To the Graduate Council:

I am submitting herewith a thesis written by Linda Burros Smalley entitled "The Effects of Item Format on Reliability and Student Performance in a Classroom Test." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Educational Psychology.

Schuyler W. Huck, Major Professor

We have read this thesis
and recommend its acceptance:

Donald V. Whitmore

Kathleen A. Lawler

Accepted for the Council:

L. Evans Peek
Vice Chancellor
Graduate Studies and Research

THE EFFECTS OF ITEM FORMAT ON RELIABILITY AND STUDENT PERFORMANCE IN A CLASSROOM TEST

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Linda Burros Smalley

December 1981

ACKNOWLEDGMENTS

The author wishes to express her appreciation to Dr. Schuyler Huck for his support and guidance throughout this study. The author is also grateful to Dr. Kathleen Lawler and Dr. Donald Wheeler for their helpful comments and suggestions.

A special acknowledgment is extended to Sheldon Clark, Ann Nowlin, and Dr. Mark Hector for serving as test reviewers during the stages of test construction. Appreciation is also expressed to John Jozwiak, Dr. Ben Layne, and Dr. John Lounsbury for allowing the tests to be administered to their students.

ABSTRACT

This study was conducted to examine the effects of item format upon the variables of reliability and student performance while holding constant the factor of item content. The four item formats compared were completion, multiple choice, true-false, and a focused true-false version which utilized underlining of a word or phrase in an attempt to reduce true-false ambiguity.

The test items were developed with the objective of assessing knowledge of research design and internal validity issues. Each of the 28 items was expressed in four formats with stem wording held constant. Four test forms were created, each of which contained four seven-item sections of differing formats. The content order was identical throughout the four forms, but the order of presentation of the item format was varied with a 4 x 4 balanced Latin square layout. The tests were randomly distributed to a total of 80 graduate students from four different classes during a regular classroom period.

Statistical analysis consisted of performing a Latin square analysis of variance for each of the dependent variables, reliability and performance. Significant Fs at the .05 level of significance for both variables were followed by mean analysis for each variable with Tukey's procedure for pairwise comparisons.

The reliability results showed that the reliability of the completion format was higher than that of the true-false format, while the multiple choice, focused true-false, and true-false reliabilities were not found to be significantly different from each other.

With regard to the performance variable, completion items were found to be more difficult than those items expressed in true-false or focused true-false format. The study did not find significant differences in student performance between the true-false, focused true-false, and multiple choice formats. The underlining technique did not result in any apparent improvement of the true-false items since the focused true-false items were not found to be significantly different from the true-false items for either the reliability or the performance variable.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	Need for the Study	1
	Purpose of the Study	3
	Research Questions	3
	Terminology	3
II.	REVIEW OF THE LITERATURE	5
	General Comments about Item Types	5
	Studies Comparing Item Formats	8
III.	METHOD	13
	Subjects	13
	Item Construction	14
	Test Formation	16
	Procedure	16
IV.	RESULTS	19
	Reliability Results	19
	Performance Results	23
V.	DISCUSSION	27
	Conclusions	27
	Discussion	28
	Limitations of the Study	31
	Suggestions for Future Research	32
LIST OF REFERENCES		33
APPENDIX		38
VITA		40

CHAPTER I

INTRODUCTION

The introductory chapter of this thesis serves to address the need for the present study and to state the study's purpose and the research questions which were asked. A final section of this chapter defines the terminology specific to this research project.

A survey of relevant literature follows and focuses upon both general comments in testing literature about item formats as well as specific evidence which has been gathered from research comparing item formats. The method by which the experiment was conducted is delineated followed by data analysis and technical details. Finally, the results are discussed accompanied by conclusions and recommendations for future research.

Need for the Study

Testing is a vital part of any academic environment. In academic settings, teacher-made tests may be used to direct students' study, measure achievement, and provide student feedback. Through good testing techniques, the teacher can enhance the quantity and quality of learning (Hopkins and Antes, 1979).

Item-writing rules and general comments concerning item formats and suggestions for improving test quality are included in many measurement texts and testing-related literature, but the authors frequently comment upon the lack of empirical evidence to support such rules or generalizations (Wesman, 1971; Ebel, 1972; Frisbie, 1973; O'Neill and Johnson,

1976). Specifically, many of these authors and researchers have noted a lack of research which gathers data from teacher-made tests. The fact that such a void exists currently was suggested by Michael Beck, a researcher who recently spoke to a National Institute of Education conference and urged for the study of tests devised by teachers (Harris, editor, 1980). It is interesting to note that a similar recommendation occurred in an article by Andrews and Bird (1938) in which they pointed out the need for facts gathered directly from college examinations rather than general information tests. Wesman (1971, p. 83) lamented the lack of significant research on item-writing problems and concluded:

> Improvements in teacher-made tests are within the ability of the average teacher to achieve; and improved tests can make a significant contribution not only to the field of testing, but through more effective testing, to the learning experience of the examined student as well.

One of the main decisions for teachers who construct their own tests is the choice of which item format to use. Tinkleman (1971) addressed this issue and stated that the test planner must use the item type which is particularly appropriate to the specific purposes of the test. However, this advice is rather vague and is difficult to follow, since, as Frisbie (1973, p. 297) stated, "empirical evidence to suggest that one item type is more advantageous than the other is lacking." Similarly, O'Neill and Johnston (1976) emphasized the necessity for further research so that teachers can knowledgeably select the type of test item which will produce the best educational effects for their own course material, students, and academic objectives.

