This is to certify that the

dissertation entitled

THE COMPARISON OF ALTERNATE-CHOICE AND TRUE-FALSE FORMS USED IN CLASSROOM EXAMINATIONS

presented by

NANCY ANN MAIHOFF

has been accepted towards fulfillment
of the requirements for

PH.D. degree in MEASUREMENT, EVALUATION, AND RESEARCH DESIGN

William A McBrews
Major professor

Date MARCH 13, 1986

O-12771

A COMPARISON OF ALTERNATE-CHOICE AND TRUE-FALSE ITEM FORMS USED IN CLASSROOM EXAMINATIONS

By

Nancy Ann Maihoff

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 COMPARISON OF ALTERNATE-CHOICE AND TRUE-FALSE ITEM FORMS USED IN CLASSROOM EXAMINATIONS

By

Nancy Ann Maihoff

The purpose of this study was to compare two-choice alternate-choice and true-false items, and correct answer first (AC_{ci}) and incorrect answer first (AC_{ic}) versions of alternate-choice items on difficulty, discrimination, reliability, and criterion-related validity; and to investigate the practicability of judging the better versions of the alternate-choice and the true-false items.

Three tests were administered to students in a freshman level natural science course. Form A and Form B of Tests I and II contained identical sets of multiple-choice items, 10 alternate-choice items, and 10 true-false items. The alternate-choice and true-false items on Form B were the content equivalent of the true-false and alternate-choice items on Form A. The same 247 students took both Tests I and II.

All alternate-choice items were converted to AC_{ci} and AC_{ic} versions; true-false items were converted to true form (TF_t) and false form (TF_f). Two experienced course instructors were asked to judge which alternate-choice and which true-false versions would best maximize the chances of an informed student correctly answering the item and an uninformed student incorrectly answering the item.

Form A and Form B of Test III consisted of alternate-choice and identical sets of multiple-choice items. Form A contained 10 AC_{ci} and 10 AC_{ic} items; Form B contained the respective content equivalent of these AC_{ic} and AC_{ci} items. There were 102 students who took Test III.

The alternate-choice items were found to be less difficult than the true-false items. Both item forms were equally discriminating and reliable, and both were equally related to final course grade.

The agreement of the judges on the better item version of the alternate-choice and true-false items was only 5 percent greater than that expected by chance.

No differences were found between the two alternate-choice versions on difficulty, discrimination, reliability, or criterion-related validity.

For all three tests, significant interaction effects were found for item position and item content. Control of these two variables is strongly recommended in further research of this type.

Dedicated

to the memory of

Robert L. Ebel

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to the many special people who helped to make this endeavor possible. My gratitude is expressed to Dr. Wm. A. Mehrens, Dr. Lou Anna Kimsey-Simon, Dr John E. Hunter, Dr. Alain F. Corcos, and Dr. Susan E. Phillips for their advice and assistance as members of my doctoral committee.

A special word of appreciation must go to Mr. Michael J. Valiga and Dr. Julie P. Noble of ACT, and Mr. Steven E. Pokorny for their valuable advice and assistance. Also, I wish to thank my students at the California School of Professional Psychology who gave me the strong encouragement to begin this educational endeavor, and a very special thanks to Ms. Thelma A. Weiner who helped me realize it could be attained.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

CHAPTER I

STATEMENT OF THE PROBLEM

College instructors have a variety of item forms available for construction of classroom achievement tests. The essay and the objective-type items such as the short-answer, matching, multiple-choice, and true-false items forms have been most commonly used to sample students' knowledge of subject matter taught in the classroom.

Of these item forms, the use of true-false items on educational achievement tests has been controversial. On one side of the controversy are those who contend that the problem of guessing, the low reliability, and the difficulty in preparing good true-false items should be evidence enough that the true-false items should be abandoned in favor of three-, four, or five-response multiple-choice items (Ahmann & Glock, 1967; Gronlund, 1965; Brown, 1970; and Frisbie, 1971).

On the other side of controversy are those (Ebel, 1979; Smith, 1958; Burmester & Olson, 1966) who point out that true-false items not only have respectably high reliabilities, but they are more efficient than multiple-choice items. This high efficiency means that more true-false questions can be asked within a specific time period than can multiple-choice items, and as a result, true-false tests can provide a much broader sampling of students' knowledge of subject matter. There are those instructors who are willing to accept a lower reliability to obtain a more content-valid examination.

Recently, Ebel (1982) proposed the use of what he termed the 'alternate-choice' item as a replacement for the true-false item. This alternate-choice item form described by Ebel is a modified two-choice multiple-choice item in which the two responses are included within the stem of the question. The following are examples of this item form (Ebel, 1982, p. 272, p. 274):

> An eclipse of the sun can only occur when the moon is 1) full* 2) new.
>
> The density of ice is 1) greater 2) less* than that of water.

The placement of the two responses is not restricted to the end of the stem, but there is freedom to place the responses where they might fit best in the structure of the stem.

Ebel's proposal to replace the true-false item with the alternate-choice item sprang from the results of research conducted in one of his college courses in educational measurement. The findings showed that tests composed of alternate-choice items were less difficult, more discriminating, and more reliable than tests composed of true-false items. Also, because of its brevity, the alternate-choice item form provided an efficiency similar to that of the true-false item in measuring examinees' knowledge.

Earlier, Smith (1958) examined what he called the 'double-choice' item form. The double-choice item is written in a form similar to the alternate-choice item constructed by Ebel, but with more distinct punctuation. Several examples follow (Smith, 1958, p. 387-388):

> The two divisions of the autonomic nervous system are the: (a) central and skeletal (b) sympathetic and parasympathetic*.

> An animal is trained to choose a 2-inch square in preference to a 1-inch square. When later confronted with a 2-inch square and a 4-inch square, he will probably choose the: (a) 2-inch (b) 4-inch*: square.

These double-choice items produced reliabilities ranging from .82 to .90 in tests containing more than 200 items. Smith found that the items were easier to write and that more could be written within a given time period than three-, four-, and five-choice multiple-choice items. In addition, student reaction to the use of the double-choice items was extremely positive.

This initial evidence indicates that the alternate-choice item form may be a superior substitute for the much disparaged true-false item form without losing the positive quality of item efficiency characteristic of the true-false item.

Need for the Study

Few empirical studies have been conducted that compare the alternate-choice type item form to the true false item. Those that have compared these items (Ruch & Stoddard, 1925; Charles, 1926) have found results that differ from those of Ebel (1982). Clearly, further empirical study is needed not only to help clarify the status of the alternate-choice item relative to the true-false item, but also to investigate in greater depth the item characteristics of the alternate-choice item form.

Purpose of the Study

The purpose of this study was threefold: 1) to compare the difficulty level, discrimination level, reliability, and criterion-related validity of the alternate-choice item form with the content equivalent true-false item form; 2) to investigate the practicability of judging whether the alternate-choice item version with the correct

answer listed first (AC_{ci}) or the version with the incorrect answer first (AC_{ic}) is the better form of the item, and whether the true form of a true-false item (TF_t) or the false form (TF_f) is the better form of this item version, and 3) to examine the effects of placing the correct answer first (AC_{ci} version) or the incorrect answer first (AC_{ic} version) on item difficulty, item discrimination, reliability, and criterion related validity of the item.

The inter-comparisons of item characteristics of the alternate-choice and true-false items are presented in Part I of this study, the item version judgment in Part II, and intra-comparison of the item characteristics of versions AC_{ci} and AC_{ic} of the alternate-choice item in Part III.

Hypotheses

Part I. The major hypotheses in Part I of this study were:

H_1: The alternate-choice items will be less difficult than the content equivalent true-false items.

H_2: The alternate-choice items will show higher discrimination than the content equivalent true-false items.

H_3: The reliability of the alternate-choice items will be greater than the reliability of the true-false items.

H_4: The criterion-related validity of the alternate-choice items will be greater than the criterion-related validity of the

ase items.

Part II. The major hypotheses in Part II of this study were:

H_1: Agreement between two departmental colleagues as to the better version of an alternate-choice item will be no better than chance (50%).

H_2: Agreement between two departmental colleagues as to the better version of a true-false item will be no better than chance (50%).

Part III. The major hypotheses in Part III of this study were:

H_1: Version AC_{ci} and version AC_{ic} of the alternate-choice item will not differ in difficulty level.

H_2: Version AC_{ci} and version AC_{ic} of the alternate-choice item will not differ in discrimination level.

H_3: Reliabilities of version AC_{ci} and version AC_{ic} of the alternate-choice item will not differ.

H_4: Criterion-related validities of version AC_{ci} and version AC_{ic} of the alternate-choice item will not differ.

Definition of Terms

Criterion-related validity. In Part I of this study, criterion-related validity is defined as the product moment correlation between the respective total score of the alternate-choice items and the true-false items and the criterion of total weighted score in the course. In Part III of this study, content-related validity is defined as the product moment correlation between the respective total score of AC_{ci} items and the AC_{ic} items and the criterion total weighted score in the course. In Part I, the scores of the true-false and alternate-choice items were removed from the weighted course total scores prior to calculating the correlation coefficient, and in Part III, the weighted

scores of both versions of the alternate-choice items were removed from the course total scores prior to calculating the correlation coefficient.

Item difficulty, p. The difficulty of an item is defined as the proportion of students answering the item correctly. This proportion is represented by the letter p.

Test difficulty. The difficulty of a test is defined as a mean score of the test.

Item discrimination, D. The discrimination level of an item, D, is defined as the value obtained for the item by subtracting, of all the students who answered the item correctly, the proportion of the 27 percent lowest scoring students from the proportion of the 27 percent highest scoring students. The Index D ranges from -1.00 to 1.00.

Item discrimination, r_{pbis}. The discrimination level of an item, r_{pbis}, is defined as the product moment correlation coefficient of the continuous variable of total score and the dichtomous variable of correct answer (1) or incorrect answer (0) for the item.

Test Reliability. The reliability, r_{tt}, of a test in this study is defined as the degree of internal consistency of the test as estimated by the Kuder-Richardson-20 formula.

Overview of Dissertation.

