguous than true-false items.

Need for this Study

As mentioned in the beginning of this chapter, there has been a continuing debate among measurement specialists on the psychometric advantages of various test item formats. In that debate, true-false and multiple-choice items are among the most commonly discussed (Wesman, 1971). Proponents of true-false items note that this form is comparatively easy to write and is quite efficient, but concede that true-false items tend to be viewed as ambiguous and best suited to testing factual recall. Multiple-choice items have the advantage of versatility and can be written to measure almost any cognitive objective. They are, however, less efficient than some other objective forms, such as true-false, and are among the most difficult to write in that it is often difficult to provide a sufficient number of plausible distractors (Mehrens & Lehmann, 1984).

Ebel (1980) has recently proposed a unique test item form termed alternate-choice items. He contends that alternate-choice items reflect the advantages of both true-false and multiple-choice items in that they are very efficient, quite versatile, and can be written to measure the important elements of any subject matter. Ebel also suggests that as a result of the unique format of the item form, alternate-choice items are easier to write than either multiple-choice or true-false items.

This unique item form proposed by Ebel appears to show potential as an important addition to the repertoire of test item forms that teachers and other item writers have at their disposal, but since alternate-choice items are new, there has been little research conducted with them. As previously noted, researchers have investigated the psychometric properties of multiple-choice, true-false, and various other test item forms in a variety of testing situations and subject matter disciplines. The accumulated knowledge gained from these studies has helped item writers to provide more precise and valid measurements using these item formats. A need exists for more psychometric investigation of alternate-choice items. The empirical data that studies of this type can provide is vital for educators and other test item writers to evaluate the usefulness of alternate-choice items and to determine the potential of this new item form for measuring knowledge in various settings.

Purpose of this Study

The purpose of this study was to evaluate the performance of alternate-choice items relative to multiple-choice items in a large-scale college course testing program. The study compared the reliability and

concurrent validity of alternate-choice and multiple-choice test items written to measure understanding of concepts and relationships in the same content area. The difficulty and discrimination of the two item forms was also explored and information on the efficiency of alternate-choice and multiple-choice items was collected. This information was used to determine the optimal length of the alternate-choice tests. The reliabilities of these lengthened tests were then compared to the reliabilities of the multiple-choice tests with testing time held constant. Because examinees' preference and perceptions have been considered an important evaluative criteria (Frisbee & Sweeney, 1982; Ward, 1982), data was collected in this study to investigate examinees' preference for alternate-choice and multiple-choice items.

The information provided by the results of this study will be useful in the evaluation of the psychometric merits of alternate-choice items. The study is, however, not without limitations. In discussing the reasons for the relatively small number of studies researching test items, Wesman (1971) notes:

> That research studies have contributed little to item writing is not very surprising. The inherent difficulties in conducting penetrating and generalizable studies may not be insurmountable, but they are far from easily resolved. It is the sophisticated recognition of these difficulties that is largely responsible for the paucity of attempts at basic research. (1971, p. 84)

In addition to the limitations common to many research studies associated with random sampling and sample size and composition, this study has at least two other limitations that Wesmann (1971) cites as often associated with basic research on test items. They are the inability 1) to control for the fact that some concepts or subjects may lend themselves more readily to one test item type or format than they do to others, and 2) to account for the fact that the skill of the item writer may account for a large portion of any differences detected. These limitations are discussed more fully in the final section of this paper, but are mentioned here in order to provide an additional dimension to the problem.

Hypotheses

The research hypotheses of the study were:

1. There is no difference in the difficulty indices of alternate-choice and multiple-choice tests.
2. There is no difference in the discrimination indices of alternate-choice and multiple-choice tests.
3. Examinees will attempt the same number of alternate-choice and multiple-choice items in a fixed time period.

4. There is no difference in the internal consistency reliabilities of alternate-choice and multiple-choice tests.
5. The Pearson product-moment correlation between individuals' alternate-choice and multiple-choice test scores is 1.00 when corrected for attenuation.
6. There is no difference in examinees' preference for alternate-choice and multiple-choice tests.

Overview

In Chapter II the literature relevant to the general problem and to specific hypotheses is reviewed. The design of the study, the sample, the instrumentation, and the method of analysis are presented in Chapter III. In Chapter IV the results of the study are discussed. This 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

The relative merits of various test item formats is considered an important subject and is the topic of a great deal of debate among educators and measurement specialists. This is evidenced by the fact that research pertaining to this debate can be found in the literature from the 1920's, when objective test items first became popular, to the present. Chapter II of this study presents a general survey of the research comparing different test item formats.

One of the most persistent and consistently raised questions in achievement testing concerns the concurrent validity of various test item formats and asks whether the item format employed affects the attribute measured by the test. Researchers have attempted to investigate this question by comparing item types that require examinees to select a response with items that require examinees to produce a response. The first section of Chapter II, therefore, is devoted to reviewing the literature comparing test item formats, such as multiple-choice, that require examinees to select a response from a list of options with items requiring examinees to supply a correct response from memory,

commonly referred to as free-response or constructed-response items. Researchers have also been greatly concerned with comparing the reliability that may be achieved using various objective test item formats. Since alternate-choice items are a new form, no specific mention is made of them in the literature. Accordingly, the second section of this chapter includes studies comparing the reliabilities of other popular test item formats. The third section of this chapter contains a review of research dealing with the testing time required by examinees to respond to different test item forms. In the final section, a number of research studies are cited pertaining to examinees' preference when they are presented with two or more types of items. A general summary at the end of the chapter provides a review of Chapter II and an introduction to Chapter III.

Studies Comparing Multiple-Choice Test Items with Constructed-Response Type Items

Teachers and other test users often report an intuitive belief that different types of test items measure different types or levels of knowledge. They believe that test items requiring examinees to produce an answer to a test question comprise not only an inherently more difficult task, but one requiring an entirely

different mental process (cognitive task) than test items requiring examinees to recognize and select the correct answer from a list of options. Several researcheres have sought to gather evidence relevant to this hypothesis. For example, Heim and Watts (1967) and Cook (1955) compared multiple-choice and completion vocabulary items and found the principle difference between the two forms is that the items requiring examinees to supply the answer tended to be somewhat more difficult. These results correspond to those of a similar study by Andrews and Bird (1938) in psychology terminology.

Rowley (1974) compared student responses to multiple-choice and free-response tests of vocabulary and mathematics. The results of his study also suggest that the free-response items were generally more difficult. In addition, Rowley discovered that the use of multiple-choice items to test vocabulary may favor examinees high on testwiseness and/or risktaking. In measuring mathematical achievement, no evidence was found to suggest that the multiple-choice scores differed in any systematic way from the free response scores. Rowley speculates that the risktaker may have been able to benefit from informed guesses on the multiple-choice vocabulary test to a greater extent than on the multiple-choice mathematics test where his/her guesses

were more truly random. He points out that testwise students who do not know the correct answer to a vocabulary item can often eliminate one or more of the options on the basis of partial knowledge and guess among those remaining. This is seldom the case with mathematics items, especially when the student is required to perform some mathematic operation to arrive at the answer. The results of a study by Rocklin and Thompson (1985) supported Rowley's conclusions and also detected an interaction between test anxiety, test difficulty and item format. As might be expected, students high on test anxiety found free-response items more difficult and anxiety producing than multiple-choice items.

The format of items testing mathematics was also the subject of a study performed by Oosterhof and Coats (1984) in which they investigated the difficulty and internal consistency of free-response and multiple-choice items used in an undergraduate course in business finance. The results of their study indicated that the free-response mathematics items were more difficult and reliable than the multiple-choice items employed. The authors note that the comparatively better performance of the completion items used in their study may have been at least partially due to the fact that the probability of

an ignorant examinee guessing the correct answer to a completion-type mathematics item is extremely low.

Ward (1982) compared the reliability (coefficient alpha) and concurrent validity of free-response and multiple-choice verbal aptitude test items. His study confirmed earlier findings that free-response items tend to be somewhat more difficult than multiple-choice items and concludes:

> This study has shown that it is possible to develop open-ended forms of verbal aptitude item types that are approximately as good, in terms of score reliability, as multiple-choice items and that require only slightly greater time limits than do the conventional items. These open-ended items, however, provide little new information. There was no evidence whatsoever for a general factor associated with the use of a free-response format. (1982, p. 9)

Ward emphasizes that the results provide no evidence of a verbal production factor and also notes that the results of the concurrent validity study indicated that both multiple-choice and free-response items can be valid measures of verbal aptitude.

Traub and Fischer (1977) also found high correlations across multiple-choice and free-response formats. Their results were similar to Ward's (1982) in that they concluded that the item format employed does not affect the attribute measured by the test. Similarly, Choppin and Purvis (1969) found multiple-choice and open-ended items equally valid in the measurement of their students' knowledge of literature.

The studies cited above provide a strong indication that multiple-choice and free-response test items can be written to measure the same thing. While the studies cited generally agree that free-response items tend to be slightly more difficult, there is no evidence to indicate that requiring students to supply a response to a question comprises a different or higher-order mental task than requiring them to select a response from a list of options.

General Studies Comparing the Reliability of Various Objective Test Item Formats

Researchers have expressed interest in comparing various objective test item formats from the early part of the century to the present. In 1921 Toops conducted what may have been the first study of this type. He compared the reliabilities of several general information tests composed of fifty items, each cast into free-response, multiple-choice and true-false forms. Each of six groups took each test with the order of administration randomly assigned. The split-half reliabilities reported for the tests were very similar ranging from .507 to .556.