Purpose of the Study

This study focused upon a classroom test comprised of four types of test items. The purpose was to examine the effects of item format upon the variables of reliability and student performance while holding constant the factors of item content and wording. The four item formats used were completion, multiple choice, true-false, and a true-false variation. The true-false variation utilized underlining of a word or phrase in an attempt to reduce ambiguity by focusing the attention of the examinee on the important part of the statement (Thorndike and Hagen, 1969).

Research Questions

The research questions were as follows:

1. In a classroom test, does the use of different item formats result in significantly different reliabilities?
2. Is there a significant difference in students' performance on different item types if content and wording is held constant?
3. Can true-false items be improved through an underlining technique designed to reduce ambiguity?

Terminology

The four forms of test items which were used in the present study were completion, multiple choice, true-false, and a focused true-false item. The completion format and its closely related variations are referred to in the literature as short-answer, recall, free-response, supply, or fill-in-the-blank items. For this study, the term completion was used to denote a declarative statement containing a blank. The

examinee was directed to fill in the blank with a word or phrase to make the sentence a true statement. The multiple choice form, also known as a selection or recognition type item, consisted of an incomplete declarative statement, the stem, followed by three to five response alternatives (one correct response and two to four distractors). The true-false item, which some authors call alternate-choice, was a statement which the student was directed to judge as true or false. The focused true-false version was identical to the true-false item except that underlining was added to the sentence in the same position as the blank occurred in the completion item of identical wording.

Test reliability, for the purposes of this study, was a measure of internal consistency and was calculated by the Kuder-Richardson Formula 20. Performance was defined as the numerical score (number of items correct) obtained by a student on each section of the test.

CHAPTER II

REVIEW OF THE LITERATURE

There is considerable diversity of opinion in the testing literature regarding the relative desirability of one item format over another. Some authors suggest that the particular subject matter and/or course objectives dictate which type of item should be used. Other authors contend that a question could be converted from one format to another without a serious change in what is being measured or in the precision of the measurement. The literature survey which follows first reviews many of the general comments about item types to illustrate the many ways in which different authors view a particular item format. The second portion of this chapter presents a survey of research which has studied the effects of item formats upon the variables of interest for the present study — reliability and performance.

General Comments about Item Types

The completion item is viewed by some authors as an objective type of item (Stanley, 1964; Green, 1975; Ahmann and Glock, 1975) while others (Hopkins and Antes, 1979; Hills, 1976) group the completion format with the short essay and other free response items and thus view it as a subjective type item. Noll, Scannell, and Craig (1979) claim that the completion item has advantages of both essay and objective test items. Among the alleged advantages of completion items are that they (1) reduce the possibility of getting a correct answer by guessing, (2) are useful for measuring factual knowledge, (3) are easy to construct,

(4) involve the student as an active participant, (5) are good for computation problems, (6) permit wide sampling of objectives, (7) are not difficult to score, and (8) are a more valid indication of students' knowledge because the student must produce rather than merely identify the correct answer. Various authors and texts listing the disadvantages of completion items include that such items (1) are more difficult to score because some teacher judgment is required, (2) tend to measure only recall or isolated facts, (3) are difficult to construct so that the student knows exactly what is expected of him, (4) are more time-consuming to score, and (5) are often difficult to construct so that there is only one correct response. It is interesting to note the contradictions in the preceding lists: that which one author construes as a weakness of completion items may be viewed by yet another writer as an asset.

The multiple choice item is viewed favorably by most authors. The adjectives used to describe this item format include versatile, flexible, apparently most popular of objective item types, most valuable, and generally applicable. Gay (1980) states that multiple choice items are generally preferred because scoring is much more efficient and free from ambiguity. Green (1975), Hopkins and Antes (1979), and Stanley (1964) agree that multiple choice items can be used to measure a wide range of cognitive levels from knowledge to comprehension, application, and judgment. Other cited advantages include that multiple choice items (1) are less affected by guessing, (2) can be used for diagnostic purposes by analysis of distractor selection, and (3) can be used with a wide variety of situations and content. However, multiple choice items

are time consuming to construct and probably require more teacher skill in test construction. Additionally, since multiple choice items require more reading time on the part of the examinee, fewer questions can be asked and fewer objectives sampled than with other objective item formats.

An examination of what authors have to say about true-false items reveals diversity of opinion and some direct controversy. Wesman (1971), Ahmann and Glock (1975), and Storey (1966) agree that the advantages of true-false items are outweighed by their limitations. Wesman (1971) comments that the true-false item has been recognized as one of the least desirable of the objective item formats. The chief disadvantages cited are that true-false items (1) are ambiguous, (2) test trivial pieces of information, (3) are very susceptible to guessing, (4) cannot be used to measure complex understandings, (5) are not suited to educational diagnosis, (6) are not easy to construct, and (7) lead to unreliable measurement. Robert Ebel, a prolific writer and defender of true-false items, is joined by Hills (1976), Green (1975), and Hopkins and Antes (1979) in declaring that many of these criticisms are unfounded. These authors contend that true-false items (1) may be written to measure complex understanding and interpretation, (2) allow wide sampling of content and objectives, (3) are completely objective, (4) are not as susceptible to guessing as critics claim, and (5) if carefully written, can yield reliable and valid measures of educational achievement.