In Chapter II, the literature relevant to the general research of multiple-choice item forms and to specific comparative research covering true-false and two-choice/alternate-choice type item forms is reviewed. In addition, literature related to research on item conversion methods and item sequence effects is reviewed, in that this literature is relevant to the item conversion methods and item

sequencing procedures used in this study. In Chapter III, the sample, the instrumentation, the procedures used, the operational form of the hypotheses, and the analyses used are presented. The results of the study for each part by each hypothesis are presented in Chapter IV. The final chapter, Chapter V, contains a summary of the study, a discussion of the findings, the limitations of the study, and suggestions for future research.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The review of the literature is organized and presented in the three areas relevant to this research. The first section of this chapter concerns the review of applicable comparative studies of objective item forms on item characteristics of difficulty, reliability, and/or validity and the more specific studies that compared the two-choice multiple-choice item and true-false item forms on these item characteristics. The second part of the review of literature concerns the methods used to convert items from one form to another item form. The third part of the review of literature concerns the effects of item sequence on the item characteristics of power tests. A chapter summary follows at the end of this chapter.

Studies Comparing Two-choice/Alternate-choice and True-false Item Forms

One of the earliest studies to examine the two-choice item was carried out by Ruch and Stoddard (1925). These investigators constructed five versions of a 100-item examination to test general information knowledge in history and social sciences. Each 100-item examination was composed of two forms, Form A and Form B. Each form contained fifty parallel items. Split-half reliabilities for each examination were calculated using Form A versus Form B. Version I of the 100-item examination was composed entirely of recall items; Version II of five-choice multiple-choice items; Version III of three-choice

multiple-choice items; Version IV of two-choice multiple-choice items; and Version V of true-false items. Examples of the item forms found in these version of the tests follow (Ruch & Stoddard, 1925, pp. 90-91):

Recall

1. The American Revolution began in the year ___1775___

Five-choice

1. The American Revolution began in 1762
   1775 1783 1789 1812 ___1775___

Two-choice

1. The American Revolution began in 1762
   1775 1789 ___1775___

True-False

1. The American Revolution began in 1775 True False

The examinee was to write the correct answer in the space to the right of the recall, five-, three- or two-choice item. The correct answer was to be underlined or circled for the true-false items.

The senior classes of 24 Iowa high schools were divided alphabetically by surnames into four groups, each containing approximately 135 students. The recall examination was administered to all students on the first day. The following day the five-choice multiple choice test was administered to Group I; the three-choice multiple-choice test, to Group II; the two-choice multiple-choice test, to Group III; and the true-false test, to Group IV.

Of particular interest to the researchers in this study were the cooperative merits of the two-choice multiple-choice item form and the true-false item form, given that their pure guessing probabilities are 50:50. The results of this study are summarized in Table 1.

Table 1

Summary of Results from Ruch and Stoddard Study

Item form	50 Item $\underline{r}_{tt}$[b]	100 Item $\underline{r}_{tt}$[b]	Mean (SD) Form A	Mean (SD) Form B	$\underline{n}$	$\underline{r}_{xy}$[c]	No. per 100 recall	Adjusted $\underline{r}_{tt}$[d]
Recall	0.811	0.896	12.18 (6.37)	10.85 (7.86)	562	-	100	0.896
5-choice	0.796	0.886	27.20 (5.73)	22.80 (7.68)	137	0.861	177	0.901
3-choice	0.598	0.748	30.61 (5.73)	26.41 (6.06)	134	0.755	139	0.806
2-choice	0.737	0.749	35.64 (5.73)	31.98 (5.88)	135	0.713	164	0.902
True-False	0.55	0.714	30.06 (5.98)	27.67 (6.84)	133	0.480	183	0.820

Note: Statistics reported are for scores uncorrected for chance.
[a]Form A vs Form B.
[b]Reliability using Spearman-Brown formula.
[c]Concurrent validity with recall items as criterion.
[d]Reliability adjusted for number of items administered per time period.

Although there were no statistical tests performed on the differences between reliabilities of the forms, Ruch and Stoddard concluded that there are no practical differences in reliability, particularly between the three-choice, two-choice, and true-false item tests. Upon examination of the mean scores, the true-false version appears to have been more difficult than the two-choice version of the examination. Also, the authors noted that considerably more two-choice items and true-false items could be administered within a given time period than could other item types.

The researchers noted that the correlation of the two-choice and true-false items with the recall items seemed to indicate that there were differences in the kinds of mental processes brought into play when answering these items:

> It should always be remembered that we cannot assure that restatement of recall items in multiple-response or true-false forms does not alter their relative difficulties. However, such a lack of perfect agreement as our coefficients show probably cannot be accounted for by the matter of changed difficulties but must be in large part indicative of differences in the pedagogical and psychological characteristics of the several types. (Ruch & Stoddard, 1925,pp. 92-93)

In a subsequent study, Charles (1926) administered a 100-item general psychology examination to 747 college students (this study was later published with Ruch: Ruch & Charles, 1928). Five versions of the examination were constructed, each version contained 50 items in Form A, and 50 parallel items in Form B. Version I of the examination contained only recall items; Version II contained only five-choice multiple-choice items; Version III only three-choice multiple-choice items; Version IV, only two-choice multiple-choice items; and Version V, only true-false items.

The following are examples of the item forms used (Charles, 1926, p. 399):

Recall

1. A synapse is a junction between two _____neurones_____

Five-choice

1. A synapse is a junction between two (1) dendrites, (2) axones, (3) muscles, (4) neurones, (5) bones. _____4_____

Three-choice

1. A synapse is a junction between two (1) neurones, (2) dendrites (3) axones. _____1_____

Two-choice

1. A synapse is a junction between to (1) neurones, (2) dendrites. _____1_____

True-False

1. A synapse is a junction between two dendrites. True False

In the recall version of the test, the examinees were instructed to write the correct answer in the space provided to the right of the question. In the five-, three-, and two-choice versions of the test, the examinees were instructed to write the number of the correct answer in the space provided. In the true-false version, 'True' or 'False' was to be underlined.

The recall version of the examination was administered on Day 1 to all students. On Day 2, the examinations were sequentially arranged and distributed so that every fourth student received the same version of the test. A summary of the results is shown in Table 2.

The results of this study showed the two-choice item form to be considerably less reliable than the true-false item form (see Table 2). When difficulty levels were compared, the true-false item

form was shown to be more difficult than the two-choice item type, a similar result to that found by Ruch and Stoddard (1925). Again, when compared to other item types, substantially more two-choice and true-false items could be administered within a given time period.

Table 2

Summary of Results from Charles Objective Item Study

Item form	50 Item $\underline{r}_{tt}$[a]	100 Item $\underline{r}_{tt}$[b]	Mean (SD) Form A	Mean (SD) Form B	$\underline{n}$	$\underline{r}_{xy}$[c]	No. per 55 minutes	Adjusted $\underline{r}_{tt}$[b]
Recall	0.603	0.752	11.33 (2.04)	15.60 (2.92)	747	0.311	154	0.824
5-choice	0.680	0.809	27.64 (3.10)	31.31 (2.94)	182	0.213	220	0.903
3-choice	0.624	0.768	32.82 (2.73)	37.90 (2.64)	188	0.313	258.5	0.896
2-choice	0.477	0.646	37.10 (2.07)	40.39 (2.09)	188	0.319	285	0.839
True-False	0.602	0.751	32.90 (2.27)	34.47 (2.33)	189	0.226	302.5	0.901

Note: Statistics reported are for scores uncorrected for chance.
[a]Form A vs Form B.
[b]Reliability using Spearman-Brown formula.
[c]Predictive validity using final grades as criterion.
[d]Reliability adjusted for number of items administered per 55 minutes.

Predictive validity of the item subtest scores with final grades was quite low for all subtest scores, with the highest correlation ($\underline{r}_{xy}$ = 0.319) between the two-choice subtest and final grades, and the two lowest between the five-choice ($\underline{r}_{xy}$ = 0.213) and the true-false ($\underline{r}_{xy}$ = 0.226) scores and final grades. Charles considered these low

correlations the result of limited variability in final grades given by instructors.

In a more recent study, Ebel (1982) compared what he termed the "alternate-choice" test item with the true-false item. The alternate-choice item form proposed by Ebel is similar in construction to the two-choice multiple-choice item form used by Charles except that in the alternate-choice form the responses can be placed in positions other than at the end of the item. Ebel provides several examples (pp.272-273):

> An eclipse of the sun can only occur when the moon is 1) full 2)new.
>
> Most teachers believe that if tests were to be abandoned there would be 1) serious 2) very few educational losses.

In his study, Ebel administered two 25-item end-of-unit tests to 28 students enrolled in an educational measurement course. One of the tests was composed of true-false items, the second test of alternate-choice items. Each test was composed of independent samples of items that did not ask the same questions. For each of the eight units comprising the course, either the true-false or alternate-choice version was administered first. By the end of the term, both the true-false and the alternate-choice versions had each been administered first four times.

The scores were summed for the eight alternate-choice form tests and the eight true-false form tests. The results of this study are shown in Table 3.

Table 3

Summary of Ebel's Alternate-choice and True-false Item Study

Item form	25 Item KR-20 r_{tt}	100 Item KR-20 r_{tt}[a]	Mean	SD	Index Diff[b]	Index Disc[c]
Alternate-Choice	0.67	0.89	19.99	3.08	0.20	0.30
True-False	0.47	0.78	18.80	2.81	0.25	0.28

Note: All statistics reported are averages over eight unit tests, and scores are not corrected for chance.
[a]Reliability using Spearman-Brown formula.
[b]Difficulty is expressed as $1 - p$
[c]Based on D, the difference between the upper-lower 27% scores.

As in the previous studies, the true-false version of the test appears to have been more difficult than the alternate-choice version. Reliability for the alternate-choice items was considerably higher than for the true-false items, and it appears that both item forms equally discriminate.

Item Conversion Procedures

When items are converted from one form to another; for example, from four-choice to alternate-choice, there has been concern that this conversion might effect item characteristics such as reliability, validity, difficulty, or discrimination. In reviewing item conversion methods, Owens, Hanna and Coppedge (1970) found that three methods most commonly used were: the judgmental, the frequency, and the discrimination methods.

In the judgmental method, the item author supplies the distractor or distractors considered most plausible. The frequency and discrimination methods are more empirical in nature. According to the

frequency method, the most frequently chosen incorrect response or responses are used as the respective distractor or distractors in multiple-choice items, or for false versions of true-false items. With the discrimination method, the incorrect responses are used that discriminate most highly between high scorers and low scorers.