In another study, Charles (1926) compared the split-half reliabilities of five-, three-, and

two-response multiple-choice tests and a true-false test with the reliability of a free-response test. He administered 50 factual information items of introductory psychology to each subject in completion form followed by 50 items in one of the other forms. Charles did not perform any statistical significance tests, but did conclude that there existed little difference of practical significance between the reliabilities of the item formats.

Ruch and Stoddard (1925) employed a design identical to Charles' but used items intended to measure knowledge of history and social science. They discovered that the split-half reliabilities for the 100 item tests composed of five-, three- and two-choice multiple-choice items and true-false items were .886, .748, .849, and .714, respectively. The researchers found that the number of items that students were able to answer varied with the number of options posed in the item. Students were able to answer approximately 1.5 true-false items for every single five-choice multiple-choice item. Therefore, they elected to recalculate the reliabilities equating them for testing time using the familiar Spearman-Brown Prophecy Formula. The corrected relabilities were estimated to be .901, .806, .902, and .820 for the five-, three-, and two-choice and the

true-false items respectively. They offered no explanation for the especially good performance of the two-choice items.

Watson and Crawford (1930), Copeland and Gilliland (1943), and Eurich (1931) compared the reliability of multiple-choice and true-false tests and reported conflicting results. Watson and Crawford (1930) compared the formats on high school physics unit tests and reported higher reliability estimates (split-half) with multiple-choice items. Copeland and Gilliland (1943) found higher Kuder-Richardson-20 reliability estimates for a true-false test than for a multiple-choice test of child psychology when they equated the reliability coefficients for testing time using the Spearman-Brown Formula. Eurich (1931) performed two experiments comparing true-false and multiple-choice items on educational psychology. He reported that the internal consistency reliability estimate of the multiple-choice test was substantially higher in one trial and approximately equal to the true-false test reliability in the other trial. He concluded that the reliabilities of multiple-choice tests are consistently as high and usually higher than the reliabilities of true-false tests.

Burmeister and Olson (1966) sought to determine whether a test composed of college-level natural science

true-false items had the same desirable psychometric characteristics as the multiple-choice form. They concluded that true-false items could be constructed that discriminate almost as well as multiple-choice items. They also noted that true-false items tended to be somewhat less difficult than multiple-choice items. The difference in difficulty may have been due, at least partially, to the guessing effect. Grosse and Wright (1985) note that the effect of guessing is to some degree, dependent on the number of alternatives offered by the items employed. However, as the number of items increases, the probability of attaining a passing or acceptable score on any item format significantly decreases.

Ebel (1971) also studied the comparative reliability and validity of true-false and multiple-choice tests. He constructed two forms of a natural science test, each form composed of 44 true-false items and 44 multiple-choice items. Ebel notes that the mean discrimination indices tended to be higher for the multiple-choice tests. The Kuder-Richardson 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 Prophecy Formula for

a double-length test under the assumption that two true-false items can be attempted for every multiple-choice item attempted.

In the investigation of concurrent validity, Ebel correlated students' scores on the two forms under the assumption that if both item types are measuring the same psychological dimension, the correlation between them would be unity. In order to compensate for the unreliability of the tests, he applied the correction for attentuation and reported that the corrected correlations between the multiple-choice and true-false subtests on the two forms were 1.20 and .80. Ebel concluded that the results of his study support the hypotheses that: (a) true-false and multiple-choice tests are equally reliable when testing time is equated, and (b) there is no difference between the concurrent validity of multiple-choice and true-false tests.

Frisbee (1974) performed a study very similar to Ebel's but with a considerably larger sample. The Spearman-Brown Prophecy Formula was used in both investigations to adjust the reliabilities of the true-false tests in order to equate them with the multiple-choice tests on the basis of testing time rather than number of items. Ebel (1971) arbitrarily used a ratio of 2:1 (true-false items: multiple-choice items) as

mentioned above, and Frisbee used an experimentally derived ratio of 3:2 in adjusting the Kuder Richardson Reliability estimates in his study. Perhaps as a function of the ratio used in equating the tests, the results of Frisbee's study did not confirm those of Ebel's (1971) study. Rather, the reliabilities of the multiple-choice tests in Frisbee's study were consistently higher than the true-false tests.

In the interest of identifying the optimal number of alternatives for increasing reliability, Williams and Ebel (1957) compared the reliability of tests composed of items having four, three, and two alternatives via the Kuder Richardson 20 method. They concluded:

> For tests of equal working time . . . three choice vocabulary test items gave a test of equal reliability, and two choice items a test of higher reliability, in comparison with the standard four-choice item. (p. 59)

Costin (1970) reported somewhat different results. His study indicated that three choice items tend to be the most reliable. Ramos and Stern (1973) and Hogben (1975) investigated only four and five choice items and found the five alternative items superior in reliability. It should be noted that Ramos and Stern did not equate the tests for testing time, as many other researchers did, but rather, compared tests of equal number of items. More recently, Straton and Catts (1980)

conducted a study seeking to identify the optimal number of alternatives that should be provided in multiple-choice test items to maximize reliability. They estimated reliability using the analysis of variance method and concluded:

> The findings of this study lend support to the notion that tests composed of three choice items are equivalent or superior to tests of four or two choice items when test reliability is used as the basis for comparison. (1980, p. 364)

Straton and Catts (1980) suggest that for many applications, three-choice items are to be preferred. They point out that, compared to four-choice items, three-choice items are easier to write and the distractors taken as a set are more plausible. Also, students should be able to complete more items in a fixed period of time thus ensuring greater coverage of subject matter. As a result, test reliability for three-choice items should be at least as high as that achieved with four choice items.

Grier (1975) reviews the results reported surrounding the debate of the optimal number of alternatives and concludes that shorter items with fewer alternatives are frequently found to be more reliable than longer items with more alternatives such as five-choice items. He also notes that in addition to this advantage of increased reliability that there are at

least two other advantages associated with shorter items. The first is that it is easier to write one or two plausible distractors than it is to write three or four. The second is that shorter items tend to be more efficient in that students spend less testing time reading and interpreting them than with longer, more complex items. The results of Budesue and Nevo's (1985) research on item efficiency tended to support Grier's (1975) conclusions. They found a strong and consistently negative relationship between examinees' rate of performance and the number of options for the items. However, their research did not support the findings of Straton and Catts (1980) that three-choice items yield higher reliability estimates than items with four or five options.

The research cited in this section spans the period from the 1920's when objective test items first became popular to the present. While the research surrounding any of the questions discussed provides no definitive results, it does appear to support a few general conclusions. First, the research seems to support the thesis that any item format can be successfully employed to test any subject matter. Second, it seems apparent that while free-response items tend to be somewhat more difficult than multiple-choice items, there is little

evidence to indicate the existence of a verbal production factor. Finally, it appears that shorter, less complex items with fewer options produce reliability estimates that are often as good or better than longer, more complex items with more options. The results of this research review provide support for the study of alternate-choice items as they clearly fall into the category of shorter, less-complex test items with fewer alternatives.

Studies Comparing Amount of Testing Time

Most researchers comparing different forms have considered the efficiency of the forms an important variable. Efficiency is defined as the number of items to which examinees are generally able to respond in a given unit of testing time. Efficiency is important because when comparing the reliability of two or more test item forms, researchers have generally considered it appropriate to equate the tests for testing time (using the familar Spearman-Brown Prophecy Formula) and to compare the reliability of these theoretically lengthened tests.

In two studies previously cited, Charles (1926) and Ruch and Stoddard (1925) reported differing results when they compared time required by examinees to respond to

test items of various formats. Charles (1926) reports that his subjects were able to respond to 1.4 true-false items for every five choice multiple-choice item and 1.2 true-false items for every three-choice item. The corresponding ratios from the Ruch and Stoddard study are 1.6 and 1.3. Williams and Ebel (1957) stated that subjects finished faster as the nubmer of response alternatives diminished, but they did not indicate how much faster. In another study, Ebel (1971) reported that subjects typically attempted two true-false items for every one multiple-choice item attempted. However, there appears to be a general agreement in the results of other studies (Watson & Crawford, 1930; Copeland & Gilliland, 1943; Frisbee, 1974) that three true-false items can usually be attempted for every two multiple-choice items attempted.

In a more recent study, Ward (1982) compared the speed of examinees' performance on verbal aptitude items set in an open-ended (free-response) format and a multiple-choice format. He explains that all items were pretested to ensure that adequate time limits would be permitted during data collection to avoid problems associated with test speededness. On the basis of pretesting, Ward determined that 75% of the examinees were able to complete 20 multiple-choice items in 12 minutes and 20 open-ended items in 15 minutes (36 seconds and 45 seconds per item respectively).

Frisbee and Sweeney (1982) compared the relative merits of multiple-true-false (MTF) items with multiple-choice (MC) items. They explain that MTF items are a cross between multiple-choice and true-false items, consisting of a question or problem posed in the stem followed by a series of statements pertaining to it. Examinees respond true or false to each of the statements. In discussing the results, they note:

> The MFT format would appear to have several advantages over the MC format: a greater number of responses can be obtained in a given time period; the longer test is likely to be more reliable; a greater range of content can be examined because of the length; and a more valid measure should be obtained because of increased reliability. In addition, students might indicate greater preference for MTF than MC because the additional test length gives more opportunities for them to show what they have learned. (p. 29)

In the interest of identifying the relative efficiency of multiple-true-false items, Frisbee and Sweeney constructed two content-parallel test forms, each containing 50 multiple-choice items (five-choice) and 250 sets of multiple-true-false items (each composed of a stem and five true-false propositions). Students were told in advance that the tests consisted of both multiple-choice and multiple-true-false items. Since multiple-true-false items had been used previously in course exams, students were quite familiar with them. The researchers explained that the course tests were

carefully timed. After ten minutes of testing time had elapsed, students were stopped and asked to circle the number of the test item they had just attempted. Testing then resumed without further interruptions. The ratio in their study was 3.44 multiple-true-false items to one five-choice multiple-choice item.