The fourth type of item which was included in this study was a variation of the true-false item in which a word or phrase in the statement was underlined. This focused true-false item, based upon a suggestion from Thorndike and Hagen (1969), was an attempt to reduce ambiguity

by focusing the attention of the examinee on the critical part of the statement thus eliminating the problem of non-critical words being open to interpretation. Use of underlining, according to Thorndike and Hagen, permits the use of more complex statements.

Studies Comparing Item Formats

The majority of generalizations, recommendations, and criticisms cited in the previous section have not been either substantiated or disproved by empirical evidence. The number of recent studies comparing item formats is not great, nor do the existing studies lead to any definitive conclusions.

A proliferation of studies comparing item formats occurred in the late 1920s. At that time, objective-type tests were considered "new" and were the subject of considerable investigation. In 1932, Kinney and Eurich published a comprehensive summary of many of those early studies in which they compiled information from 33 studies conducted between 1921 and 1931. Based upon the evidence they reviewed, the authors concluded that recall and multiple choice tests ranked highest in validity as compared with true-false and essay tests. Additionally, recall and multiple choice tests appeared superior in reliability; and objective tests, in general, were found to be at least as reliable as essay tests. True-false tests ranked high on comprehensiveness of sampling and pedagogic value. Recall tests were found to be the most difficult for students; however, of the four studies cited which compared difficulty, two found multiple choice easiest while the other two studies found true-false items to be the easiest.

The Kinney and Eurich article is certainly the most comprehensive article in the literature in that the authors compared a wide range of characteristics for several different item types. Since that time, however, there have been relatively few investigations comparing item formats. A survey of comparative studies which deal with concerns relevant to the present study follows.

One of the variables of interest for the present investigation was that of reliability. According to Ebel (1972, p. 379) reliability is often "the most significant measure of quality of a classroom test." An early study by Andrew and Bird (1938) compared the reliability of completion items with the reliability of multiple choice items on a college test. While some studies have attempted to compare different item formats which measure similar content or objectives, the study by Andrew and Bird maintained identical phrasing across the two types of items. Under actual examination conditions, the completion tests were shown to have higher reliability.

Burmester and Olson (1966) converted multiple choice items into true-false items for a natural science college examination. They reported a high reliability (.86) for the true-false form of the test (although there was no statistical comparison with the reliability of the multiple choice items). These two investigators concluded that true-false items "may be written to have the desirable characteristics usually associated with well-written multiple choice items" (Burmester and Olson, 1966, p. 467).

Several other studies have compared the reliabilities of true-false versus multiple choice items. When two tests were equated in terms of

time necessary for taking the test, Copeland and Gilliland (1943) found a multiple choice test to be slightly less reliable than a true-false test based upon the same subject matter. Recent studies by Green (1979) and Irvin, Halpern, and Landman (1980) found little, if any, difference in the reliabilities of multiple choice versus true-false tests. The latter study has limited generalizability, however, since the items were used orally in a test given to retarded students. Additionally, only 88% of the multiple choice items were directly analogous in content to the true-false test items.

A study comparing multiple choice versus true-false reliabilities by Ebel (1971) yielded contradictory results. In a natural science achievement test administered to college students, Ebel compared multiple choice with true-false items in two forms of the test. In one case, the true-false items had a slightly higher reliability. In the second case, the true-false reliability was lower than that of the multiple choice items. Frisbie (1971), a student of Ebel, conducted a similar study on a much larger scale for his doctoral research. Employing high school students as subjects and using social studies and natural science achievement test items, Frisbie created eight forms of a test. One half of each test contained multiple choice questions, and the other half of each test contained true-false items of a different content. Statistical tests showed that the reliabilities of the multiple choice tests were significantly greater than those of the true-false tests. A later investigation by Frisbie (1974) using empirically lengthened true-false tests replicated his earlier findings. Similarly, Oosterhof and Glasnapp (1974) found true-false items to be less reliable than corresponding four-option multiple choice items in a college classroom test.

Another variable of interest for the present study was that of performance. It would seem that a comparison of performance across different item formats would necessitate that the items be of identical content. While most of the cited studies utilized items of similar content or questions based upon the same objectives, many did not maintain identical wording and thus may have subtly altered item content.

Studies comparing student performance on completion versus multiple choice test items appear in agreement that completion items are more difficult than multiple choice items (Andrew and Bird, 1938; Cook, 1955; Heim and Watts, 1967; and O'Neill and Johnston, 1976). However, Cook (1955) noted that the responses to pairs of items of the two types based upon the same question were highly correlated. Additionally, Sax and Collett (1968) suggested that making close discriminations among novel distractors on a multiple choice item requiring application of knowledge may be more difficult than the corresponding recall item, whereas on simple material the multiple choice item may be easier.

One study was found which compared performance on multiple choice versus true-false items. Oosterhof and Glasnapp (1974) found that, when a correction for guessing was applied, multiple choice items were easier for students.