A small number of studies have been conducted comparing the effects of these methods on item reliability, validity, discrimination, and difficulty. Loree (1948) studied the relative effects of using the judgmental and the frequency methods for selecting multiple-choice item distractors for tests in three subject areas: arithmetic, health knowledge, and word meaning.

Three forms of each test were developed. Form A consisted of multiple-choice items for which the examiner "conceived" of the distractors. Loree defines the term 'conceived' as any process by which the test constructor obtains distractors to multiple-choice items, other than by obtaining direct evidence of the kinds of responses students make to a specific test item when that item is presented in a free response form (p. 6). In other words, the examiner judges without empirical method which distractors would be the most appropriate to include in the multiple-choice item.

Form B of the examinations consisted of the same items in recall form. Form C consisted of the same items, but these were in multiple-choice form in which the distractors had been developed from the most frequent response errors to the recall items on Form B.

All forms of each test were administered to high school students in the Chicago area. Form B, containing the recall items, was administered first. Form C was then developed from the results of Form B, and both

Form A and Form C were administered to the same students three weeks later.

To investigate the effects of item conversion methods on concurrent validity, the total score correct for each form was correlated with every other form. The results are shown in Table 4. There were no significant differences between the correlations of these three forms for any test.

The effects on the difficulty level of the two item conversion methods were tested by t-tests. For the Arithmetic Problems Test, the means of Form A, Form B and Form C were significantly different from

Table 4

Summary of Results from Loree Item Conversion Study

Test	10 Item $\underline{r}_{tt}$[a]		20 Item $\underline{r}_{tt}$[b]		Mean			Concurrent Validity		
	Form A	Form C	Form A	Form C	Form A	Form B	Form C	$\underline{r}_{ab}$	$\underline{r}_{bc}$	$\underline{r}_{ac}$
Arithmetic	0.802	0.820	0.844	0.901	11.03	8.39	9.57	0.817	0.847	0.844
Health Knowledge	0.451	0.369	0.622	0.539	14.84	8.17	10.57	0.578	0.580	0.744
Word Meaning	0.764	0.714	0.866	0.833	28.03	14.56	19.24	0.812	0.735	0.790

Note: Statistics reported are for scores uncorrected for chance.
[a]Parallel Split Half reliability
[b]Reliability using Spearman-Brown formula.

each other. For the other two tests, the mean on Form A was significantly higher (easier) than on Form B (see Table 4). There were no differences in the reliability of Form A and Form C for any of the tests.

Owens, Hanna, and Coppedge (1970) investigated the effects on reliability of multiple-choice items converted from recall form by the judgmental method, the frequency method, and the discrimination method. A 33-item geometry recall test that had been administered to 357 high school students was the source from which 17 four-choice multiple-choice questions were developed. The same multiple-choice items in each of the three forms of the examination had exactly the same stems and the same correct responses. Only the three distractors differed from form to form.

For the judgmental form of the exam, each of 13 secondary mathematics teachers supplied three distractors they judged most appropriate for each item. The three most frequently mentioned distractors for each item were retained.

The item analysis for the recall test was used to develop the other two forms of the examination. For the frequency form of the exam, the three most frequently produced incorrect answers were selected as the distractors. For the discrimination form of the exam, the three distractors were selected that discriminated the highest.

First a parallel recall test and then the three forms of the multiple-choice examinations were administered to 1875 students enrolled in high school geometry. The examinations were sequentially ordered so that every third student received the same form of the test. After the administrations, the students' examinations were divided into three

equivalent groups based on recall exam scores and on which form of the multiple-choice examination the student took.

The reliability coefficients for the judgmental, frequency, and discrimination forms were .556, .620, and .614, respectively, and concurrent validity coefficient with the recall examinations were .617, .646, and .647, respectively (when corrected for attenuation they increased to .982, .973, and .979). The validity coefficients were tested for homogeneity, and no significant differences were found. Although the differences in reliabilities were not tested, the investigators concluded there was little practical difference between them.

Burmester and Olson (1966) used the frequency method in a study to determine if college-level natural science alternate-response items could show the same desirable item characteristics as previously administered five-choice multiple-choice items. It must be noted that the two responses available in these alternate-responses items were 'Acceptable' and 'Unacceptable', and thus were actually a variation of the true-false item form.

A total of 85 multiple-choice items was selected on the basis of performance on previously administered final exams. The average difficulty, $\underline{p}$, of these items was .57, and the average discrimination index (Flanagan) was .45. Based on the item analysis for each item, a multiple-choice item was converted to true form if the distractors were approximately evenly selected. The correct answer was simply added to the stem to make the item a true statement. If a distractor was shown to be particularly attractive, the item was converted to a false statement by adding this incorrect answer to the stem.

The resulting 37 'acceptable' (true) items and 48 'unacceptable' (false) items were administered to 110 students in the same natural science course. The KR-20 reliability for the alternate-choice items was .86. The distribution of difficulty levels and the discrimination levels of the multiple-choice and alternate-choice items are shown in Table 5.

The results show that alternative-response items are less difficult, and that they discriminate as well as the five-choice items, and show high reliability ($\underline{r}_{tt} = .86$). The investigators also found that more alternate-choice than five-choice items could be administered within a given time, thus providing for greater sampling of content areas or educational objectives.

More recently, Frisbie (1971) conducted a three-phase study to compare the reliabilities of true-false and multiple-choice high school general knowledge science and social studies tests. In phase I, four-choice multiple-choice items were converted to true-false items using the judgmental and discrimination methods. For the judgmental method, five high school science teachers and five high school social studies teachers were used to judge the best distractors of multiple choice items from a published standardized test in their respective fields. If four out of five judges agreed on the best distractor, it was used to make a false true-false item; if fewer than four judges agreed on a distractor, the item was made true true-false item. This resulted in a social studies test of 41 false and 29 true statements and a natural science test of 45 false and 25 true statements.

For the discrimination conversion method of this phase, the original social studies and natural science multiple-choice tests were

Table 5

Summary of Burmester and Olson Alternate-choice Study

Item Statistic	Multiple-Choice f	Alternate-Choice f
Difficulty Index *p*		
81-100	4	27
61-80	39	42
41-60	27	12
21-40	14	4
00-20	1	0
Discrimination Index (Flanagan)		
61-80	13	11
41-60	38	33
21-40	34	28
00-20	0	13
Discrimination Index *D*[a]		
61-80	-	2
41-60	-	21
21-40	-	35
00-20	-	27

[a]Discrimination index *D* was not available for multiple-choice items.

each administered to 100 students, and the discrimination index, D, was calculated for each distractor. If D was less than .20, or did not differ from the other indices by more than .09, the item was converted to true; otherwise, the distractor with the largest D was used to make the item false. The resulting 70-item true-false social studies test that was developed contained 33 true and 37 false statements; the 70-item science test also had 33 true and 37 false statements.

In phase II, the social studies and science true-false tests were administered to a sample of students to identify ambiguous items and improve them.

In phase III, each of the four tests, social studies-judgmental method (SJ), social studies-discrimination method (SD), science-judgmental method (NJ), and science-discrimination method (ND), was divided into two forms, and the original multiple-choice items were added to each test. Form A of the SJ test contained multiple-choice items 1-35 and the converted SJ true-false items 36-70. Form B contained the converted SJ true-false items 1-35 and the multiple-choice items 36-70. The other tests were arranged similarly.

A total of 1018 high school students was administered one of these eight forms. Eight minutes after the beginning of the exam, students were asked to stop and write the number of the item on which they were working. The median number of multiple-choice items attempted within this time period was 17.04, and the median number of true-false items attempted was 25.59. Thus, in this sample, students attempted three true-false items for every two multiple-choice items.

When the KR-20 reliability of the true-false and multiple-choice items for each form was tested using paired t-tests, the true-false

items were shown to be consistently lower in reliability than the multiple-choice items. The KR-20 reliability of the true-false items converted by the judgmental method was than compared to the KR-20 reliability of the true-false items converted by the discrimination method. A paired *t*-test showed no significant difference between the reliability of items converted by these two methods.

Williams and Ebel (1957) used the discrimination method to convert four-choice multiple-choice Iowa Test of Educational Development vocabulary items to three- and two-choice items. The least discriminating distractor was dropped from the four-choice item to convert it to a three-choice form, and the two least discriminating distractors were dropped to convert to a two-choice item. Three forms of the test were developed, each containing exclusively either four-, three-, or two-choice items.

These three forms were arranged in an alternate sequence and administered to all students in 6 four-year Iowa high schools. Students were given 30 minutes to complete as many items as possible. The score was based on the number correct.

The results (see Table 6) show that there were no significant differences among the reliabilities of these three forms (using the first 85 items of each form). The difficulty level of the item decreased as the number of responses decreased; the discrimination of the items also decreased accordingly. In addition, considerably more two-choice than four-choice or three-choice items could be completed within the 30-minute time period.

Item Sequence

When examinations are administered under crowded classroom conditions, college instructors, to prevent cheating, often develop two or more forms of the examination in which the same items are arranged in different sequences. Several authorities in the measurement field contend that there exists a sequence effect (Cronbach, 1970) or serial error (Stanley, 1961) caused by these different item arrangements. Sequence effect is the discouragement a student feels as the result of failing to answer an item, and serial error is the failure to answer items that follow a particularly troublesome item.

One common hypothesis is that the underlying dynamic and cause of this problem is test anxiety of the examinee (Mckeachie, Pollie, & Speisman, 1955; Mandler & Sarason, 1952; and Alpert & Haber, 1960). If the student encounters a particularly troublesome item, the student's anxiety level increases, and this increase adversely affects subsequent

Table 6

Summary of Williams and Ebel Item Conversion Study

Item form	n items finished	KR-20 r_{tt}	Mean (SD)	Index Diff	Index Disc[a]
Four-choice	85	.945	42.14 (16.70)	.496	.487
Three-choice	94	.941	49.29 (15.78)	.585	.471
Two-choice	129	.929	58.22 (14.01)	.684	.412

Note: Test statistics are based on the first 85 items of each test.
[a]Discrimination Index D.

performance. To minimize test anxiety, then, helps the student to maximize test performance, and measurement texts often suggest that items be arranged from least difficult to most difficult. When different item forms are used, it is suggested that the item forms be ordered from the most simple to the most complex (e.g. true-false, matching, short answer, multiple choice, interpretative, and essay questions), and that within each item form, the items be arranged from least difficult to most difficult (Gronlund, 1977).