In another study comparing the relative merits of alternate-choice items with true-false items, Ebel (1982) noted that students seemed to be able to complete about the same number of each type of item in a given time period. This study, described as preliminary, provided no empirical data.

The research reviewed in this section is not conclusive nor does it provide any definitive results or information pertaining to the efficiency of any particular test item format beyond the obvious observation that students are able to answer more short and simple test items than they are long and complex items in a given time period. It does, however, highlight the fact that researchers (1) have considered the efficiency of an item type as an important consideration and (2) have been able to gauge item efficiency with some consistency. This information has been useful to researchers in estimating the reliability of tests of theoretical length. Recent empirical

evidence of the type described in this section is lacking for alternate-choice items. Therefore, provision was made in this study to collect data on the ratio of the number of multiple-choice to alternate-choice items to which examinees are able to respond in a given time period.

Studies Considering Examinees' Preference for an Item Type

In addition to the factors mentioned above, researchers have also considered examinees' preference as a worthwhile element in the overall evaluation of an item format. It is hoped that such information may provide insight into examinees' ability to interpret and understand the question posed in the item, and to gauge students' acceptance of the item format. In one such case, Ebel (1982) conducted an informal survey of students enrolled in an introductory measurement course who had encountered both true-false and alternate-choice items. He reports that students generally expressed belief that the alternate-choice items were less ambiguous and equal in difficulty to the true-false test items used in the course. In addition, more than half of the students surveyed expressed a preference for alternate-choice items. Ward (1982) also considered examinee preference in his study comparing different

forms of verbal aptitude test items. He noted that students' perceptions that multiple-choice items were less difficult was confirmed by the data. Perhaps as a result of the perception that multiple-choice items were easier, students also expressed a preference for them.

Students' perceptions were also considered an important variable in a study of the effects of incorporating humor in test items (McMorris, Urbach, & Connor, 1985). The researchers reported that the inclusion of humorous items did not affect the students' scores, yet students favored the inclusion of humorous items and perceived them to be easier. Item statistics did not support students' perceptions of the difficulty of these items. Rosenfeld and Anderson (1985) also studied the effects of including humor in test items. Their results were similar to those of McMorris, Urbach, and Connor (1985) in that all students viewed humor positively. However, Rosenfeld and Anderson did detect a significant sex difference in perception of the items and performance on them. The college males viewed the experimental items as much more humorous than their female counterparts did, but also scored lower. The researchers speculated that the males who perceived the items as extremely funny may have been more distracted by them than the female participants were.

Frisbee and Sweeney (1982), in a study previously cited comparing multiple-true-false and multiple-choice test items, asked the examinees for their perceptions of the relative difficulty of the two forms and for their preference. Results indicated that students' perceptions closely matched empirically derived tabulations of item difficulty. Also, nearly half (47.8%) indicated a preference for multiple-true-false over multiple-choice items. About one-third (32.0%) preferred multiple-choice and the remainder of the examinees expressed no preference of one type or the other.

Campbell (1961) and Benson and Crocker (1979), and Green (1984) concentrated their investigations on students' perceptions of item difficulty. All of these studies concluded that item format and students' reading ability were significantly related to students' perceptions of item quality. In the same vein, Ebel (1980) was interested in students' perception of test item ambiguity. He suggests that, in many cases, students' complaints that a test item is ambiguous are the result of the students' own lack of knowledge or incomplete knowledge of the subject matter. Ebel refers to this condition as extrinsic ambiguity. Conversely, he explains that intrinsically ambiguous items are ambiguous due to the fact that the item is imprecisly worded or has

some other defect inherent in the item. Ebel (1980) emphasizes this distinction between intrinsic and extrinsic ambiguity in order to highlight the effect students' knowledge of the subject matter has on their perception of item quality. An ill-prepared student views many items as very difficult and/or of poor quality.

While the research cited seems to indicate that examinees' perceptions (especially of item difficulty) may be significantly influenced by reading ability and subject matter knowledge, it also indicates that, as a rule, examinees are able to judge the relative difficulty of item types fairly accurately. It seems clear that examinees' perceptions should be considered an important element in the overall evaluation of a novel item type. If an item type does not have face validity and at least nominal familiarity to and acceptance of the examinees, use of the item type may serve to disturb or disrupt the "psychological set" of the examinees as they prepare for and take examinations (Frisbee & Sweeney, 1982).

Summary

In Chapter II a general survey of research surrounding the debate of the relative merits of various test item formats was provided. The review affirms that alternate-choice items constitute a potentially important

addition to the repertoire of test items employed by educators and other test users. As such, they merit further study. In Chapter III the methodology is presented; including a description of the students who participated in the study, and the procedures employed to test the hypotheses.

CHAPTER III

DESIGN AND PROCEDURES

Introduction

This research study was designed to examine the reliability and concurrent validity of alternate-choice and multiple-choice tests of educational psychology. The difficulty and discrimination of the two item forms was compared, and students' preference for alternate-choice or multiple-choice test items was explored. In this chapter the sample is described and the instrumentation is discussed. Some examples of the items employed in the study are also provided, as well as sections outlining the methodology and the analysis. The chapter concludes with a brief summary and an introduction to Chapter IV.

Hypotheses

The hypotheses of the study were:

1. There is no difference in the difficulty indices of alternate-choice and multiple-choice tests.
2. There is no difference in the discrimination indices of alternate-choice and multiple-choice tests.
3. Examinees will attempt the same number of alternate-choice and multiple-choice items in a fixed time period.

4. There is no difference in the internal consistency reliabilities of alternate-choice and multiple-choice tests.
5. The Pearson product-moment correlation between individuals' alternate-choice and multiple-choice test scores is 1.00 when corrected for attenuation.
6. There is no difference in examinees' preference for alternate-choice and multiple-choice tests.

Subjects Participating in the Study

The subjects that participated in this study were undergraduate students at Michigan State University enrolled in the Spring Term of 1983 in a course titled Teacher Education 200, The Individual and the School. This one-term, four credit course is required of all elementary and secondary education majors at Michigan State and serves primarily as an introduction to educational psychology. On the average, the course has 100 to 200 students enrolled in five to ten sections. Course sections have common content, textbooks, tests and term papers and are taught by supervised teaching interns. In a typical term the majority of the students in the course are sophomores and juniors. About half of the students are elementary education majors and the other half are secondary education majors.

Subjects participating in the study were not randomly selected. Rather, all students enrolled in the course were included in the study. While it is generally agreed that complete randomization is necessary to guarantee the external validity of a study (Campbell & Stanley, 1963), it was not possible in the present case. Nevertheless, the subjects included in the study did appear quite representative of college students enrolled in other courses in the College of Education at Michigan State University.

Two groups of subjects were actually used in the study, since two phases of testing were required for instrument development and data collection. The first group was used to try out newly created alternate-choice test items. This group of subjects was composed of all students enrolled in Teacher Education 200 in the Winter Term of 1983. The second group used in the actual data collection was composed of all students enrolled in the course in the Spring Term of 1983. The final group of participants in the study was composed of 112 students enrolled in six sections.

Instrumentation

The multiple-choice items that were used in the study were drawn from a pool of currently existing items

used to construct unit and final examinations for the course. Items in this pool had been prepared with great care and used in various tests for the previous six terms. These items had been analyzed, revised and improved on the basis of expert judgement (judgement of the course instructors and educational psychology faculty) and item analysis. The items had also been classified according to topics and keyed to the objectives of the course. They appeared to provide adequate sampling of the curriculum. As a result of the care taken in their preparation, analysis and revision, these multiple-choice items were judged to be technically sound, having appropriate difficulty, discrimination and content validity. A few examples of these items are given below.

Bruner would place a child who makes use of visual images to organize his thoughts at which stage of cognitive development?

A. Concrete
B. enactive
*C. Iconic
D. Symbolic

Bill West, a high school social studies teacher wants his students to "become good citizens." In order to make this goal a reality he must first:

a. provide opportunties for citizenship to occur.
*b. specify how good citizenship is to be demonstrated.
c. contact good citizens in the community to serve as models.
d. identify appropriate rewards for achieving good citizenship.

Cognitivists take which of the following positions on the role of errors in learning?

a. Errors are threats to learning.
*b. All responses provide some feedback.
c. Errors are events that must be overcome.
d. Teachers should eliminate errors that they can anticipate.

Alternate-Choice Item Generation Procedure

Since a similar large pool of high-quality alternate-choice items pertaining to the course content was not available, it was necessary to create one. Alternate-choice test items to meet that need were written with the assistance of the course coordinator and course instructors. In addition to the multiple-choice items, the following resources were used in writing the alternate-choice items:

a) The course text and teachers manual.

b) The course study guide, manual and term projects.

c) Common student problems and questions.

d) Ideas from the course instructors and educational psychology staff.

e) The multiple-choice test items.

As noted in Chapter I, the alternate-choice item format as proposed by Ebel (1982) is a unique item form. There is an important difference between the conventional two-choice multiple-choice form mentioned in the

literature (Straton & Catts, 1980; and others) and the unique form pospoed by Ebel. The pool of multiple-choice items described above was useful in highlighting important concepts and ideas, and served as an aid in identifying topics for writing alternate-choice items. However, the alternate-choice items were conceived and written independently of the multiple-choice items. They were <u>not</u> fashioned simply by eliminating two or three of the least attractive distractors from four or five choice items as conventional two-choice items often are.