In summary, the research comparing the effects of item format is sparse, not very recent, often contradictory and inconclusive. The present study sought to improve upon and add to the empirical base provided by the cited studies in the following ways: (1) to include in the study all three of the objective item formats, (2) to additionally include a true-false variation with underlining designed to reduce

true-false ambiguity, (3) to hold constant the factor of item content by using identical stem wording for all formats, and (4) to utilize in the study a teacher-made test rather than a standardized instrument and thus to be relevant to classroom test construction concerns.

CHAPTER III

METHOD

The present study utilized an experimental design in which the independent variable, item format, was manipulated in order to examine the effects upon the dependent variables, reliability and performance. A balanced 4 x 4 Latin square layout, with its inherent counterbalancing and controls for carryover effects, was used for the formation of the four test forms. The counterbalancing and holding constant the factor of item content combined with extensive use of randomization (both of subjects to conditions and in the test formation) were employed to minimize the effects of extraneous variables and thus increase the internal validity of the experiment. The use of subjects of both sexes from three different colleges and at two different institutions was an attempt to increase the external validity of the study.

Subjects

The test was administered to four classes of students in three different colleges at two universities. All 80 subjects were graduate students who had been exposed to research design material.

The specific subject information is as follows: 25 subjects (5 male, 20 female) from Educational Psychology 5220, Interpreting Published Articles: Research Design at The University of Tennessee; 12 subjects (11 male, 1 female) from Psychology 5530, Issues in Applied Psychological Measurement at The University of Tennessee; 15 subjects (4 male, 11 female) from Nursing 5210, Applied Nursing Research at The University

of Tennessee; 28 subjects (6 male, 22 female) from Foundations of Education 790, Methods of Research in Education at Georgia State University.

Item Construction

The test was created with the objective of assessing knowledge of research design and internal validity issues. The majority of the items were drawn from a pool of previously used items in a research design course. The professor of the course served as a consultant during the stages of item development.

A multiple choice or a completion stem was chosen from the item pool and converted into the remaining item formats using identical wording. There were a total of 28 items each expressed in four formats resulting in a grade total of 112 items. Ultimately, four test forms were created, each of which contained four seven-item sections of differing formats.

The multiple choice version of a test item consisted of a stem followed by three to five response alternatives. The decision to vary the number of response options was based upon recommendations from several textbook authors. In each multiple choice item, the response alternatives were always arranged in logical order (numerically, alphabetically, etc.) to avoid any systematic bias in the position of the correct response.

The completion item consisted of the multiple choice stem followed by a blank. The examinee was instructed to fill in the blank with a word or phrase. The directions further indicated that the examinee should supply only one answer for each item.

The true-false format was formed by combining the stem with one of the multiple choice response alternatives. Within each seven-item section of the test, the decision to have an item be true or false was determined randomly with only the stipulation that there by four false items and three true items in each true-false section of the test. Including more false than true items is a generally accepted test construction principle for increasing reliability and decreasing the potential effects of an acquiescent response set. The multiple choice stem was combined with the correct response to form the true items. For a false item, a team of judges selected the distractor which would combine with the stem to make the most plausible false statement.

The focused true-false versions were identical to the true-false format except that underlining was added in the same position as the blank occurred in the completion item of identical wording. An example of each of the four item formats follows.

Completion item: When significant differences are attributable to subjects dropping out of the study rather than to the experimental treatment, this threat to internal validity is called _____.

Multiple choice item: When significant differences are attributable to subjects dropping out of the study rather than to the experimental treatment, this threat to internal validity is called

A) history.
B) maturation.
C) mortality.
D) selection.

True-false item: When significant differences are attributable to subjects dropping out of the study rather than to the experimental treatment, this threat to internal validity is called selection.

Focused true-false item: When significant differences are attributable to subjects dropping out of the study rather than to the experimental treatment, this threat to internal validity is called <u>selection</u>.

A general recommendation for improving tests and detecting errors and ambiguities is that teachers have their tests reviewed by colleagues. In adherence with this recommendation, the test items were subjected to independent review by three individuals. The items (in all formats) were reviewed by two graduate students who had taken research design courses as well as a course in test construction. In addition, the multiple choice items were reviewed by a professor who teaches a research design course. After the review process, several items were revised, and the final tests were assembled with order and grouping of items being determined randomly.

Test Formation

The design used for the formation of the four test forms was a randomly selected 4 x 4 balanced Latin square. This information is presented in Figure 1. The content order was identical throughout the four tests, but the order of presentation of each item type was varied with the balanced Latin square layout. The tests were formed in this way for several reasons. A balanced Latin square not only avoids an order of presentation bias, but it also controls for carryover effects since each treatment is preceded and followed by every other treatment only once. This repeated-measures design also insured that every student was tested over identical content and that each student received an equal number of each item type.

Procedure

The test was administered to the four groups of subjects during a regular classroom period. For one of the groups the test served as a

Test Form	Item Number 1-7	8-14	15-21	22-28
1	Multiple choice	True-false	Focused true-false	Completion
2	True-false	Completion	Multiple choice	Focused true-false
3	Completion	Focused true-false	True-false	Multiple choice
4	Focused true-false	Multiple choice	Completion	True-false

Figure 1

Test Formation

regular midterm examination. The other groups of subjects were asked to take the test as an assessment of their knowledge of internal validity and research design issues. The four instructors who administered the test felt that everyone had ample time to complete the examination.

Within each group of subjects, the four forms of the tests were randomly distributed to the students. Such random distribution permits the assumption that students who received the questions in one format were comparable in knowledge and skill to those who had the questions in other formats.