A number of studies of college classroom examinations have been conducted examining the effects of arranging items by difficulty level. All examinations in these studies were power tests. The earliest study was conducted by Brenner (1964), who administered a series of examinations to students in educational psychology classes over two quarters. During the first quarter, Q1, a total of 320 multiple-choice test items were written and administered over a series of four examination periods. For each period, there were two forms of the examination administered. Each form contained 40 items that had been randomly arranged using a table of random numbers. The purpose of administering these items was to obtain a difficulty value (percentage correct, uncorrected for chance), and discrimination value (point-biserial) for each item.

During the subsequent quarter, Q2, the same items, or a subset thereof, was again administered to a new group of students. Items administered during the first testing period of Q1 were administered during the first testing period of Q2. Those items administered the second, third, and fourth testing periods of Q1 were administered during the respective second, third, and fourth testing periods of Q2.

During the first testing period of Q1 and Q2, the same examination form of 40 randomly arranged items (Form IB) was administered to check for reliability (stability) of item difficulties. Two other forms of the same items were also administered during this first testing period. Form IA contained items arranged by difficulty from easy to hard, and Form IC contained items arranged from hard to easy.

For the second testing period of Q2, two forms of a test were developed from the pool of eighty items tested during the second testing period of Q1. From this pool, Brenner selected the 10 easiest items, the 10 hardest items, and 20 items that in combination best reflected the course content and were highest in discrimination. Form IIA contained the 10 easiest items (in order of increasing difficulty), followed by 20 randomly arranged course content items and then the 10 hardest items. Form IIB contained the ten hardest items (in decreasing order of difficulty) followed by 20 randomly arranged course content items and then the 10 easiest items.

For the third testing period of Q2, two forms of an examination were constructed from items of only one of the examinations administered during Q1. Form IIIA contained items arranged by difficulty from easy to hard; Form IIIB contained items arranged from hard to easy.

From the 80 items available for the fourth testing period of Q2, 40 items were selected in the same manner as those for Forms IIA and IIB. However, in Form IVA, <u>all</u> the items were arranged from easy to hard, and on Form IVB <u>all</u> the items were arranged from hard to easy.

Brenner calculated the following test statistics for each form of each examination: mean score correct on the examination, the KR-8

reliability coefficient, and the mean discrimination (average point-biserial correlation between item and total test score). Significant differences between average difficulty, average discrimination, and reliability for each pair of examination forms for each testing period were determined using the t-test. The results are shown in Table 7.

Table 7

Results of Brenner Item Arrangement Study

Exam form	Mean score Difficulty	Reliability	Discrimination[a]
IA	21.18	.578	.220
IB	21.04	.553	.232
IC	20.90	.598	.218
IIA	20.04	.753	.283 *
IIB	23.93	.674	.250 *
IIIA	26.14	.778	.309
IIIB	26.33	.805	.326
IVA	23.69	.836	.284
IVB	24.17	.747	.289

[a]Average point-biserial correlation between item and total test score.
*$p<.02$ using a paired t-test.

Clearly, difficulty-based item arrangement did not affect the mean scores of the examination nor the reliability, and for the most part, did not effect discrimination. The significant difference found in discrimination of Form IIA and IIB was not explained by Brenner.

Brenner did not report if the item difficulties of Form IB were reliably replicated from Q1 to Q2, however, research by Carter (1942), Davis (1951), and Gibbons (1940) has shown that item difficulty values are highly reliable across administrations.

Monk and Stallings (1970) analyzed the data available on 11 tests administered to students in a college-level basic geography course.Each test had two forms, and each was administered at some time between 1965 and 1968. The number of items in each test varied from as few as 80 to as many as 200. The objective of producing two forms of each test was to reduce the likelihood of two students in adjacent seats working on the same question. The same items were used in both forms of each test; however, the patterns of item arrangement differed in each form. In one form, the test items were grouped by item form (true-false, matching and multiple-choice). In the second form, the arrangement of the item-form groups was altered and the order of the individual items within each group was changed.

The significant differences between the mean scores for the two forms of a test were assessed by using the *t*-test. The number of items, means, standard deviations, and KR-21 reliabilities for each form are shown in Table 8. Significant differences were found between the mean scores of only two pairs of tests. Monk and Stallings pointed out, however, that one of these significant pairs of tests (7 and 8) had been corrected with a key containing ten errors. There were only slight variations that existed in the reliabilities of each pair of tests.

Huck and Bowers (1972) conducted two studies on the effects of item sequence and *p* values (proportion of examinees who answered an item correctly). In the first study 10 forms of a 60-item final examination were administered to 120 college students enrolled in an introductory psychology course. The items were the same on each form, the difference was only in the item order on each form. The item order on each form was such that six balanced Latin Squares were formed.

Table 8

Results of the Monk and Stallings Item Arrangement Study

Test Pair	Number of test items	Mean score	SD	n	KR-21 r_{tt}
1	100	69.66	13.44	89	.892
2	100	72.10	11.10	77	.845
3	100	73.11	16.56	10	.938
4	100	75.21	11.14	90	.858
5	100	68.46*	18.15	94	.944
6	100	73.49*	10.74	84	.840
7	100	69.97**	10.92	124	.826
8	100	62.62**	9.14	132	.727
9	100	60.28	12.46	123	.854
10	100	62.21	11.74	121	.838
11	80	58.75	8.67	68	.802
12	80	58.61	8.21	71	.707
13	80	52.36	10.12	66	.834
14	80	50.55	9.68	69	.811
15	200	134.43	25.07	63	.935
16	200	129.13	23.21	63	.920
17	80	49.70	8.19	138	.728
18	80	48.53	8.29	118	.731
19	80	48.74	9.86	118	.814
20	80	48.29	9.86	121	.813
21	200	130.02	23.36	120	.921
22	200	129.93	22.26	118	.916

*$p<.01$
**$p<.001$

A special ANOVA procedure (Cochran & Cox, 1957, pp. 133-139) was used to test for residual effects of item order on the *p* values. None of the *F*-values for the six Latin Square designs was significant.

In their second study, Huck and Bowers administered six forms of a 50 item midterm to 162 students in the same introductory psychology course. The items were arranged to form six Latin Square designs. The special ANOVA procedure again showed no significant sequence effects for any of the Latin Square designs.

A recent study was conducted by Plake (1980). Three forms of a 96 item multiple-choice examination were administered to students in a course in psychiatric nursing. The items for this examination were selected from a pool of items from a test that had already been administered and for which item difficulty and reliability statistics were available. The item difficulties ranged from .20 to .96, and the KR-20 reliability of the originally administered test was .85. On the first form of the examination, the 96 items were placed in difficulty order from easy to hard. Item order for the second form was random, and the order for the third form was spiral cyclical. In the spiral cyclical form, every four items were arranged from easy to hard; thus, the cycle of easy to hard was repeated every fifth item.

One-half of the examinations contained directions that explained the respective item ordering and that gave test-taking strategies for that item order. The other one-half did not contain these directions.

A set of questions were added to the end of the examination which asked students to rate, on a 1-5 scale, the fairness of the test, the perceived difficulty of the test and to estimate their performance on the test.

A 3 X 2 multivariate ANOVA was performed on the dependent measure of total score on the examination, rated fairness, perceived difficulty, and performance estimates. There were no significant interaction effects for item order and directions, nor were there significant main effects for item order or for directions.

Klosner and Gellman (1973) administered three forms of a 75 item multiple-choice final examination to students enrolled in an educational measurement course. Each form contained the same items. Form S contained items arranged in the order the subject-matter was presented in the course. Form S X D contained items grouped by subject-matter, but within each subject matter topic the items were arranged by difficulty. Form D contained items grouped only by easy to hard difficulty level.

To match for ability, students were ranked according to their midterm grade and divided into triads. Students within each triad were randomly assigned to take one of the three forms of the examination. The means, standard deviations and reliability of each of the three test forms are shown in Table 9.

Using the median midterm examination grade, students were then split into a high achieving group and a low achieving group. A 2 X 3 ANOVA of the test scores showed only achievement grouping to be significantly related to total test score. Neither the main effect of item order, nor the interaction of item order with achievement level was significant.

Smouse and Munz (1968) developed three forms of a 100 item multiple-choice final examination for a course in introductory psychology. Each form contained the same items but differed in the

Table 9

Results of the Klosner and Gellman Item Order Study

Test Form	n	Mean	SD	KR-21 r_{tt}
Form S (Subject)	18	60.06	5.98	.675
Form S X D (Subject and Difficulty)	18	61.67	5.10	.586
Form D (Difficulty)	18	59.72	6.08	.680

difficulty-based order of items. In the respective forms, the item order was from easy to hard, hard to easy, and randomly mixed.

These three forms were administered to two randomly assigned groups: a high test-taking anxiety group, and a normal test-taking anxiety group. For the normal test-taking anxiety group, the usual test taking atmosphere was maintained; however, for the high test-taking group:

> The anxiety-provoking treatment consisted of informing the Ss that because of "widespread cheating" on previous examinations, should their individual performances drop significantly below previous examination scores, they would have to take a special oral examination administered by the coordinator of the introductory psychology sections. Further, the examination was administered by a professor rather than the graduate instructor and was proctored by assistants who continually circulated among the Ss. (Smouse & Munz,1968, p. 182)

Stapled at the end of each examination was the Multiple Affect Adjective Check List (MAACL) developed by Zucherman (1960). The purpose of administering the MAACL was to measure the amount of anxiety felt in each test situation.

A 2 X 3 ANOVA for unequal n's was performed on the total number of

items answered correctly and on the MAACL scores. The item arrangements, the anxiety treatments, and the interactions between these two independent variables were not significant.

In a subsequent study, Munz and Smouse (1968) administered the same three forms of the 100 multiple-choice item final examination to another group of students enrolled in the introduction to psychology course. Prior to the examination, each student had completed the Achievement Anxiety Test (AAT) developed by Alpert and Haber (1960) and was placed into one of four achievement anxiety type groups based on their AAT scores: Facilitators, Debilitators, Non-affecteds, and High-affecteds. According to Alpert and Haber, a facilitator is an individual whose test performance is facilitated by the anxiety-provoking situation; the debilitator is an individual whose test performance is depressed by the anxiety-provoking situation; the non-affected is an individual whose test performance is not affected by the anxiety provoking situations; and the high-affected is an individual who has the potential as both a facilitator and a debilitator.