The alternate-choice items were written using a procedure suggested by Ebel (1980). In that procedure, he lists the first step in writing alternate-choice items as the identification of an important concept or idea which can be stated simply and definitively in a declarative sentence or proposition. This step is common to the creation of all items. With the concept statement in mind, Ebel instructs the item writer to analyze the statement for a central or critical element which could be conceived as a semantic differential and presented as polar alternatives to test the students' knowledge of the concept. According to Ebel, the response alternatives should:

1. Involve a critical element of the proposition.
2. Be distinctly different, opposite in meaning, or mutually exclusive.

3. Be parallel in meaning and structure (members of the same class of ideas).

4. Avoid inclusion of the word "not."

5. Be plausible.

6. Be definitely correct or incorrect.

7. Complete the sentence sensibly.

8. Be presented in natural or alphabetical order.

9. Present no relevant clues. (Ebel, 1980, p. 115)

For example, one application of this procedure is seen in creation of an alternate-choice item to test the students' knowledge of theories of human development. All of the development theorists studied in the course emphasize that the rate of attainment or progress through various developmental stages tends to vary from person to person. The sequence or order through which people pass through the stages, however, is fixed and does not vary. This concept was identified as an important concept which could be stated simply and definitively. The critical element of this concept was then identified as the sequence of the stages. Two alternatives which involved the critical element were conceived and the following item was written.

> The sequence at which different people pass through developmental stages is *a) fixed. b) varied.

The course coordinator and instructors who assisted with the item writing were surprised and delighted at the

ease with which they were able to generate items which, often after only minor revision, were judged to be quite acceptable items. As items were written, they were reviewed for flaws by measurement specialists. These items were also edited by the course instructors and educational psychology faculty consultants for subject matter relevance and accuracy. Based on these reviews, items were either retained, revised, or discarded. Using this method, a bank of approximately 300 alternate-choice test items was created. A few examples of these alternate-choice test items are given below.

> Piaget believed that learning is most likely to occur in the presence of cognitive
> *a) conflict. b) harmony.
>
> Most teachers agree that rote learning and drill are *a) efficient b) inefficient activities for enhancing the learning of young children.
>
> Cognitivists believe that a child who is given candy for doing his/her homework is a) more *b) less likely to learn as much over the long run as a child given no reward.

In order to ensure that all items used in the final data collection were technically sound and also to provide an opportunity to "iron out" the details for the final data collection, a pilot study was conducted in the Winter Term of 1983. While the newly created alternate-choice items had been subjectively analyzed by subject matter and measurement experts, a more complete analysis required that the items be subjected to actual

try-out and item analysis, as had the multiple-choice items used in the course.

Since each unit test and the final examination of the pilot study was to be composed of two content-parallel equivalent subtests, one subtest composed of alternate-choice items and the other composed of multiple-choice items, it was necessary to take special care in test construction. A table of test specifications containing unit objectives, cognitive level to be tested, and the number of items allocated to measure each objective was designed for each test (see Appendix A). These tables of specifications were created with the aid of the instructors and the educational psychology faculty consultants. The test specifications were used in the process of item generation to ensure that the alternate-choice and the multiple-choice items were measuring the same level of cognitive complexity and that the tests composed of the two item formats were as content parallel as possible.

Analysis of the pilot study data (see Table 1) revealed that, as a rule, the mean item difficulty, discrimination and the reliability of alternate-choice tests were only slightly lower than the multiple-choice tests. In the pilot study it was possible to try-out the newly written alternate-choice test items and on the

basis of item analysis and student reaction solicited informally, to refine them further. Foils that were not attractive were replaced, and items that were identified as either very difficult or very easy or that displayed low discrimination were marked for reevaluation. Accordingly, many of the original items were revised or rewritten. It was necessary to discard others.

Table 1

Psychometric Characteristics of the Items Employed in the Pilot Study

[a]Test	Mean Diff.	Mean Disc.	K-R 20 Rel.
Test 1 AC	64	26	.479
Test 1 MC	70	26	.509
Test 2 AC	83	24	.629
Test 2 MC	74	29	.650
Test 3 AC	81	18	.522
Test 3 MC	72	28	.552
Test 4 AC	75	24	.486
Test 4 MC	76	25	.595
Final AC	75	21	.605
Final MC	66	28	.705

[a]N of items = 30 for Tests 1-4, N=50 for the Final Exam

Research Design

In the Spring Term of 1983 the data for the study were collected. A counter-balanced design was employed

in which three unit examinations and a final examination was administered to all students enrolled in the course. Each unit examination was composed of two content-parallel subtests of thirty items; one composed of alternate-choice items and the other composed of multiple-choice items. The final examination was composed of 50 alternate-choice items and 50 multiple-choice items.

In addition to comparing the psychometric properties of alternate-choice and multiple-choice items, the study was also designed to investigate the efficiency of alternate-choice items. Specifically, the study sought to identify the ratio of the number of alternate-choice items to multiple-choice items to which students were able to respond in a given time period. Such information was necessary to compare the reliabilities of the two forms equated for testing time. In order to gain this information, a time study was conducted in the Spring Term of 1983. Students in all sections were timed and asked to mark the item they had just completed at the end of five minutes and at the end of ten minutes. The students were then informed that they would not be interrupted further and were instructed to proceed with the examination.

Since students' perceptions of the tests and preference for alternate-choice or multiple-choice items was also deemed an important consideration, a survey was conducted at the end of the term to explore the students' perceptions and preferences. Students were asked to respond frankly and to write additional comments on the form. Care was taken to assure the students of anonymity in their feedback. On the basis of remarks made by students under the comments section of the survey and verbally to the instructors of the course it was concluded that the students reported honestly and took the task seriously.

Analysis

In the interest of gaining information necessary to test the hypotheses, each of the examinations was scored and analyzed. The item analysis programs employed provided information on examinees' scores, item and test difficulty and discrimination, and estimates of internal consistency reliability (Kuder-Richardson 20) necessary to evaluate the first three hypotheses. The definitions and formulas for these indices and coefficients may be found in an introductory measurement text.

Multivariate analysis of variance (MANOVA) with alpha set at .05 was employed in the study to determine

whether the multiple-choice items differed significantly from the alternate-choice items in difficulty and discrimination. Norusis (1985) explains that MANOVA allows a researcher to test the differences between two or more groups on two or more dependent variables. In this study, item format was the independent variable and item difficulty and item discrimination were the dependent variables. Norusis (1985) also notes that the MANOVA procedure yields a main effect F for each of the dependent variables and for the interaction effects. If a significant MANOVA F is found the researcher may elect to perform follow-up univariate analyses in an attempt to determine which levels of the dependent variable are significant and contributing to the MANOVA F. Follow-up analyses were not performed in this study, since the statistical significance of the differences among the four alternate-choice examinations and among the four multiple-choice examinations in difficulty and discrimination were not of interest.

The third hypothesis pertained to the efficiency of the alternate-choice and multiple-choice items. To test that hypothesis frequency distributions indicating the number of items to which subjects responded in the allotted time were constructed. The ratio of the means of the two distributions was tested for significance using the chi square test for goodness of fit.

Shavelson (1981) explains that the purpose of the chi square test is to determine whether the observed distribution differs systematically from the theoretically expected distribution, or whether the differences may be attributable to chance. In the present case, the observed ratio of the number of alternate-choice to multiple-choice items to which students were able to respond was tested against the hypothesized ratio of 1:1, indicating no difference.

In the interest of testing the fourth hypothesis, the Kuder-Richardson 20 reliability estimates for the alternate-choice and multiple-choice test were equated using the familiar Spearman-Brown correction formula. The equating procedure was based on the information gathered in the time trials previously mentioned, and the alternate-choice tests were theoretically lengthened and equated by testing time with the multiple-choice tests. A procedure proposed by Feldt (1980) was employed to test these equated coefficients.

Feldt explains that the statistic for conducting the test of $H: p_1 = p_2$ when the coefficients are obtained form the same sample is as follows:

$$t_{N-2} = \frac{(W-1)(N-2)^{1/2}}{(4W\ (1-r^2_{x_1x_2}))^{1/2}}$$

Where: $W = \frac{1-r_1}{1-r_2}$

N = examinees

$r_{x_1x_2}$ = correlation between the tests.

The test is based on the usual assumptions associated with the two-factor random model analysis of variance and is generally considered to provide a conservative test of the hypothesis that coefficient alpha (or Kuder-Richardson 20 if the items are scored dicotomously) is the same for two tests or measurement procedures.

The procedure proposed by Feldt was used to test each of the four sets of reliability coefficients of the study. It is important to recognize that many statisticians look with disfavor on the practice of performing consecutive tests that cannot be regarded as strictly independent, because this practice potentially increases the possibility that at least one of the tests might reach significance by chance (Norusis, 1985). Accordingly, in the interest of maintaining a conservative test and minimizing Type 1 errors, the .01 confidence level was selected as the decision point for rejecting the fourth hypothesis.

Hypothesis 5 pertains to the concurrent validity of the alternate-choice and multiple-choice format. To test this hypothesis a Pearson product-moment correlation coefficient was calculated beween examinees' scores on the multiple-choice and alternate-choice subtests on each of the four examinations. 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).

$$r_{tt} = \frac{r_{xy}}{\sqrt{r_{xx}\ r_{yy}}}$$

Where: r_{tt} = true correlation between scores on x and y
r_{xy} = correlation between observed scores on x and y
r_{yy} = reliability coefficient for the y scores

Theoretically, if the two item formats are measuring the same construct, the correlation between them when corrected for attenuation will equal one. The corrected correlation coefficients were tested to determine if their values were different from unity using a procedure which Lord (1959) suggests. Lord notes that the appropriate test statistic in situations of this type is chi square (X^2) distributed with one degree of freedom. In order to provide a conservative test, alpha was preset at the .01 level of significance.