CHAPTER IV

RESULTS

The analysis of the reliability and performance data was conducted by using a Latin square analysis of variance with an alpha level of .05. Following significant Fs for both variables for the factor of item format, Tukey's procedure for pairwise comparisons was utilized for testing the reliability and performance means for the item formats, again with an alpha level of .05.

Reliability Results

For each section of the test, reliabilities were estimated through use of the Kuder-Richardson Formula 20:

$$r = \frac{k}{k-1}\left(1 - \frac{\Sigma pq}{s^2}\right)$$

where k is the number of items, p is the proportion of correct responses per item, q is the proportion of incorrect responses per item, and s^2 is the variance of the scores for that section of the test. Within each section, the obtained reliability was based upon an n = 20. In three of the sections, the variance term was smaller than the Σpq term, resulting in a negative number being obtained for r. Since the theoretical lower limit of r is zero, these values were entered as zeros. Further discussion of these results occurs in Chapter V.

When comparing the reliability of true-false with multiple choice or completion tests, some provision must be made either empirically or

via the Spearman-Brown Prophecy Formula for equating the testing time for the item formats. This is necessary since, for a given period of time, more true-false items can be answered than can multiple choice or completion. Ebel (1971) arbitrarily used the factor of two (true-false: multiple choice) for adjusting the reliability of the true-false tests. Other studies have attempted to determine this ratio. The ratio of true-false to multiple choice items answered has been reported as 1.43 (Copeland and Gilliland, 1943), 1.5 (Frisbie, 1971, 1974), 1.73 (Oosterhof and Glasnapp, 1974), and 2.4 (Green, 1979). Since most of these reported values (with the exception of Green's value whose study was not very comparable to the others cited or to the present study) cluster around 1.5, a value which was replicated by Frisbie, the 1.5 value was chosen for use in this investigation as the factor by which the reliability of the true-false items would be adjusted.

Figure 2 then presents the reliability values for each section of the test with adjusted values being shown for the true-false and focused true-false sections. The adjusted values were obtained through use of the Spearman-Brown Prophecy Formula with m = 1.5:

$$r_a = \frac{mr_o}{1 + (m-1)r_o}$$

where r_o is the original reliability, m is the factor by which the test is to be lengthened, and r_a is the adjusted reliability value.

To analyze the reliability values presented in Figure 2, the data were first transformed with a Fisher's r to z transformation. An analysis of variance for a single Latin square was then performed. The results of the analysis are presented in Table 1.

Test Form	Item Number 1-7	8-14	15-21	22-28
1	Multiple choice r = .42	True-false r = .00	Focused true-false r = .40	Completion r = .61
2	True-false r = .30	Completion r = .57	Multiple choice r = .56	Focused true-false r = .22
3	Completion r = .58	Focused true-false r = .59	True-false r = .58	Multiple choice r = .59
4	Focused true-false r = .00	Multiple choice r = .25	Completion r = .71	True-false r = .00

True-false and focused true-false sections show reliability values which have been adjusted by a factor of 1.5 with the Spearman-Brown Prophecy Formula.

Figure 2

Reliability Values for Each Section of the Test

Table 1

Analysis of Variance Summary Table for Reliability Data

Source of Variation	Sum of Squares	df	Mean Square	F
Item format	.54	3	.18	6.68*
Rows	.31	3	.10	3.85
Columns	.22	3	.07	2.69
Residual error	.16	6	.03	
Total	1.24	15		

*Significant at .05.

A significant F for the treatment, item format, led to a comparison of the mean reliabilities for the four item formats by use of Tukey's procedure. The ordered means (transformed back from z to r values) for the reliabilities were as follows:

Completion	Multiple Choice	Focused True-False	True-False
.62	.46	.32	.24

The results of the test indicated that the reliability of the completion format was significantly different from the reliability of the true-false format. No other significant differences were found.

Performance Results

A value for performance for each of the 80 students was determined by simply totaling the number of correct responses for each section of the test. The mean values for performance for each section based upon an n = 20, along with the variance, are presented in Figure 3.

A Latin square analysis of variance for the replicated case was conducted for the performance data. The summary table for this analysis is presented in Table 2.

Preliminary testing of the Latin square residual term showed a significant F. This result indicates that interactions may be present in one or more of the main effects. To test the main effects (columns and item format) using the Error (b) term cound have resulted in positively biasing the tests. The author, therefore, followed Myers' (1979) recommendation and tested the main effects against the Latin square residual term.