A 3 X 4 ANOVA was performed on mean score correct on the final examination. No significant differences were found for the main effects of item order and achievement-anxiety type. A significant but unclear interaction between anxiety types and the random and easy to hard arrangement was found.

Marso (1970) administered three forms of a 103 item multiple-choice comprehensive final exam to students enrolled in several sections of an introductory educational psychology course. One form contained the items in random order. The other two forms contained the items grouped by course content. One form contained the items grouped in the order

they were presented during the course; the other form contained the items grouped in the opposite order in which they were presented during the course.

Several days prior to the final examination, each student was given a set of test anxiety scales developed by Carrier and Jewel (1966) which was used to classify the students as to level of test-taking anxiety. Students were divided into an upper, a middle, and a lower third base on the total anxiety scale score for the sample.

A 3 X 3 ANOVA for unequal n's was performed on the final examination scores and on the total time in minutes taken to complete the examination. Only the main effect of test anxiety was significant for final examination scores, showing students with high anxiety to perform more poorly than middle or low anxiety students. Neither main effects nor interactions between the main effects were significant for the number of minutes taken to complete the exam.

Chapter Summary

The first series of studies reviewed in this chapter concerned the effects of item form on the item characteristics of difficulty, reliability, and/or validity. When the two-response type multiple-choice item form has been compared to the true-false item form, the two-response type item has consistently been found to be less difficult than the true-false item. When the reliabilities of these two item forms are compared, the results are less consistent: of the three studies, one study showed the two-response item type to be more reliable than the true-false item (Ruch & Stoddard, 1925), one study showed no practical difference (Ebel, 1982), and one study showed the two-response item type to be less reliable than the true-false item (Charles, 1926). From the

results of concurrent validity studies of the two-choice and the true-false items with recall items, Ruch and Stoddard (1925) suggest that answering the two-choice, and particularly the true-false item, may require different mental processes.

The next series of studies reviewed concerned the methods of converting items from one item form to another item form, and the possible effect of the conversion method used on the item characteristics of difficulty, discrimination, reliability, and/or validity. When items were converted from recall to multiple-choice form by use of the frequency and judgmental methods (Loree 1948), or by these two methods and the discrimination method (Owens, Hanna, and Coppedge, 1970), no significant differences were found in the reliability of the multiple-choice items, or in the concurrent validity of the recall or multiple-choice items. The results were different for the difficulty of the multiple-choice items, however. Items converted by the judgmental method were found to be significantly easier than items converted by the frequency method (Loree,1948).

When items were converted from multiple-choice form to true-false type item form by use of the frequency method (Burmester and Olson, 1966), the true-false form was found to be less difficult, equal in discrimination to its original five-choice multiple-choice form, and highly reliable. When the judgmental and discrimination methods were used to convert multiple-choice items to true-false form (Frisbie, 1971), no differences were found in the reliability of the true-false items converted by these methods. It should be noted, however, that the true-false items were significantly lower in reliability than the original four-choice items.

When four-choice multiple-choice items were converted to three-choice and to two-choice form by use of the discrimination method (Williams & Ebel, 1957), no significant differences were found in the reliability of the converted items. It was found, however, that the difficulty and discrimination of the items decreased as their number of responses decreased.

The last series of studies reviewed concerned the effects of item arrangement, or item sequence, on the test characteristics of power tests. It has been hypothesized that, because of anxiety caused by the failure to answer an item (sequence effect) and/or the failure to answer items that follow a troublesome item (serial error), students perform differently on tests in which the items are arranged differently. Several studies have examined the effects of item order on difficulty, discrimination, and/or reliability without controlling for anxiety of students. Monk and Stallings (1970) compared two forms of each of eleven tests on difficulty and reliability. Each item form differed only as to the arrangement of items. No differences in reliability were found, and only two of the eleven pair of tests differed significantly in difficulty.

The effects of arranging items by random or by various degrees of difficulty were examined by several researchers (Brenner, 1964; Huck & Bowers, 1972; Klosner & Gellman; 1973, and Plake, 1980) No significant differences were found in difficulty of the tests administered, no matter what item arrangement was used. In addition, Brenner found no differences in the reliability of nine test forms, each of which contained a different item order.

The remaining studies reviewed concerned the effects of item order

on test difficulty that also included some measure of student anxiety in the research design. In one study (Smouse & Munz, 1968) no significant differences were found for student scores on tests containing items ordered differently by difficulty, nor in the scores of high and low anxiety students. In a subsequent study (Munz & Smouse, 1968), a significant but unclear interaction was found between anxiety types and item order for total test scores. In a later study (Marso, 1970), significant differences were found in test scores of the anxiety groups but not for item order. Thus, although anxiety may play a part in student performance on a test, there is little evidence that this anxiety is related in any way to item order.

CHAPTER III

DESIGN AND PROCEDURE

Introduction

This research study was conducted in three parts. Part I was designed to examine the difficulty and discrimination levels of alternate-choice and true-false item forms; the reliabilities of the true-false subtest scores and the alternate-choice subtest scores; and the criterion-related validities of true-false and alternate-choice items as measured by the correlation between each respective subtest score and final grades. Part II was designed to investigate the practicability of judging the best version (TF_t or TF_f) of the true-false item and to determine whether the alternate-choice version with the correct-answer-first (AC_{ci}) or with the distractor-first (AC_{ic}) is the best form of the item. Part III was designed to examine the effects of placing the correct answer first (AC_{ci}) or placing the incorrect answer first (AC_{ic}) in the alternate-choice responses on difficulty, discrimination, reliability, and criterion-related validity.

Part I

Sample. The students that participated in Part I of this research were from seven sections of a natural science course offered at Michigan State University in the fall quarter of the 1983-84 academic year. The lectures of the seven sections were team-taught by the same two professors and, although the lab sessions were not team taught, the same material was covered during the quarter.

The sample of students who participated did so because their instructors agreed to include the alternate-choice and true-false items on their midterm and final examinations. These students cannot be considered a random sample, since they were not chosen in such a way that each student taking this natural science course in the fall quarter had an equal and independent probability of being selected. It can be argued, however, that these seven sections can be considered representative of the population of students taking this course in the fall quarter, particularly in regard to the cognitive skills required to master the material taught and to take the course examinations. This argument is based on the manner in which students select the sections of this course.

Briefly all lower division students are required to enroll for a specific number of general education credits in the area of biological and mathematical sciences. The majority of incoming freshman students enroll in this natural science course, Natural Science 115, in the fall quarter to fulfill part of their minimum requirements in this area (according to personal communication with the Assistant Provost for Undergraduate Education, July 1, 1984). As a result, there are approximately 36 sections of this course offered each fall quarter.

Most often the section chosen by the student is based either on the time it is offered, or the space availabile in a given section. It is rare that these incoming freshmen choose a course based on the instructor teaching it. There are many other criteria or combinations of criteria used by students.

Thus the self-selection process in this course tends to be multidimensional and non-systematic in nature. Therefore, although students in the seven sections chosen for this study were not randomly sampled, the self-selection process used by students is not likely to result in systematic differences among sections in relation to the variables relevant to this study.

Of the 255 college freshman students enrolled in these sections, 247 took both the midterm (Test I) and the final examination (Test II). Both tests contained the alternate-choice and true-false questions. Table 10 shows the number of students who took Form A and Form B of each test. To achieve a balanced research design, there was one student randomly eliminated from the group who took Form B of Test I and Form A of Test II.

Materials. The examination items used in this study were drawn from an item pool of approximately 1600 questions. All items were written to test the knowledge, understanding, and application of scientific principles and methodology of biological science and to test philosophy, historical perspectives, and social implications of the biological sciences. Approximately 400 of these items were applicable to the genetics and human reproduction emphasis of the course. All items in the item pool were developed by faculty teaching natural science courses at Michigan State University. Some items were written

Table 10

Number of Students Who Took Form A and Form B of Tests I and II

	Test II	
Test I	Form A	Form B
Form A	55	68
Form B	68[a]	55

[a]One student was randomly eliminated from this group to achieve a balanced design.

25 years ago, others were written the term preceding this study. (Some of the items have been used innumberable times, some used only a few times.) All items in the item pool had been administered to students at least once and had shown some measure of difficulty and a positive discrimination. Item statistics were not available as they were neither stored with the items nor retained in files. For security reasons, none of the items in the natural science item pool are keyed with the correct answer.

Some of the items which are applicable to genetics and human reproduction are in matching form or in a key-type multiple-choice form (the student matches items from a four- or five-item key to subsequently listed statements). However, the majority of the 400 potential items are in four- or five-choice form. The following are typical of the items found in the item pool applicable to the area of genetics and human reproduction:

EXAMPLE ONE

The reason(s) why it required 150 years to develop and clarify the cell principle (theory) was

a) poor equipment for studying the cell.

b) poor communication between scientists.

c) "getting the idea" of a unified structure and function of the cell.

d) a and b above.

e) a, b, and c above.*

EXAMPLE TWO

Amino-acids are carried to ribosomes by

a) messenger RNA.

b) transfer RNA*.

c) proteins.

d) cytoplasmic DNA.

e) nuclear DNA.

EXAMPLE THREE

The most effective way of making human chromosome counts is

a) by examining the egg and sperm.

b) by utilizing cultured and treated red blood cells.

c) by utilizing cultured and treated white blood cells*.

d) by examining liver tissue.

e) none of these are effective.

EXAMPLE FOUR

In a DNA molecule, one strand contains the following sequence of bases: A-G-A-T-C. Which of the following represents the complementary sequence on the other strand?

a) C-C-T-A-G b) A-G-A-T-C c) T-C-T-A-G* d) U-C-U-A-G e) none of these

Item Conversion Procedures. The process of converting midterm examination items from multiple-choice to alernate-choice and true-false forms began shortly after the start of fall quarter. The item conversion process for the final examination began shortly after the administration of the midterm examination. The process used for both examinations was identical.