The last hypothesis (Hypothesis 6) pertains to examinee preference for tests composed of alternate-choice and multiple-choice items. In order to gain information necessary to evaluate examinees' preference, a questionnaire was constructed and administered to the examinees at a class period near the end of the term. Examinees' responses were tabulated and analyzed to identify examinee preference. The chi square

goodness of fit test previously described was used to test examinees' preference for tests composed of multiple-choice or alternate-choice items.

Summary

The 112 subjects who participated in the final testing phase of this study were described as representative of undergraduate college students enrolled in the College of Education at Michigan State University. In a counter balance design, each subject was administered three unit examinations and a final examination composed of content parallel alternate-choice and multiple-choice subtests.

Examinees' responses were analyzed and difficulty indices, discrimination indices, and reliability coefficients were calculated. MANOVA was employed to test the differences in the difficulty and discrimination indices of the alternate-choice and multiple-choice subtests. The ratio of the number of alternate-choice and multiple-choice items to which subjects were able to respond in a given unit of testing time was calculated; and the procedure proposed by Feldt (1980) was employed to test the differences in the reliability coefficients of the alternate-choice and multiple-choice tests for significance. Also, the correlation between individuals' alternate-choice and multiple-choice subtest scores was

calculated and corrected for attenuation for each testing form to provide information on the the concurrent validity of the alternate-choice test items. Finally, a questionnaire was administered and responses analyzed to evaluate examinees' preference for alternate-choice or multiple-choice items.

CHAPTER IV

RESULTS OF THE STUDY

Introduction

The results of the study are presented in Chapter IV. The chapter is divided into major sections corresponding to the research hypotheses. The first section deals with the comparisons of the difficulty and discrimination of the two item forms. Section two pertains to the findings regarding the number of multiple-choice and alternate-choice items to which examinees were able to respond in the time trials. The third section contains the results relevant to the reliabilities of the multiple-choice and alternate-choice subtests. Results that reflect on the concurrent validity of the subtests composed of the two item formats are reported in the fourth section, and results pertaining to students' perceptions of alternate-choice items are presented in the fifth. The chapter concludes with a brief summary.

Results Concerning Difficulty and Discrimination

The data were analyzed using MANOVA of the Statistical Package for the Social Sciences (Nie, Hull, Jenkins, Steinbrenner, & Brent, 1985) to determine

whether significant differences existed in the difficulty and discrimination indices of the alternate-choice and multiple-choice items. The results of the analysis are presented in Table 2. An inspection of Table 2 reveals that the tests composed of alternate-choice items tended to be easier than the multiple-choice tests. This is evidenced by the fact that the means of the item difficulty indices for the alternate-choice tests were generally higher than those of the corresponding multiple-choice tests, indicating a greater proportion of the examinees answered the alternate-choice items correctly. These differences were found to produce a significant ($F > .05$) main effect for item difficulty in the MANOVA analysis. It was, therefore, necessary to reject the first hypothesis of no difference in the difficulty of the two formats.

The results of the analysis of the discrimination of the tests were similar to those of the difficulty analyses. The data presented in Table 2 shows that the multiple-choice items tended to be somewhat more discriminating than the alternate-choice items. However, the results of the MANOVA analysis indicated that the difference in the mean discrimination indices of the tests composed of the two item forms was not significant at the .05 level. On the basis of the analysis of the

Table 2

Descriptive Statistics and Multivariate Analysis of Variance for Difficulty and Discrimination

Test	Mean Diff.	S.D.	MANOVA	P
	DIFFICULTY			
Test I				
Alternate-Choice	68	15.52	7.372	.017
Multiple-Choice	71	16.74		
Test II				
Alternate-Choice	72	14.59		
Multiple-Choice	69	15.21		
Test III				
Alternate-Choice	72	14.66		
Multiple-Choice	68	16.95		
Final Exam				
Alternate-Choice	75	17.71		
Multiple-Choice	65	22.04		
	DISCRIMINATION			
Test I				
Alternate-Choice	32	10.28	0.246	.620
Multiple-Choice	36	11.15		
Test II				
Alternate-Choice	26	12.16		
Multiple-Choice	34	13.08		
Test III				
Alternate-Choice	32	12.49		
Multiple-Choice	26	11.03		
Final Exam				
Alternate-Choice	24	13.78		
Multiple-Choice	26	12.44		

discrimination indices of the tests, the second hypothesis was not rejected.

Results Concerning Efficiency of Alternate-Choice and Multiple-Choice Items

The third hypothesis of the study stated that examinees can respond to the same number of alternate-choice and multiple-choice items in a fixed time period. In order to examine the data pertinent to this hypothesis, frequency distributions were constructed of the number of items students attempted in the time trials. The mean number of alternate-choice and multiple-choice items which students attempted and the ratio of the number of each form attempted in each trial are presented in Table 3.

The results of the time study indicated that the examinees were able to respond to a greater number of alternate-choice items than multiple choice items in the time allowed. In both Test I and Test II students were timed and instructed to mark the item on which they were currently working at five minutes and again at ten minutes. The ratio of the average number of alternate-choice to multiple-choice items attempted by the students in the two tests was then calculated. These ratios, which serve as an index of the relative rates of work by subjects on the two item forms, were quite stable

Table 3

Mean Number of Alternate-Choice and Multiple-Choice Items Attempted in the Time Trials

Test[a]	5 Minutes	10 Minutes
Test I		
Alternate-Choice	15.3	26.6
Multiple-Choice	10.4	17.4
RATIO (AC:MC)	1.47	1.53
Test II		
Alternate-Choice	14.4	26.7
Multiple-Choice	8.6	16.2
RATIO	1.67	1.65

[a]N=103 for Test I, n=108 for Test II.

across trials, ranging from 1.47 (one multiple-choice item to 1.47 alternate-choice items) to 1.67, with an average of 1.58. The proportion of 1:1.58 was tested using the chi square goodness of fit test to determine the probability of this proportion if the population proportion is actually 1:1 as hypothesized. The analysis revealed a chi square of 10.24 (1, N=98) which was found to be significant at the .01 level. The conclusion drawn from these results was that, in general, students attempted approximately three alternate-choice items for every two multiple-choice items they attempted.

Results Concerning the Reliability of the Alternate-Choice and Multiple-Choice Tests

The fourth hypothesis of the study proposed no differences in the reliability of tests composed of either alternate-choice or multiple-choice items. The study of the reliability of the two forms involved calculating the Kuder-Richardson 20 estimate for each of the unit examinations and the final examination. These reliability coefficients are reported in Table 4. The alternate-choice test reliability estimates were adjusted with the Spearman-Brown formula to estimate the reliabilities of tests 1.58 times as long as the original tests. The adjusted reliabilities also appear in Table 4.

Table 4

Significance Tests of the Reliability of the Alternate-Choice and Multiple-Choice Tests

Test	KR-20	[a]Corrected	t_w	P
Test I				
alternate-choice	.711	.795	.655	.514
multiple-choice	.778			
Test II				
alternate-choice	.547	.666	1.130	.261
multiple-choice	.601			
Test III				
alternate-choice	.588	.693	.201	.842
multiple-choice	.683			
Final Examination				
alternate-choice	.682	.772	1.578	.117
multiple-choice	.713			

[a]Reliabilities of the alternate-choice tests were adjusted by the Spearman-Brown formula to estimate the reliability of tests 1.58 times as long.

The difference between the corrected alternate-choice reliability coefficients and the multiple-choice test reliability coefficients were tested using the technique (t_w) proposed by Feldt (1980) to determine whether any of them were statistically significant. The t_w statistic is provided in Table 4 with the corresponding probability level. None of the differences in the reliability coefficients were found to be significant at either the .01 or .05 level. Therefore, the fourth hypothesis of no difference in the reliability coefficients of the two item forms was not rejected.

Results Concerning Concurrent Validity

The fifth hypothesis of interest pertained to the concurrent validity of the alternate-choice and multiple-choice item formats. In order to test this hypothesis, examinees' responses on each examination were scored to derive a separate alternate-choice and multiple-choice subtest score. A Pearson product moment correlation coefficient was calculated between students' scores on the multiple-choice (x) and the alternate-choice (y) tests. These correlation coefficients are presented in Table 5. The coefficients were then corrected for attenuation and the chi square

technique proposed by Lord (1957) was used to test whether the dis-attenuated correlation coefficients differed significantly from unity. The results of that analysis are also presented in Table 5.

Table 5

Significance Tests of the Correlation Coefficients for Multiple-Choice and Alternate-Choice Scores on the Tests

Test Form	N	r_{xy}	r^a_{tt}	X^2_1
Test I	103	.689	.877	8.222*
Test II	108	.583	.921	1.825
Test III	112	.530	.770	19.425*
Final Exam	106	.666	.898	17.162*

r^a_{tt} designates r_{xy} corrected for attenuation

*p < .01

The results revealed that three of the four disattenuated coefficients were significantly different from unity at the .01 level. The conclusion drawn from the analysis was that the corrected correlations between individuals' multiple-choice and alternate-choice scores was not perfect (equal to one), and that the concurrent validity of the two forms was not equal.