Item Number

Test Form	1-7	8-14	15-21	22-28
1	Multiple choice $\bar{X}$ = 5.00 s^2 = 1.50	True-false $\bar{X}$ = 4.45 s^2 = 1.25	Focused true-false $\bar{X}$ = 4.90 s^2 = 1.89	Completion $\bar{X}$ = 2.55 s^2 = 2.45
2	True-false $\bar{X}$ = 4.95 s^2 = 1.55	Completion $\bar{X}$ = 3.15 s^2 = 2.73	Multiple choice $\bar{X}$ = 3.90 s^2 = 2.59	Focused true-false $\bar{X}$ = 5.15 s^2 = 1.43
3	Completion $\bar{X}$ = 3.80 s^2 = 2.76	Focused true-false $\bar{X}$ = 4.25 s^2 = 2.49	True-false $\bar{X}$ = 4.80 s^2 = 2.36	Multiple choice $\bar{X}$ = 3.75 s^2 = 2.59
4	Focused true-false $\bar{X}$ = 4.50 s^2 = 1.05	Multiple choice $\bar{X}$ = 4.35 s^2 = 1.63	Completion $\bar{X}$ = 2.50 s^2 = 2.85	True-false $\bar{X}$ = 5.05 s^2 = 1.15

Figure 3

Mean Performance and Variance for Each Section of the Test

Table 2

Analysis of Variance Summary Table for Performance Data

Source of Variation	Sum of Squares	df	Mean Square	F
Between Ss				
Test Forms	1.63	3	.54	
Between subjects: Error (a)	330.49	76	4.34	
Within Ss				
Item format	165.38	3	55.13	8.65*
Columns	15.18	3	5.06	.79
Latin square residual error	38.22	6	6.37	4.62**
Subjects X columns: Error (b)	314.46	228	1.38	
Total	865.37	319		

*Significant at .05 when tested against the Latin square residual error.

**A significant F here indicates that interactions may be present in one or more of the main effects. Subsequent tests of main effects were tested against the Latin square residual term rather than the Error (b) term based upon a recommendation by Myers (1979).

Follow-up pairwise comparisons with Tukey's test were then conducted upon the performance means for the four item formats:

True-False	Focused True-False	Multiple Choice	Completion
4.81	4.70	4.25	3.00

The results of the post hoc testing indicated that the completion mean was significantly lower than the true-false and focused true-false means. There was insufficient evidence to show that multiple choice, true-false, and focused true-false differed significantly from each other.

CHAPTER V

DISCUSSION

Conclusions

The conclusions for this study with regard to the specific research questions which were asked were as follows:

1. The use of different item formats in a classroom test did result in different reliabilities. The Kuder-Richardson Formula 20 reliability coefficient for the completion format was shown to be higher than that of the true-false format. There was insufficient evidence to show differences between the reliabilities of the multiple choice, focused true-false, and true-false formats.

2. Even when students were tested over the same content with stem wording held constant, different item formats did, in some cases, result in differing levels of performance. Completion items were shown to be more difficult than those same items expressed in true-false or focused true-false format, but were not shown to be more difficult than the multiple choice items. There was insufficient evidence to indicate differences in students' performance between true-false, focused true-false, and multiple choice formats.

3. Based upon the results of the present study, it seems that the underlining technique did not improve the true-false items. Performance on the focused true-false items was not very different from student performance on the true-false items. While the focused true-false items had a slightly higher reliability than the true-false items, the difference was not great enough to be of any consequence.

Discussion

Examination of the reliability means for the four item formats shows completion to have the highest reliability, followed by multiple choice, focused true-false, with true-false having the lowest reliability. The ordering of these means is in agreement with the general findings of most of the studies cited in the literature review. The only two formats which were shown to be significantly different in this study were completion and true-false. The fact that the present study failed to find a significant difference between multiple choice and true-false reliabilities agrees with studies by Burmester and Olson (1966), Green (1979), Irvin, Halpern, and Landman (1980), and some of Ebel's research.

Another aspect of the reliability results which merits discussion is the fact that three of the sections of the test had estimated reliabilities less than zero. Ebel (1972a, p. 424) clearly states that "one of the shortcomings of the reliability coefficient is that its magnitude is not solely dependent upon the quality of the test." Among the factors that may influence the estimated reliability of a test are (1) length of the test, (2) variance of test scores, (3) abilities of students being tested, (4) test item hetereogeneity, and (5) restriction of the effective range of the test scores.

The present study attempted to control for the factor of student ability by randomly distibuting the test forms to the students. Additionally, since the item content was held constant across the four forms of the test and for the four item formats as well, the factor of item hetereogeneity on a section of a test should not be a possible

explanation for obtaining the negative reliabilities on three sections of the test.

Examination of the numbers involved in the calculation of these reliabilities indicates that the negative reliabilities occurred when the variance term was smaller than the Σpq term. A very small variance, then, was associated with three sections of the test: true-false on Test Form 1, focused true-false on Test Form 4, and true-false on Test Form 4. The scores associated with these three sections were simply not very spread out, resulting in a small variance being obtained.

However, it is interesting to note that this situation only occurred in the true-false or the variation of the true-false format. It seems plausible that two factors worked together to produce this situation: (1) the extremely short length of the test and (2) the restriction of the effective score range. A study by Brandenburg and Whitney (1972) comments that, in general, tests with scoring methods which have the largest effective score range also have the highest internal consistency coefficients. Similar conclusions (linking restricted effective score ranges with lower reliability) occur in articles by Ebel (1969) and Maberly (1967).

The effective score range is defined as the maximum score minus the expected chance score. For this study, on any seven-item section of the test, the artificially short length of the test restricted how much scores could vary on any of the four formats. It is important to realize that all of the reliabilities reported in Figure 2 are small in an absolute sense because they are based upon test sections containing only seven items. However, the true-false and focused true-false sections

appeared to be the most severely penalized due to the combination of the short length of each section in addition to their particularly restricted effective score ranges as compared with the multiple choice and completion formats.