The senior instructor initially selected 65 items that had the potential of being included on Test I, and 100 items that had the potential of being included on Test II. For each item, the senior instructor indicated the correct answer and the distractor judged the most reasonable answer given by an uninformed student. Only an item in which the correct response included a single answer or element was considered for conversion to alternate-choice and true-false form. An item that required selection of the answer from a key (key-type multiple-choice item) or an item in which the correct response contained more than one answer or element (e.g., all the above, a and b above) was considered for use only in its original multiple-choice form. This was because the extensive revision required to convert such a complex item to alternate-choice or true-false form might change the content of the item. EXAMPLE ONE is such a complex question. A total of 26 items for the mid-term and 27 items for the final examination met the criteria for conversion to alternate-choice and true-false form.

In the alternate-choice form suggested by Ebel (1982), the two responses can be placed at the very end or at any other location within the stem. This freedom of placement permitted the multiple-choice items to be converted to alternate-choice form in one of three ways. First, if the stem of an item was a statement, then the stem was kept intact

and only the correct answer and the distractor indicated by the senior instructor were joined to the end of the stem. EXAMPLE TWO is such a question, and in alternate-choice form it reads:

Amino-acids are carried to ribosomes by a) messenger RNA

b) transfer RNA* .

Second, if the item was a statement and contained duplicate wordings in both the distractor and the correct answer, the duplicate wordings were made part of the stem and only the word or words that made the statement correct or incorrect became the responses. EXAMPLE THREE contains these duplicate wordings (underlined); in alternate-choice form it reads:

The most effective way of making human chromosome

counts is by utilizing cultured and treated

a) red b) white* blood cells.

Third, when the stem of the original item was in question form, some rewriting of the stem was necessary. EXAMPLE FOUR was in question form, and to covert it to alternate-choice form, the stem was rewritten as follows:

In a DNA molecule, one strand contains the following

sequence of bases: A-G-A-T-C. The complementary sequence on

the other strand is a) U-C-U-A-G b) T-C-T-A-G*

A table of random numbers was used to determine whether the correct answer or distractor would be listed first. Items were converted from alternate-choice to true-false form by randomly eliminating either the correct response or the distractor from the alternate-choice item. Research (Frisbie, 1971; Oosterhof & Glasnapp, 1974) has shown that true-false items in false form tend to have better discriminating ability than items in true form. Ebel (1979) suggests including more

than one-half, and in some cases up to 67 percent, false-form items in a true-false test. Of the 10 true-false items included in each form of each examination, it was decided to make 60 percent ($n = 6$) of the items false. Within this parameter, a table of random numbers was used to determine whether an item was to be true or false. The items in EXAMPLES TWO, THREE and FOUR in true-false form read as follows:

EXAMPLE TWO

Amino-acids are carried to ribosomes by messenger RNA. (F)

EXAMPLE THREE

The most effective way of making human chromosome counts is by utilizing cultured and treated white blood cells. (T)

EXAMPLE FOUR

In a DNA molecule, one strand contains the following sequence of bases: A-G-A-T-C. The complementary sequence on the other strand is T-C-T-A-G. (T)

Test Form Development. Two forms of the examination were developed for both the midterm and the final examinations. For each item on an examination, the alternate-choice version was included on one form and its content equivalent true-false version on the other form. Before this could be done, however, it was necessary to ensure that the items were, in fact, equivalent in content.

Each item in its original multiple-choice form, alternate-choice form, and true-false form was submitted to two measurement experts to be judged for equivalence of content. One of the experts was a professor in Educational Measurement at Michigan State University and a nationally recognized expert in his field. The second measurement expert was a recent Ph.D. graduate in Educational Psychology, who had extensive

experience in test item development. The judges were asked to compare the alternate-choice, the true-false, and the original multiple-choice item forms to each other to determine whether the questions measured the same item content. They also were asked to write any comments they had on the sheet containing the item. All items were judged equivalent in content. There were several items, however, that were identified by the judges as having construction flaws. It was decided to eliminate these flawed items from further consideration for use on the examination. From the remaining items, the 20 clearest, non-redundant items were selected for inclusion on the examination. Appendix A contains the 20 items for each test. Each item is in multiple-choice, alternate-choice, and true-false form.

It was anticipated that few students, if any, had encountered the alternate-choice item in any examinations prior to the Natural Science 115 midterm examination. Therefore, it was decided to alternate the unfamiliar alternate-choice form items with the more familiar true-false form items. To accomplish this, the 20 alternate-choice items were randomly assigned to groups of five items each. It must be noted that on Test I it was necessary to keep questions 1, 2, 3 together due to their relation to the descriptive paragraph preceding the questions; within this grouping, the items were randomly assigned a sequence. It was also necessary to keep questions 7 and 8 in order, as question 8 referred to question 7. These two items were randomly assigned as a pair in Test I sequence. The other items were randomly assigned a sequence within each group. The content equivalent true-false items were put in the same sequence as their alternate-choice counterparts. Two groups of alternate-choice items were then randomly assigned to Form

A, their true-false equivalent forms were assigned to Form B, and vice versa. These groups of alternate-choice and true-false items were arranged in two ways: on Form A, the arrangement was alternate-choice, true-false, alternate-choice, true-false; on Form B, the arrangement was true-false, alternate-choice, true-false, alternate-choice. Thus, alternate-choice items on one form had their content equivalent true-false item in the same respective position and sequence on the other form. The arrangement of the items on each form for Test I and Test II is shown in Table 11.

The sequenced items were returned to the senior instructor who added to them a subset of 22 multiple-choice items for Test I and 65 multiple-choice items for Test II. The multiple-choice items were arranged in two different sequences by the senior instructor; one sequence was randomly assigned to Form A and other to Form B (see Table 11). The four- and five-choice multiple-choice items were placed last on each form to reduce the advantage of guessing for students who might have felt rushed near the end of the examination, even though the examination was a power test.

Procedure. Prior to the administration of the midterm (Test I), the University Committee on Research Involving Human Subjects (UCRIHS) gave permission for the conduct of this research project (see Appendix B).

For both Test I and Test II, Forms A and B were arranged in alternating sequence and administered to students assembled in the large lecture hall regularly used for lectures and examinations. The purpose of sequencing the forms was to obtain randomly equivalent groups and to discourage students from copying the answers of those sitting nearby.

Table 11

Positions of Items on Test I and Test II

Test I		Test II	
Form A	Form B	Form A	Form B
AC_1	TF_1	AC_1	TF_1
•	•	•	•
AC_5	TF_5	AC_5	AC_5
TF_6	AC_6	AC_6	TF_6
•	•	•	•
TF_{10}	AC_{10}	AC_{10}	TF_{10}
AC_{11}	TF_{11}	AC_{11}	TF_{11}
•	•	•	•
AC_{15}	TF_{15}	AC_{15}	TF_{15}
TF_{16}	AC_{16}	AC_{16}	TF_{16}
•	•	•	•
TF_{20}	AC_{20}	AC_{20}	TF_{20}
MC_{21}	MC_{21}	MC_{21}	MC_{53}
•	•	•	•
MC_{25}	MC_{25}	•	•
MC_{26}	MC_{42}	•	•
MC_{27}	MC_{27}	•	•
MC_2	MC_{41}	•	•
MC_9	MC_{31}	•	MC_{85}
MC_{30}	MC_{32}	•	MC_{21}
MC_{31}	MC_{29}	•	•
MC_{32}	MC_{30}	•	•
MC_{33}	MC_{33}	•	•
•	•	•	•
MC_{40}	MC_{40}	•	•
MC_{41}	MC_{26}	•	•
MC_{42}	MC_{28}	MC_{85}	MC_{52}

Where: AC_i is the alternate-choice version of item i.
TF_i is the true-false version of item i.
MC_i is the same four- or five-choice multiple-choice item on Form A and Form B.

Page one of each test contained instructions for taking the examination and was the same for both Forms of Tests I and II (see Appendix C). Students were given one hour to complete the 42-item Test I, and two hours to complete the 85-item Test II. Because of this generous time allotment, both Tests I and II were considered power tests.

Students marked their responses to each question on machine-scorable answer sheets. The examinations and the machine-scorable answer sheets were collected by the two instructors, separated as to Form A and Form B, and machine scored by the Michigan State University Scoring Office. The midterm and the final examination scores were weighted and merged with other weighted test and quiz scores to form a total course score for each student. This total course score was used for grade assignment to students. All information identifying the student was removed by the Director of the Scoring Office before allowing this researcher access to the data. The data were entered into the CDC 6000 version of SPSS (Statistical Package for the Social Sciences) to rearrange the items in Form B to the same sequence on Form A, to convert to correct-answer = 1 and incorrect-answer = 0, and to perform the necessary data analysis.

Part II

Participants. The senior instructor of the Natural Science 115 sections used in this study and a departmental collaborator who was well versed in measurement methodology were asked to judge the better version of each alternate-choice and each true-false item. The senior instructor has been teaching this course for approximately 20 years; the collaborator, now retired, had taught the course for more than 30 years.

Materials. Only the alternate-choice and true-false items administered in Tests I and II were used in this part of the study. Each of the alternate-choice and true-false items on Form A and Form B of these tests was converted to two versions. The alternate-choice items were converted to correct-answer-first (AC_{ci}) and incorrect-answer-first (AC_{ic}) forms, and the true-false items to true form (TF_t) and false form (TF_f).

Procedure. The senior instructor and departmental collaborator were given a packet for each form of each test that contained both versions of each item listed on a separate page, an instruction sheet, and a recording sheet (see Appendix D). The judges were asked to choose, from among the two versions of each item, the one version that in their estimation would, simultaneously, most maximize the chances of a correct answer being made by a student who knows the material, and an incorrect answer being made by an uniformed student. The judges independently recorded their choices on the recording sheets and returned the packets to this investigator. The investigator then tallied the percent of agreement.

Part III

Sample. The students that participated in Part III of this study were from three sections of Natural Science 115, all of which were team taught by the same two professors in the spring quarter of the 1983-84 academic year. Only the senior instructor had been involved in Part I of this study.

The students who participated did so because the instructors agreed to include the two versions of the alternate-choice items on their final examination. The students participating consisted of the population of

all students taking Natural Science 115 in the spring quarter of the 1983-84 academic year. There was one student repeating the course who participated in Part I the preceding quarter. A total of 102 students took the final examination (Test III). There were 51 students who took Form A and 51 students who took Form B.