Results Concerning Students' Preference for Alternate-Choice and Multiple-Choice Items

Prior to the administration of the final examination and the conclusion of the course, students were asked to complete a brief questionnaire designed to measure their perceptions of the course tests. Whether the students would prefer alternate-choice items over multiple-choice items as Ebel (1982) suggested was of special interest. The results of that questionnaire are presented in Appendix B.

The results of the questionnaire indicated that the students tended to view the course testing as well as the alternate-choice items employed quite positively. For example, over 70% of the students indicated that the tests provided a strong motivation to learn the principles taught, and that the alternate-choice items tested important concepts in the curriculum. It is interesting to note that although slightly more than half of the students perceived alternate-choice items to be more difficult and ambiguous than multiple-choice items, these perceptions were not verified by the item statistics. Question nine, however, indicated that a majority (74%) of the students responding thought that future exams should be composed of both multiple-choice and alternate-choice items. Only 25% of the students

favored either the exclusive use of multiple-choice (21%) or alternate-choice (5.0%) items on future exams.

The question most pertinent to the sixth hypothesis of the study was item number six which posed the statement, "I would prefer taking a test composed of alternate-choice items to a test composed of multiple-choice items." Only 34% of the students responded affirmatively to this statement, while 66% of them disagreed. This proportion of 66:34 was tested using the chi square goodness of fit test. The analysis revealed a chi square of 12.47 (1, N=98) which was found to be significant at the .01 level, indicating that a significant number (66%) of the students responding did not prefer taking a test composed of alternate-choice items. Because these results did not provide evidence of a preference for alternate-choice items over multiple-choice items, the fifth hypothesis was not rejected.

Summary

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

1. While the multiple-choice form was significantly more difficult that the

alternate-choice form, the difference in the discrimination of the two forms was not found to be significant.

2. Students attempted approximately three alternate-choice items for every pair of multiple-choice items attempted.
3. The Kuder-Richardson 20 reliability estimates for the multiple-choice tests were not significantly different from those of the corrected for length alternate-choice test reliability estimates.
4. The correlations, corrected for attenuation, between the multiple-choice and alternate-choice test scores were significantly different from unity for three of the four tests, indicating a difference in the concurrent validity of the two forms.
5. Students did not express a preference for alternate-choice items over multiple-choice items. The majority, however, indicated that both alternate-choice and multiple-choice items should be included on future tests.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The purpose of this study was to compare the reliability of multiple-choice and alternate-choice tests and to explore the concurrent validity of alternate-choice tests that were written to measure understandings of concepts and relationships in an introductory Educational Psychology course. The major questions of the study that were formulated as research hypotheses were:

1. Are alternate-choice and multiple-choice achievement tests that were designed to measure the same objectives equally difficult and discriminating?

2. What is the ratio of the number of alternate-choice to multiple-choice items to which examinees are able to respond in a given time period?

3. Are alternate-choice and multiple-choice achievement tests that were designed to measure the same objectives equally reliable?

4. Will the corrected for attenuation correlations between examinees' scores on the multiple-choice and alternate-choice tests equal unity?

5. Will examinees exhibit a strong preference for the newer alternate-choice format over the familiar multiple-choice item format?

Ebel (1980) proposed a novel item format termed alternate-choice items. He suggested that this new item type would compare quite favorably with other conventional test item formats with regard to difficulty, discrimination, reliability and validity. He also suggested that students tended to prefer them over more conventional item formats. A review of the literature revealed a large number of studies comparing the reliability, validity, and other psychometric properties of various test item forms. However, since the alternate-choice item format was new, very little research had been conducted on it. While this unique item form proposed by Ebel appeared to show potential as an important addition to the repertoire of test item formats that teachers and other item writers have at their disposal, little empirical research existed to substantiate Ebel's claim.

In this study a group of approximately 112 undergraduate students enrolled in an introductory educational psychology course at Michigan State University each responded to three unit tests and a final examination composed of content-parallel alternate-choice

and multiple-choice subtests. Students' responses were scored and analyzed to identify the difficulty and discrimination indices of the items. Kuder-Richardson 20 reliability estimates for the alternate-choice and the multiple-choice items were also calculated. Students were timed on two occasions to identify the number of alternate-choice and multiple-choice items to which they were able to respond in a given time period. The correlation between individuals' scores was calculated and corrected for attenuation for each unit examination and the final examination. In addition, a questionnaire was developed and administered to the subjects to explore their perceptions of alternate-choice items and their preference for either alternate-choice or multiple-choice items.

Statistical tests were performed to determined if the difficulty, discrimination and reliability of the two item types were significantly different and to determine if the value of the corrected correlation coefficients departed significantly from unity. Statistical tests were also performed to determine the probability of the observed results of the time study and of students' responses on the questionnaire.

Conclusions

The conclusions associated with the research hypotheses listed in Chapter I were:

1. The multiple-choice items were significantly more difficult than the alternate-choice items.

2. The multiple-choice items were not significantly more discriminating than the alternate-choice items.

3. Students were able to respond to approximately three alternate-choice item to every two multiple-choice items they attempted.

4. There was not a significant difference between the reliabilities of the alternate-choice and the multiple-choice tests.

5. The corrected for attenuation correlations between the multiple-choice and alternate-choice test scores were significantly different from unity for three of the four examinations.

6. Students did not express a strong preference for either item format used in the study, but generally viewed both formats positively.

Discussion

The results of the analysis of difficulty and discrimination of the alternate-choice and

multiple-choice items revealed significant differences in the difficulty but not the discrimination of the two forms. These results may be interpreted as evidence that alternate-choice items discriminate between knowledgeable and less-knowledgeable examinees approximately as well as multiple-choice items, even when they are less difficult. Because it is probably easier for item writers to produce test items that are less difficult than it is to produce items that are of medium to high difficulty, this may constitute an important advantage, especially since alternate-choice items may be comparatively easier to write than multiple-choice items in the first place. Statistical significance aside, the practical importance of these results is that they support Ebel's (1982) claim that a person who is content knowledgeable and a fairly skilled item writer can produce alternate-choice items which compare quite favorably psychometrically with other item formats.

As expected, students' "rate of work" varied with item form. Students were able to respond to substantially more alternate-choice items in a given unit of testing time than multiple-choice items. Since it has previously been demonstrated (Ebel, 1971; Frisbee, 1974) that students are usually able to respond to approximately three true-false items to every two

multiple-choice items, it was expected that students' rate of work on alternate-choice items would approximate that of true-false items. It was, therefore, not surprising to find that students answered an average of 1.58 alternate-choice items to every multiple-choice item they attempted for a ratio of about 3:2. While the primary purpose of the identification of this ratio of alternate-choice to multiple choice items was to equate the alternate-choice tests with the multiple-choice tests for testing time, it does have additional practical importance. A practical consideration of this finding is that a longer test may be administered in a given period of testing time if the items are in alternate-choice form as opposed to multiple-choice form. This would be especially important if the examiner is concerned about the adequacy with which the sample of items which comprise the test represent the universe of content. Also, a longer test can provide for a more thorough sampling of the universe and will usually constitute a more reliable measure.

The results of the reliability analysis generally followed those of the difficulty and discrimination analyses. The reliability estimates of the alternate-choice tests were slightly lower than those of the corresponding multiple-choice tests. However, when

the reliabilities of the alternate-choice tests were adjusted by the Spearman-Brown formula to equate them with the multiple-choice tests for testing time, in each case they were slightly higher than those of the corresponding multiple-choice test. None of these differences were statistically significant, yet the reliabilities of all of the tests were quite respectable for classroom tests of this type. The results of the reliability analysis provides further support for Ebel's claims regarding the usefulness of alternate-choice items. It is important to recognize that these results are somewhat confounded because test reliability is related to the difficulty and discrimination of the items involved.

The results of the validity study are more difficult to interpret. Three of the four disattenuated concurrent validity coefficients were found to be significantly different from one, indicating a difference in the concurrent validity of the two forms.

The explanation for this difference is not readily apparent. While a concerted effort was made to create alternate-choice tests that were content-parallel and that tested the same level in the cognitive hierarchy, it is possible that these results are due to systematic differences in the items of the subtests. It may also be

true that the format of an item makes it more conducive to a particular type of question than to another, or some concepts may lend themselves more readily to being tested with one item format than another. For example, alternate-choice items seem particularly well suited for posing comparative statements or propositions to the examinee. The alternate-choice items written for and used in the study may be generally directed toward concepts that are most easily adapted to the item form. These concepts may have substantive differences or vary in cognitive complexity from those most readily adapted to the multiple-choice format.

Another possible explanation for the difference in concurrent validity may be in the cognitive task of responding to the different item forms. The cognitive task of responding to the alternate-choice items may be somewhat different from that required of the multiple-choice items. Since alternate-choice items are new, the examinees had not had nearly the exposure to them as they have had to multiple-choice items, and it may be that the cognitive tasks of reading the alternate-choice item, identifying the question, evaluating the alternatives, and selecting a response may be somewhat different from the series of cognitive tasks involved in responding to a multiple-choice item.

From a practical standpoint, the correlations found between students' scores on the alternate-choice and multiple-choice formats were quite high. Students who scored high on one form also tended to score high on the subtest composed of the other item format and vice versa. Further, both item formats appeared to do a good job of testing students' achievement and to be quite valid.

The results of the questionnaire administered to the students indicated that they did not have a strong preference for alternate-choice items as Ebel (1982) predicted. Nevertheless, the examinees viewed the tests in general and the alternate-choice tests in particular, quite positively. Students indicated that the alternate-choice items were challenging and did a good job of testing their knowledge of the course content, and that most of the alternate-choice items were relevant and tested important points in the curriculum. It is interesting to note that more than one-half (60%) of the students perceived that the alternate-choice items were more difficult than the multiple-choice items. These results were not verified by the item statistics. As mentioned previously, the alternate-choice items were clearly not more difficult than the multiple-choice items. In fact, the analysis of the indices of

difficulty for the alternate-choice items showed the examinees responded correctly to them more often than they did to the multiple-choice items.