For example, a test composed of two-alternative true-false items would have an effective score range of one-half of the number of items – for this study, about 3.5 points. The number of response alternatives for the multiple choice items varied from item to item, but a reasonable estimate of the multiple choice effective score range is probably about 5 points. For the completion format, if we assume the chance score to be 0, the effective score range would be 7 points.

Maberly (1967) further elaborates that full satisfaction of assumptions for application of internal consistency formulas is rarely met but that the assumptions are fairly robust to violations. In the case of narrowly restricted score ranges, however, one or more of the basic assumptions may be violated to such an extent that the estimation of internal consistency is extremely difficult, if not impossible.

The performance results basically agreed with most authors in the general ordering of the means, with completion being most difficult and true-false being the easiest. However, upon testing the differences between means for significance, there was insufficient evidence to show that completion differed from multiple choice or that true-false,focused true-false, and multiple choice differed significantly from each other. The main performance comparison of interest for this study, true-false versus focused true-false, led to the conclusion that underlining did not make the items either significantly more difficult or significantly easier.

The variation of the true-false format included in this study utilized underlining in an effort to alleviate ambiguity, one of the frequently mentioned criticisms of true-false items. Examination of the reliability and performance results indicates that the underlining technique apparently did not improve the true-false items. The focused true-false items were slightly more reliable than the true-false items but not significantly so. There was no evidence to indicate that the focused true-false items were less ambiguous. In fact, the students had slightly higher scores on the regular true-false items, but again, the difference was so slight as to be of no significance.

Limitations of the Study

The present study utilized classes of subjects who participated on a voluntary basis, with the exception of one class of 25 students who took the test as a regular midterm examination. The factor of having the test be a "real" exam was thus diluted, and there is no way to assess the degree of seriousness or motivation with which the other 55 subjects approached the test. In addition, the particular subject matter, the fact that the sexes were unbalanced, and that all the subjects were graduate students prohibit much generalization to other subject matter or to groups of different age or sex composition. Finally, a larger sample size would have been beneficial for this study.

With regard to the test construction, the number of items in each section was so small as to present difficulties in the estimation of some of the reliability coefficients. Sections consisting of more items would have allowed for more realistic estimates of reliability and would have resulted in a less severely restricted effective score range for

the true-false and focused true-false formats. Holding stem wording constant combined with extensive use of randomization during the test formation may have resulted in (1) some questions being forced into an item which would have been better worded differently or (2) an item being randomly assigned to be a false statement when it would have made a better true statement (or vice versa). Another possible limitation was that there was no control for the number of options on the multiple choice items. Randomization resulted in some multiple choice sections containing items with an average of almost five options per item while another section contained an average of less than four options per item. For a real test, it is recommended to vary the number of options per multiple choice item; but perhaps, for the purposes of a study, it would have been wiser to hold the number of options per multiple choice item constant.

Suggestions for Future Research

Future studies comparing item formats might utilize other age groups and/or subject matter. The use of longer tests, especially in the true-false format is recommended. Empirically lengthening the true-false test by a factor of 1.5 while holding content constant could be accomplished by converting one multiple choice item into more than one true-false item. However, instead of more studies comparing item formats in an attempt to argue for the general superiority of one item format over another, perhaps future research should focus on which item formats are appropriate or particularly inappropriate for expressing which types of content, subject matter, or course objectives.

LIST OF REFERENCES

LIST OF REFERENCES

Ahmann, J. Stanley, and Marvin D. Glock. Evaluating Pupil Growth. Fifth Edition. Boston: Allyn and Bacon, Inc., 1975.

Aiken, Lewis R. Psychological Testing and Assessment. Second Edition. Boston: Allyn and Bacon, Inc., 1976.

Andrew, Dorothy M. and Charles Bird. A comparison of two new-type questions: recall and recognition. Journal of Educational Psychology, 1938, 29, 175-193.

Brandenburg, Dale C. and Douglas R. Whitney. Matched pair true-false scoring: effect on reliability and validity. Journal of Educational Measurement, 1972, 9, 297-302.

Burmester, Mary A. and Leroy A. Olson. Comparison of item statistics for items in multiple-choice and in alternative-response form. Science Education, 1966, 50, 467-470.

Cook, Desmond L. An investigation of three aspects of free response and choice type tests at the college level. Dissertation Abstracts, 1955, 15, 1351.

Copeland, John S. and A. R. Gilliland. A comparison of the validity and reliability of three types of objective examinations. Journal of Educational Psychology, 1943, 34, 242-246.

Cureton, Edward E. The definition and estimation of test reliability. Educational and Psychological Measurement, 1958, 18, 715-738.

Ebel, Robert L. Blind guessing on achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.

Ebel, Robert L. Can teachers write good true-false test items? Journal of Educational Measurement, 1975, 12, 31-35.

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

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

Ebel, Robert L. The case for true-false test items. School Review, 1970, 78, 373-389.

Ebel, Robert L. 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, 1971.

Ebel, Robert L. The relation of item discrimination to test reliability. Journal of Educational Measurement, 1967, 4, 125-128.

Ebel, Robert L. Why is a longer test usually a more reliable test? Educational and Psychological Measurement, 1972b, 32, 249-253.

Edwards, Allen L. Experimental Design in Psychological Research. Fourth Edition. New York: Holt, Rinehart and Winston, Inc., 1972.