Materials. The items used in Part III consisted of a stratified random sample of 20 of the 38 alternate-choice items administered in Tests I and II. Two of the 40 items administered in these tests were excluded from Part III because the material they tested had not been taught the spring quarter. The discrimination index, D, was used as the stratifying variable, and the 38 alternate-choice examination items were arranged from lowest to highest discriminating ability, then grouped into strata. Each stratum contained a spread of .10 of these indices, with the exception of the lowest stratum which contained a spread of .13. A 20/38 or .53 proportional sample of items was selected from each stratum. The distribution of these strata, the number of items in each stratum, and the number selected is shown in Table 12.

Test Form Development. Two forms of the final examination (Test III) were developed. From the 20 items selected for inclusion, there were 10 AC_{ci} items randomly assigned to Form A, and 10 of their AC_{ic} version assigned to Form B. There were 10 AC_{ic} items randomly assigned to Form A and their AC_{ci} versions to Form B.

The respective versions of the items were arranged in the same sequence on each form of the examination, and returned to the senior instructor for the inclusion of 46 complementary four- and five-choice multiple-choice items selected from the item pool. These 46 items were arranged in two different sequencies; one sequence was assigned to

Table 12

Strata, Number in Each Stratum, and Results of Random Sample of Items

Discrimination Index	Number in Stratum f	Number Selected f
-.03 - .09	7	4
.10 - .19	12	6
.20 - .29	6	3
.30 - .39	6	3
.40 - .49	4	2
.50 - .59	3	2
Total	38	20

Note: Proportion of items selected from each stratum was 20/38 or .53.

Form A, the other to Form B. To prevent students sitting next to each other from working on the same alternate-choice item at the same time, the 20 alternate-choice items were embedded in the examination and were assigned as items 31 to 50 on each form of the examination. Thus, item 31 on Form A read:

In scientific methodology, prediction means nearly the same as

a) expectancy* b) interpretation of data.

On Form B, item 31 read:

In scientific methodology, prediction means nearly the same as

a) interpretation of data b) expectancy*.

The arrangement of all items on Forms A and B of the final examination (Test III) is presented on Table 13. The specific arrangement of the

alternate-choice item versions is shown in Table 14.

Procedure. Forms A and B of Test III were arranged in a regular sequence and administered to the students assembled in the large lecture hall regularly used for lectures and examinations. The forms were alternately ordered to obtain randomly equivalent groups and to discourage the copying of answers from those sitting on either side of the student. Oral instructions were given regarding the taking of the examination. Students were allowed two hours to complete the 66 items.

Students marked their responses to each question on machine-scorable answer sheets. The examinations and the answer sheets were collected by the instructors, separated as to Form A and Form B, and machine scored by the Michigan State University Scoring Office. The Test III score was weighted and merged with the other weighted quiz scores to produce a total weighted course score for each student. These total weighted scores were used for course grades. All information identifying a student was removed by the Director of the Scoring Office prior to allowing this researcher access to the data. The data were entered into the CDC 6000 version of SPSS to rearrange the items in Form B to the same sequence on Form A, to convert to correct-answer = 1 and incorrect-answer = 0, and for analysis of the data.

Table 13

Position of Items on Test III

Test III	
Form A	Form B
MC_1	MC_1
•	•
MC_{12}	MC_{28}
MC_{13}	MC_{51}
•	•
•	•
MC_{16}	MC_{54}
MC_{17}	MC_{29}
MC_{18}	MC_{30}
MC_{19}	MC_{55}
•	•
•	•
MC_{30}	MC_{66}
AC_{31}	AC_{31}
•	•
•	•
•	•
AC_{50}	A_{50}
MC_{51}	MC_{14}
MC_{52}	MC_{15}
MC_{53}	MC_1
•	•
•	•
MC_{65}	MC_{13}
MC_{66}	MC_{16}

Where: AC_i is one of the version of alternate-choice item i.
MC_i is the same four- or five-choice multiple-choice item on Form A and Form B.

Table 14

Position of Alternate-choice Versions on Test III

Test III	
Form A	Form B
AC_{31ci}	AC_{31ic}
AC_{32ci}	AC_{32ic}
AC_{33ic}	AC_{33ci}
AC_{34ci}	AC_{34ic}
AC_{35ic}	AC_{35ci}
AC_{36ic}	AC_{36ci}
AC_{37ci}	AC_{37ic}
AC_{38ic}	AC_{38ci}
AC_{39ci}	AC_{39ic}
AC_{40ci}	AC_{40ic}
AC_{41ci}	AC_{41ic}
AC_{42ci}	AC_{42ic}
AC_{43ci}	AC_{43ic}
AC_{44ic}	AC_{44ci}
AC_{45ci}	AC_{45ic}
AC_{46ic}	AC_{46ci}
AC_{47ic}	AC_{47ci}
AC_{48ic}	AC_{48ci}
AC_{49ic}	AC_{49ci}
AC_{50ic}	AC_{50ci}

Where: AC_{ci} has the correct answer listed first in alternate-choice item i.

AC_{ic} has the incorrect answer listed first in alternate-choice item i.

Hypotheses

The major hypotheses stated in Chapter I are restated in operational terms in this chapter:

Part I

H_1: The mean score of the alternate-choice items will be significantly greater than the mean score of the content equivalent true-false items when tested by the $\underline{F}$ test. Alpha was preset at .05.

H_2: The item-total point-biserial ($\underline{r}_{pbis}$) correlations of the alternate choice items will be significantly higher than the item-total point-biserial ($\underline{r}_{pbis}$) correlations of the content equivalent true-false items when tested by the sign test. Alpha was preset at .05.

H_3: The KR-20 reliability coefficient of the alternate-choice items will be greater than the KR-20 reliability coefficient of the true-false items. The Feldt Test for Equality of Two KR-20 Reliabilities was used to test this hypothesis. Alpha was preset at .05.

H_4: Criterion related validity of the alternate-choice items, as defined by the product moment correlation between the alternate-choice total scores and the criterion total weighted scores (with the alternate-choice scores removed from the criterion) will be greater than the criterion related validity of the true-false items as defined by the product moment correlation between true-false total scores and the criterion total weighted scores (with the true-false scores removed from the criterion). The correlations were transformed to $\underline{z}_r$

scores and the z-test statistic for two independent correlations (Glass & Stanley, 1970, pp. 313-314) was used to test the differences. Alpha was preset at .05.

Part II

H_1: Agreement between two departmental colleagues' judgments of the better version of an alternate-choice item will be no better than chance (50 percent). In this study, the best form was defined as the form that could best maximize both the choice of a correct answer from an informed student and the choice of an incorrect answer from an uninformed student.

H_2: Agreement between two departmental colleagues' judgments of the better version of a true-false item will be no better than chance (50 percent). Again, the best form was defined as the form that could best maximize both the choice of a correct answer from an informed student and the choice of an incorrect answer from an uninformed student.

Part III

H_1: The mean score of the AC_{ci} items will not differ from the mean score of the AC_{ic} items when tested by the paired t-test. Alpha was present at .05.

H_2: The item-total point-biserial (r_{pbis}) correlations of the AC_{ci} items will not differ from the item-total point-biserial (r_{pbis}) correlations of the AC_{ic} items when tested by the sign test. Alpha was preset at .05.

H_3: The KR-20 reliability coefficient of the AC_{ci} items will not differ from the KR-20 reliability coefficient of the AC_{ic} items. The Feldt Test for Equality of Two KR-20 Reliabilities was used to test this hypothesis.

H_4: The criterion related validity of the AC_{ci} items, as defined by the product moment correlation between the AC_{ci} total scores and the criterion total weighted scores (with the AC_{ci} scores removed for the criterion), will not differ from the criterion related validity of the AC_{ic} items, as defined by the product moment correlation between AC_{ic} total scores and the criterion total weighted scores (with the AC_{ic} scores removed from the criterion).

The correlations were transformed to $\underline{z}_r$ scores and the $\underline{z}$-test statistic for two independent correlations was used to test this hypothesis. Alpha was preset at .05.

Design and Analysis

The research design for Part I and Part II of this study was the same. Each part was a comparative study in which differences in difficulty, discrimination, reliability, and content-related validity were tested for two item forms.

Latin Square Design. The administration of Form A and Form B of Tests I, II, and III was designed to obtain two randomly divided groups, Group I and Group II. Two levels of two treatments, Item Form and Item Position, were administered to students. Group I received alternate-choice items in positions 1-5, 11-15 and true-false items in positions 6-10, 16-20; Group II received true-false items in positions 1-5, 11-15 and alternate-choice items in positions 6-10, 16-20. In this 2 X 2

Latin square design, no student received more than one treatment combination of Item Form and Item Position. The design and level of treatment for each group is shown in Table 15. When groups of subjects are used in each cell of the Latin square rather than single subjects (Lindquist, 1956 termed this a Type II Latin square design), the sums of squares are separated into two Between Subjects, and four Within Subjects components. These components and degrees of freedom are shown in Table 16.

Feldt Test for Equality of Two KR-20 Reliabilities. This test is an approximate statistical test derived by Feldt (1969) to test the hypothesis that two KR-20 reliabilities are equal. The assumptions made about a test, the examinees, and the scores of test 1 (these assumptions are the same for test 2) are:

(i) The N_1 examinees are assumed to be a random sample from the examinee population.

(ii) The k_1 units are assumed to be a random sample from the population of units in the domain represented by Test 1.

(iii) Over the entire population of examinees, the quantity t_{1i} is assumed normally distributed.