In the same vein, it is interesting to observe that 59% of those responding to the questionnaire perceived the alternate-choice items to be ambiguous. Yet 45% of those responding held the contrary view that it was easier to understand and interpret the question in alternate-choice items than in multiple-choice items, and 32% indicated that the multiple-choice items were more ambiguous than the alternate-choice items.

As with students' perceptions of item difficulty, students' percepitons of item ambiguity was not confirmed by the item statistics. Items are considered ambiguous when there is not a single correct answer on which experts would agree. Ambiguous items tend to confuse examinees and, therefore, to produce lower or negative indices of discrimination. Item statistics for ambiguous items usually show the higher achieving examinees unable to select a single response as correct and often divided between two or more of the options. The item analysis statistics for the tests used in this study did not show such a pattern. While the item discrimination tended to be lower than ideal, this was more attributable to the relatively low difficulty of the items.

A possible explanation for the students' perception of ambiguity may lie in the students' understanding of the concept of ambiguity. Ebel (1980) suggested that many students do not possess a clear understanding of the term and often regard all items to which they do not know the correct answer as ambiguous. Their perceptions of ambiguity, therefore, may be at least partially attributable to low achievement. Incomplete knowledge which does not equip the student to detect subtle differences or make fine distinctions may lead the student to view many items as difficult and ambiguous. In the present case, the most likely explanation of the results may be that the lower achieving students tended to view many items, both alternate-choice and multiple-choice as difficult and ambiguous. The fact that some examinees tended to view the alternate-choice items as more difficult and ambiguous than the multiple-choice items may also be at least partially attributable to the novelty of the form.

This study was designed to compare selected psychometric indices between alternate-choice and multiple-choice items. The purpose of the study was to provide evidence as to the usefulness of the alternate-choice format in testing achievement. To that end, the results of the statistical analyses are probably

not as important as the practical importance of the study. Wesman (1971) points out that the generalizability of most research studies of item form effectiveness and item comparison studies are quite limited. He notes that this is principally due to the fact that situational variables such as the content of the items and the skill of the item writer cannot be controlled. For example, in one study, well written true-false items on natural science may appear quite superior to lower quality sentence completion items. In another study, well conceived completion type items written to test students' ability to formulate hypotheses in the social sciences may be much more reliable than a set of multiple-choice items written to test the same subject.

Even though the generalizability of the present study may be quite limited, it does constitute a demonstration of the usefulness of alternate-choice items. Also, the statistical significance tests are probably not as important as the fact that a relatively large pool of alternate-choice items were written and, after revision, were found to perform quite well and to yield quite acceptable psychometric indices of quality. It is also of importance to note that the examinees viewed the alternate-choice items positively and believed

that the items presented to them in this novel format did a good job of testing how much they knew.

Limitations of the Study

The results of this study should be interpreted with caution, for certain limitations existed in the study. Major limitations of the study fall into two major categories: 1) those associated with sampling, and 2) uncontrolled situational variables. As described in Chapter III, the participants consisted of all students (n=112) enrolled in an introductory educational psychology course at Michigan State University. The fact that the participants were not randomly selected from a defined population limits the generalizability of the study. The generalizability is further limited by the fact that the participants were a fairly homogeneous group of students majoring in Education. It is uncertain to what extent these students are representative of college students of other major fields of study or at other universities, or to what extent these results can be generalized to other age groups.

The study is further limited by uncontrolled situational variables. As previously noted, Wesman (1971) cites the fact that researchers are unable to control situational variables as the principle cause of

the failure of item effectiveness or item comparison studies to significantly advance our knowledge or art of item writing. Wesman identifies the skill of the item writer and the subject matter as two important situational variables that commonly limit studies of item comparison or effectiveness. Unfortunately, both of these variables remain uncontrolled in this study further limiting the generalizability of the results.

Suggestions for Further Research

The following suggestions are offered for further investigation into the comparative effectiveness of the alternate-choice item format.

1. The generalizability of the present study was limited by the effects of situational variables such as the subject matter tested and the skill of the item writers. It would be appropriate for future research to study the effectiveness of alternate-choice items in other situations, testing different subject matters, and written by other item writers.
2. Examinees' perception of item difficulty and conception of item ambiguity merits further investigation, possibly in relation to the examinees' knowledge of the subject matter or level of achievement.

3. The number of alternate-choice items to which examinees are able to respond in a given time period could be investigated using different subject matter with item difficulty as a control variable.
4. The results of this study suggested differences in the concurrent validity of the item formats. Research might be designed to investigate these differences further.
5. The cognitive task of responding to various item formats may be quite different and may constitute a significant source of item difficulty. Additional research into the cognitive aspects of responding to test items could be fruitful.

APPENDICES

APPENDIX A

TABLES OF SPECIFICATIONS

Table of Specifications

INTRODUCTION AND UNIT I: COGNITION

	Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1.	Teacher as a decision maker	1				1
2.	Theory into practice, e.g., Physical Development		1			1
3.	Piaget's theory of learning	1	1	1		3
4.	Piaget's four Developmental periods	3	1	1		5
5.	Relation between language and thought	1		1		2
6.	Bruner's perspective on learning: Discovery	1	1	1		3.5
7.	Ausubel: Didactic Instruction	1	1	1		3.5
8.	Klausmeier's Model of Concept Attainment	1	1			2
9.	Attention to Stimuli	3	1	1		5
10.	Acquisition and Retention	3	1	1		5

INTRODUCTION AND UNIT I: COGNITION (cont.)

	Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
11.	Improving Performance	4	1	1		6
12.	Retrieval and Transfer	2	1	1		4
13.	Memory: Short-Term and Long-Term	1	1	1		3
14.	Social-Economic Status	1	1			2
15.	I.Q. Testing and its uses	1	1	1		3
16.	Sex Roles and Sexism	1	1	1		3
17.	Cognitive styles	2	1	1		4
18.	Differentiation	1	1			2
19.	Guilford's Model of the Intellect	1				1
20.	Creativity and its enhancement	1	1			2
	TOTAL	30	17	13	1	51

Table of Specifications

UNIT II: BEHAVIORAL THEORY AND SOCIAL-EMOTIONAL DEVELOPMENT

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1. Conditions of Learning	2	1			3
2. Simplistic Theories of Learning	1				1
3. Early Behavioral Research	1				1
4. Quantitative Behavioralism	1				1
5. Skinner: Theory of Reinforcement	2	1	1		4
6. Bandura: Social Learning Theory	2	1	1		4
7. Modeling: Characteristics	2	1	1		4
8. Reinforcement: Characteristics	2	3	1		6
9. Using reinforcement	2	3	1		6
10. Reducing Undesirable	4	1	1	1	7
11. Social-Emotional Development	1	1			2

Table of Specifications

UNIT II: BEHAVIORAL THEORY AND SOCIAL-EMOTIONAL DEVELOPMENT

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1. Conditions of Learning	2	1			3
2. Simplistic Theories of Learning	1				1
3. Early Behavioral Research	1				1
4. Quantitative Behavioralism	1				1
5. Skinner: Theory of Reinforcement	2	1	1		4
6. Bandura: Social Learning Theory	2	1	1		4
7. Modeling: Characteristics	2	1	1		4
8. Reinforcement: Characteristics	2	3	1		6
9. Using reinforcement	2	3	1		6
10. Reducing Undesirable	4	1	1	1	7
11. Social-Emotional Development	1	1			2

Table of Specifications

UNIT III: INSTRUCTION

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1. Educational Goals and Learning Objectives	3	1			4
2. Taxonomies of Educational Objectives	2	1	1		4
3. Types of Learning Outcomes	4				4
4. Information Processing	1		1		2
5. Designing Instruction and adapting needs	1	1			2
6. Self-fulfilling prophecy	1		1		2
7. Effects of Inappropriate Teacher Expectations	3	2	1		6
8. Characteristics of Appropriate Teacher Expectations	2	1	1	1	5
9. Student Self-Concept	2	1	1		4
10. Teaching for personal growth	1	1			2

UNIT III: INSTRUCTION (cont.)

	Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
11.	Organizing instruction, classroom variables	1	1	1		3
12.	Successful Classrooms	1				1
13.	Characteristics of Teaching Effectiveness	3	1		1	5
14.	Classroom Climate and Effective Teaching	2	1	1		4
15.	Professional Teacher Development	1	2			3
16.	Sources for Teacher Improvement	1	1	1		3
17.	Feedback and its uses	1	1	1		3
18.	Creating, Maintaining and Restoring Appropriate classroom behaviors	1				1
	TOTAL	31	15	9	2	57

Table of Specifications

UNIT IV: MOTIVATION AND MANAGEMENT

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1. Definition of Behavior Motivation	1				1
2. Need Theory: Murray and Maslow	2	1	1		4
3. Cognitive Theory: Weiner	1				1
4. Intrinsic Theory: Hunt	1	1			2
5. Achievement Motivation	1	1	1		3
6. Attribution Theory	1	1			2
7. Task involvement and achievement	1	1			2
8. Motivational Tasks	3	2	2	1	8
9. Cooperation and Competition	1	1			2
10. Creating desirable student behavior	2	1			3
11. Maintaining Students on Task	1	1			2

UNIT IV: MOTIVATION AND MANAGEMENT (cont.)

	Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
12.	Restoring students to desirable behavior	2	1			3
13.	Ideal Teacher Attitudes and Behavior	3	2	1		6
14.	Characteristics of a Favorable Teaching Environment	1	1			2
15.	Leadership styles	2	1			3
16.	Management Techniques	2	1	1		4
17.	Classroom physical characteristics and tasks	1				1
18.	Public Law 94-142	1	1	1		3
19.	Mainstreaming and special problems	2	1			3
20.	Individual Educational Programs	1				1
	TOTAL	30	17	8	1	56

Table of Specifications

UNIT V: EVALUATION

(Note: Unit V topics will be texted on the Final Examination along with a survey of the other Units.)

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
1. Errors of Measurement	1				1
2. Test Reliability	1	1			2
3. Test Validity	1		1		2
4. Norm referenced tests	1	1			2
5. Criterion referenced	1				1
6. Interpreting test results	1		1		2
7. Descriptive Statistics	3	1			4
8. Distribution Statistics	2	1			3
9. Normal Curve	3	1	1		5
10. Test Length and Content Coverage	1	1			2
11. Instructional Objectives and Test Construction	1	1			2

UNIT V: EVALUATION (cont.)

Unit Content	Knowledge (Recall)	Comprehension (Translation, Interpretation, Extrapolation)	Application (Abstract to Concrete and vice versa)	Analysis (Elements, Relations, Organizing Principles)	TOTAL
12. Essay tests: construction and scoring	1	1			2
13. Objective test items	1	2			3
14. Norm vs. criterion measures	1	1			2
15. Rating scores	1	1			2
16. Function of grades	1				1
17. Types of grade schedules	1	1	1		3
18. Grades and Educational Philosophy	1		1		2
TOTAL	23	13	5		41

APPENDIX B

QUESTIONNAIRE SUMMARY

Test and Item Questionnaire Summary

		Percent Agree/ Strongly Agree	Percent Disagree/ Strongly Disagree
1.	The tests given in this course provide a strong motivation for me to study.	71%	29%
2.	Alternate-choice items tend to be more difficult than multiple-choice items.	60%	40%
3.	Most of the alternate-choice items test important points in the curriculum.	77%	23%
4.	Many of the alternate-choice items are ambiguous.	59%	41%
5.	Alternate choice items challenging and do a good job testing how much I know.	59%	41%
6.	I would prefer taking a test composed of alternate-choice items to a test composed of multiple choice items.	34%	66%
7.	It is usually easier to interpret and understand the question posed in alternate-choice items than in multiple-choice items.	45%	55%

	Percent Agree/ Strongly Agree	Percent Disagree/ Strongly Disagree
8. Multiple-choice items tend to be more ambiguous than alternate-choice items.	32%	68%
9. In the future, unit exams should be composed of:		
a. all alternate-choice items	5%	
b. all multiple-choice items	21%	
c. part alternate-choice items and part multiple-choice items.	74%	

BIBLIOGRAPHY

BIBLIOGRAPHY

Andrews, D. M. & Bird, C. (1938). Comparison of two new-type questions: Recall and recognition. Journal of Educational Psychology, 29, 175-193.

Bensen, J. & Crocker, L. (1979). The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 39, 225-233.

Budescu, D. U. & Nevo, B. (1985). Optimal number of options: An investigation on the assumption of proportionality. Journal of Educational Measurement, 22 (3), 183-196.

Burmeister, M. A. & Olson, L. A. (1966). Comparison of item statistics for items in multiple-choice and in alternative-response forms. Science Education, 12, 467-470.

Campbell, A. C. (1961). Some determinants of the difficulty of nonverbal classification items. Journal of Educational Psychology, 21, 899-913.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental design for research. Chicago: Rand McNally.

Charles, J. W. (1926). A comparison of five types of objective tests in elementary psychology. Unpublished doctoral disseration, University of Iowa, Iowa City.

Choppin, B. H. & Purvis, A. C. (1969). Comparison of open-ended and multiple-choice items dealing with literary understanding. Research in the Teaching of English, 3, 15-24.

Cook, D. L. (1955). An investigation of three aspects of free-response and choice-type tests at the college level. Dissertation Abstracts International, 30, (9-10A) 4310.

Copeland, J. S. & Gilliland, A. R. (1943). Comparison of the validity and reliability of three types of objective examinations. Journal of Educational Psychology, 34, 242-246.

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

Ebel, R. L. (1971). 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.

Ebel, R. L. (1975). Can teacher write good true-false items? Journal of Educational Measurement, 12, 31-35.

Ebel, R. L. (1980). Some advantages of alternate-choice items. Unpublished manuscript. College of Education, Michigan State University, East Lansing, MI.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. Journal of Educational Measurement, 19 (4), 267-277.

Eurich, A. C. (1931). Four types of examinations compared. Journal of Educational Psychology, 22, 268-278.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45 (1), 99-105.

Frisbee, D. A. (1974). The effects of item format on reliability and validity: A study of multiple-choice and true-false tests. Educational and Psychological Measurement, 34, 885-892.

Frisbee, D. A. & Sweeney, D. D. (1982). The relative merits of multiple- true-false achievement tests. Journal of Educational Measurement, 19, 29-35.

Ghiselli, E. C. (1964). Theory of psychological measurement. New York: McGraw-Hill.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. Educational and Psychological Measurement, 44, 551-562.

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

Grosse, M. W. & Wright, B. D. (1985). Validity and reliability of true-false tests. Educational and Psychological Measurement, 45, 1-6.

Hays, W. L. (1973). Statistics for the social sciences. New York: Holt, Rinehart, and Winston.

Heim, A. W. & Watts, K. P. (1967). Experiment on multiple-choice versus open-ended answering in a vocabulary test. British Journal of Educational Psychology, 37, 339-346.

Hogben, D. (1975). The reliability, discrimination, and difficulty of work knowledge tests employing multiple-choice items containing three, four, or five alternatives. Australian Journal of Education, 17, 63-68.

Hughes, R. & Trimble, W. (1965). The use of complex alternatives in multiple-choice items. Educational and psychological Measurement, 24, 117-126.

Huck, S. W. (1978). Test performance under conditions of known item difficulty. Journal of Educational Measurement, 15, 117-126.

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 22 (3), 207-220.

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

Maihoff, N. A. & Mehrens, W. A. (1985). A comparison of alternate-choice and true-false items forms used in classroom examinations. Paper presented at the National Council on Measurement in education Annual Meeting, Chicago.

McMorris, R. F., Urbach, S. L., & Connor, M. C. (1985). Effect of incorporating humour in test items. Journal of Educational Measurement, 22 (4), 127-134.

Mehrens, W. A. & Lehmann, I. J. (1984). Measurement and evaluation in education and psychology. New York: Holt, Rinehart & Winston.

Mendelson, M. A., Hardin, J. H., & Canady, S. D. (1980). The effects of format on the difficulty of multiple-completion test items. Harvard Educational Review, 50, 73-80.

Millman, J. (1978). Determinants of item difficulty: A preliminary investigation. CSE Reports No. 114 Ed 163 071.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Brent, D. (1975). Statistical package for the social sciences. New York: McGraw-Hill.

Norusis, M. J. (1985). SPSS-X: Advanced statistics guide. New York: McGraw-Hill.

Oosterhof, A. C. & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. Applied Psychological Measurements, 8 (3), 287-294.

Plake, B. S. & Huntley, R. M. (1984). Can grammatical clues result in invalid test items? Educational and Psychological Measurement, 44, 687-692.

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

Rocklin, T. & Thompson, J. M. (1985). Interactive effects of test anxiety, test difficulty, and feedback. Journal of Educational Psychology, 77 (3), 368-372.

Rosenfeld, P. & Anderson, D. (1985). The effects of humorous multiple-choice alternatives on test performance. Journal of Instructional Psychology, 12 (1), 3-5.

Rowley, G. L. (1974). Which examinees are most favored by the use of multiple choice tests? Journal of Educational Measurement, 11 (1), 15-21.

Ruch, G. M. & Stoddard, G. D. (1925). The comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 16, 89-103.

Shavelson, R. J. (1981). Statistical reasoning for the behavioral sciences. Boston: Allyn and Bacon, Inc.

Straton, R. G. & Catts, R. M. (1980). A comparison of two, three, and four choice item tests given on total number of choices. Educational and Psychological Measurement, 40, 357-364.

Toops, H. A. (1921). Trade tests in education. Teachers College Contribution to Education. New York: Teachers College, Columbia University, No. 115.

Traub, R. E. & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice test. Applied Psychological Measurements, 1, 355-369.

Ward, W. W. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurements, 6, 1-11.

Watson, D. R. & Crawford, C. C. (1930). Four type of tests. High School Teacher, 7, 282-283.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.) Educational measurement (pp. 81-129). Washington, DC: American Council on Education.

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

Wilson, V. L. (1982). Maximizing reliability in multiple-choice questions. Educational and Psychological Measurement, 42, 69-72.

Winne, P. H. & Belfry, J. M. (1982). Interpretive problems when correcting for attenuation. Journal of Educational Measurement, 19 (2), 125-133.

General References

Bloom, B. S. (1956). Taxonomy of educational objectives; handbook I: Cognitive domain. New York: Longmans, Green and Co.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Ebel, R. L. (1979). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Hopkins, K. D. & Stanley, J. C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lee, W. (1975). Experimental design and analysis. San Francisco: Freeman and Company.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addisson Wesley.

Thorndike, R. L. (1971). Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Winer, B. J. (1971). Statistical principles in experiemental design. New York: McGraw-Hill.