Frisbie, David A. Comparative reliabilities and validities of true-false and multiple choice tests. Ph.D. thesis, Michigan State University, 1972.

Frisbie, David A. Multiple choice versus true-false: a comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, David A. The effect of item format on reliability and validity: a study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892.

Gay, Lorraine R. The comparative effects of multiple-choice tests versus short-answer tests on retention. Journal of Educational Measurement, 1980, 17, 45-50.

Green, John A. Teacher-Made Tests. Second Edition. New York: Harper and Row, Publishers, 1975.

Green, Kathy. Multiple choice and true-false: reliability and validity compared. Journal of Experimental Education, 1979, 48, 42-44.

Harris, Emily C. (Ed.). Report on Education Research, June 11, 1980, 12, 1-2.

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

Henrysson, Sten. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational Measurement. Second Edition. Washington, D.C.: American Council on Education, 1971, 130-159.

Hills, John R. Measurement and Evaluation in the Classroom. Columbus: Charles E. Merrill Publishing Company, 1976.

Hopkins, Charles D. and Richard L. Antes. Classroom Testing: Construction. Itasca: F. E. Peacock Publishers, Inc., 1979.

Huck, Schuyler, Sheldon Clark, and Pei Yang. The figure-ground problem in true-false tests: and some possible solutions. Paper presented at the Seventeenth Annual Southeastern Invitational Conference on Measurement in Education, Greensboro, 1978.

Irvin, Larry K., Andrew S. Halpern, and Janet T. Landman. Assessment of retarded student achievement with standardized true/false and multiple-choice tests. Journal of Educational Measurement, 1980, 17, 52-58.

Kinney, L. B., and A. C. Eurich. A summary of investigations comparing different types of tests. School and Society, 1932, 36, 540-544.

Kirk, Roger E. Experimental Design: Procedures for the Behavioral Sciences. Belmont: Brooks/Cole Publishing Company, 1968.

Kuder, G. Frederic and Marion W. Richardson. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Maberly, Norman C. Characteristics of internal consistency estimates within restricted score ranges. Journal of Educational Measurement, 1967, 4, 15-19.

Myers, Jerome L. Fundamentals of Experimental Design. Third Edition. Boston: Allyn and Bacon, Inc., 1979.

Neter, John, and William Wasserman. Applied Linear Statistical Models. Homewood: Richard D. Irwin, Inc., 1974.

Noll, Victor H., Dale P. Scannell, and Robert C. Craig. Introduction to Educational Measurement. Fourth Edition. Boston: Houghton Mifflin Company, 1979.

O'Neill, George W., and James M. Johnston. An analysis of test item types as a determinant of student academic performance and study behavior. Journal of Personalized Instruction, 1976, 1, 123-127.

Oosterhof, Albert C., and Douglas R. Glasnapp. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. Journal of Experimental Education, 1974, 42, 62-64.

Sax, Gilbert, and LeVerne S. Collet. An empirical comparison of the effects of recall and multiple-choice tests on student achievement. Journal of Educational Measurement, 1968, 5, 169-173.

Stanley, Julian C. Measurement in Today's Schools. Fourth Edition. Englewood Cliffs: Prentice-Hall, Inc., 1964.

Stanley, Julian C. Reliability. In R. L. Thorndike (Ed.), Educational Measurement. Second Edition. Washington, D.C.: American Council on Education, 1971, 356-442.

Storey, Arthur G. A review of evidence on the case against the true-false item. Journal of Educational Measurement, 1966, 59, 282-285.

Terwilliger, James S., and Kaustubh Lele. Some relationships among internal consistency, reproducibility, and homogeneity. Journal of Educational Measurement, 1979, 16, 101-108.

Thorndike, Robert L., and Elizabeth Hagen. Measurement and Evaluation in Psychology and Education. Third Edition. New York: John Wiley and Sons, Inc., 1969.

Tinkleman, Sherman N. Planning the objective test. In R. L. Thorndike (Ed.), Educational Measurement. Second Edition. Washington, D.C.: American Council on Education, 1971, 46-80.

Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), Educational Measurement. Second Edition. Washington, D.C.: American Council on Education, 1971, 81-129.

APPENDIX

APPENDIX

INSTRUCTIONS FOR FOCUSED TRUE-FALSE ITEMS

True-False: Presented below are seven items. Each of the statements is either true or false. If you think an item is true, circle the "T" right before the statement; if you think it is false, circle the "F." To assist you in focusing on the part of the statement that is critical to the item being either true or false, a word or phrase has been underlined in each item. You are to assume that the non-underlined part is true and should question the truth of only the word(s) that is/are underlined. For example, for any item you think is false, you should be able to think of (not write down) a word or phrase that could go into the underlined part to make it a true statement.

VITA

Linda Burros Smalley was born in Knoxville, Tennessee, on August 29, 1948. She attended Knoxville City and Knox County Schools and was graduated from Karns High School in June 1966. In August 1969, she received a Bachelor of Arts degree in English from The University of Tennessee, Knoxville. As an undergraduate, the author was selected for membership in Alpha Lambda Delta, Phi Kappa Phi, and Phi Beta Kappa honorary fraternities. After several years teaching experience in secondary schools, the author entered the Graduate School of The University of Tennessee, Knoxville, in 1978. She received the Master of Science degree with a major in Educational Psychology in December 1981.