(iv) Over the entire examinees-by-units matrix for Test 1, the e_{1ij} are assumed homogenious in variance and are normally distributed, independently of each other and of t_{1i} (Feldt, 1969, p. 365)

Where:

N_1 = The number of examinees

k_1 = The number of items on the test 1

t_{1i} = The true score in deviation form, of the examinee, where $E(t_{1i} = 0)$

e_{1ij} = Measurement error, where $E(t_{1ij}) = 0$ for examinees and $E(e_{1ij})$ for item j for person i of Test 1

The statistic W is obtained using the following formula:

$$\underline{W} = \frac{1 - r_2}{1 - r_1}$$

where in Part I

r_1 = reliability of the alternate-choice items

r_2 = reliability of the true-false items

where in Part III

r_1 = reliability of the AC_{ci} items

r_2 = reliability of the AC_{ic} items

The statistic $\underline{W}$ is approximately distributed as a central $\underline{F}$ with $\underline{N}_1 - 1$, and $\underline{N}_2 - 1$ degrees of freedom only when $\underline{N}_1$ and $\underline{N}_2$ are greater than 100. When $\underline{N}_1$ or $\underline{N}_2$ is less than 100, the degrees of freedom must be adjusted by use of the following formulas:

$$\underline{v}_2 = \frac{2A}{A - 1} \qquad \underline{v}_1 = \frac{2A^2}{2B - AB - A^2}$$

Where:

$$A = \frac{df_4}{df_4 - 2} \cdot \frac{df_2}{df_2 - 2}$$

$$B = \frac{(df_1 + 2)(df_4)^2}{(df_4 - 2)(df_4 - 4)(df_1)} \cdot \frac{(df_3 + 2)(df_2)^2}{(df_2 - 2)(df_2 - 4)(df_3)}$$

and,

$df_1 = N_1 - 1$ $\qquad$ $df_3 = (N_2 - 1)(k_2 - 1)$

$df_4 = N_2 - 1$ $\qquad$ $df_2 = (N_1 - 1)(k_1 - 1)$

Table 15

Latin Square Research Design Used in Part I and Part III of this Study

		Treatment A_1	Treatment A_2
Treatment	B_1	G_I	G_{II}
	B_2	G_{II}	G_I

Where in Part I:

A_1 = AC item form

A_2 = TF item form

B_1 = Item position 1-5, 11-15

B_2 = Item position 6-10, 16-20

G_I = Group I_I

G_{II} = Group II_I

Where in Part III:

A_1 = AC_{ci} item form

A_2 = AC_{ic} item form

B_1 = Item position 1-5, 11-15

B_2 = Item position 6-10, 16-20

G_I = Group I_{III}

G_{II} = Group II_{III}

Table 16

Analysis of Total Sum of Squares for the 2 x 2 Latin Square Design

Source	df	Sum of Squares
Between - Subjects	$an - 1$	SS_S
AB	$a - 1$	$SS_{AB(b)} = SS_G$
error (b)	$a(n - 1)$	$SS_{error(b)} - SS_S - SS_G$
Within - Subjects	$an(a - 1)$	$SS_{wS} = SS_T - SS_S$
A	$a - 1$	SS_A
B	$a - 1$	SS_B
AB (w)[a]	$(a - 1)(a - 2)$	$SS_{AB(w)} = SS_{AB} - SS_G$
error (w)	$a(a - 1)(n - 1)$	$SS_{error(w)} = SS_{wS} - SS_A - SS_B - SS_{AB(w)}$
Total	$a^2n - 1$	SS_T

Note: From Design and Analysis of Experiments (p.278) by E.F. Lindquist, 1953, Boston: Houghton Mifflin.

[a]In a 2 x 2 design this term vanishes as a source of sums of squares because
df = (2 - 1)(2 - 2 = 0).

CHAPTER IV

RESULTS

This chapter is divided into three major sections. The results of Part I of this study are presented in the first section. In Part I, the alternate-choice and true-false items are compared on difficulty, discrimination, reliability, and criterion-related validity.

The results of Part II are presented in the second section. In this part of the study, the practability of judging the best form of the alternate-choice item (AC_{ci} or AC_{ic}) and the best form of the true-false item (TF_t or TF_f) are explored.

The results of Part III are presented in the third section. In this part of the study, the alternative-choice items with the correct answer listed first (AC_{ci}) and the incorrect answer listed first (AC_{ic}) are compared on difficulty, discrimination, reliability, and criterion-related validity.

Part I

During the initial exploration of the data, a repeated measures analysis of students' scores across Test I (midterm exam) and Test II (final exam) showed that the students performed differently on the two tests (see Table 17). As a result, it was decided to treat Test I and Test II as independent substudies within Part I. This posed no problem because there was no overlap in the material tested up to the midterm and the material tested after the midterm.

Table 17

Results of Repeated Measures Analysis of Tests I and II

Source	df	MS	F	p
Between-Subjects				
Constant	1			
error	245			
Within-Subjects				
Test (I & II)	1	13.21	5.71	.05
error	245	2.31		
Form (A & B)	1	370.75	154.54	.001
error	245	2.40		
Test x Form	1	16.13	6.29	.05
error	245	2.56		

The treatment of each test as a substudy required twice as many statistical analyses to test the hypotheses than originally planned. To adjust for Type I error, the alpha level stated in each hypothesis in Chapter III was increased from .05 to .025.

The first operational hypothesis to be tested stated:

H_1: The mean score of the alternate-choice items will be significantly greater than the mean score of their content equivalent true-false items when tested by the F-test.

The means (M), standard deviations (SD), and other item statistics for each Item Form for each Item Position within the Latin Square design are shown in Table 18. The results of the Latin square analysis for Test I are shown in Table 19; the results for Test II are shown Table 20.

For both Test I and Test II the mean score of the alternate-choice items was found to be significantly greater than the mean score of the true-false items. Given these results, Hypothesis 1 was supported.

Although item position was not of primary interest in this study, the results deserve discussion. In Test I, the mean score of items in Item Position 1-5,11-15 was significantly greater than the mean score of items in Item Position 6-10,16-20. The reverse was true for Test II, where the mean score of items in Item Position 6-10,16-20 was significantly greater than the mean of items in Item Position 1-5,11-15. These results suggest that there is an interaction effect between item position and item content on student performance on these two examinations.

Table 18

Latin Square Design and Item Statistics for Tests I and II

Item Position	Item Form AC	Item Form TF
TEST I		
Items 1-5, 11-15	Group I M = 7.19 SD = 1.71 $\overline{r}_{pbis}$ = .392 r_{tt} = .413 r = .592 r_{tt_a} = .876	Group II M = 6.03 SD = 1.70 $\overline{r}_{pbis}$ = .367 r_{tt} = .271 r = .443 r_{tt_a} = .788
Items 6-10, 16-20	Group II M = 6.76 SD = 1.42 $\overline{r}_{pbis}$ = .325 r_{tt} = .036 r = .425 r_{tt_a} = .272	Group I M = 4.94 SD = 1.57 $\overline{r}_{pbis}$ = .326 r_{tt} = .111 r = .278 r_{tt_a} = .555
TEST II		
Items 1-5, 11-15	Group I M = 6.52 SD = 1.68 $\overline{r}_{pbis}$ = .380 r_{tt} = .333 r = .545 r_{tt_a} = .859	Group II M = 5.48 SD = 1.42 $\overline{r}_{pbis}$ = .313 r_{tt} = .000 r = .272 r_{tt_a} = .000
Items 6-10, 16-20	Group II M = 7.37 SD = 1.59 $\overline{r}_{pbis}$ = .365 r_{tt} = .329 r = .388 r_{tt_a} = .831	Group I M = 5.48 SD = 1.91 $\overline{r}_{pbis}$ = .411 r_{tt} = .462 r = .628 r_{tt_a} = .896

NOTE: N = 123 for each cell
r_{tt} = KR20 reliability
r = correlation between each Item Form and the course grade criterion
r_{tt_a} = reliability adjusted to 100 items by the Spearman-Brown formula

Table 19
Latin Square Analysis of Test I (Midterm)

Source	df	MS	F	p
Between-Subjects	245	2.95	1.01	
AB (b)	1	13.34	4.58	n.s.
error (b)	244	2.91		
Within-Subjects	246	3.62	1.61	n.s.
A (Item Form)[a]	1	270.783	120.40	.001
B (Item Position)[b]	1	71.075	31.60	.001
AB (w)	1	0.00	0.00	n.s.
error (w)	244	2.249		
Total	491			

[a]Item Form AC vs TF

[b]Item Position 1-5,11-15 vs 6-10,16-20

Table 20

Latin Square Analysis of Test II (Final)

Source	df	MS	F	p
Between-Subjects	245	3.46	.99	
AB (b)	1	0.59	.17	n.s
error (b)	244	3.47		
Within-Subjects	246	2.89	1.44	
A (Item Form)[a]	1	116.09	57.62	.001
B (Item Position)[b]	1	104.73	51.98	.001
AB (w)	1	0.00	0.00	n.s.
error (w)	244	2.02		
Total	491			

[a]Item Form AC vs TF

[b]Item Position 1-5,11-15 vs 6-10,16-20

The second operational hypothesis to be tested stated:

H_2: The item-total point-biserial ($\underline{r}_{pbis}$) correlations of the alternate-choice items will be significantly higher than the item-total point-biserial ($\underline{r}_{pbis}$) correlations of the content equivalent true-false item when tested by the sign test.

A $\underline{r}_{pbis}$ coefficient was computed for each alternate-choice item and the total alternate-choice score of its respective test. Similarly a $\underline{r}_{pbis}$ coefficient was calculated for each true-false item and the total true-false score of its respective test. These values are found in Appendix E. The $\underline{r}_{pbis}$ of each of the 20 content-equivalent alternate-choice and true-false items was placed side by side and a sign test used to test for differences in discrimination ability of these two item forms. For Test I, $\underline{T} = 11$, $\underline{p} > .05$,, and for Test II, $\underline{T} = 13$, $\underline{p} > .05$, where $\underline{T}$ is the number of times the $\underline{r}_{pbis}$ of the alternate-choice item was greater than the $\underline{r}_{pbis}$ of the content-equivalent true-false item. Given these results, Hypothesis 2 was rejected. Note that the average $\underline{r}_{pbis}$ ($\overline{\underline{r}}_{pbis}$) for each cell in the Latin Square design is shown in Table 18.

The third operational hypothesis to be tested stated:

H_3: The KR-20 reliability coefficient of the alternate-choice items will be greater than the KR-20 reliability coefficient of the true-false items. The Feldt Test for Equality of Two KR-20 Reliabilities was used to test this hypothesis.

The KR-20 reliabilities ($\underline{r}_{tt}$) of the 5 alternate-choice items and the 5 true-false items were computed for each cell in the Latin square design. These KR-20 reliability coefficients are shown in Table 18. The reliabilities of these content equivalent item forms were then

tested for equality by use of the Feldt Test for Equality of Two KR-20 Reliabilities (Feldt, 1969). In Test I, for Item Position 1-5,11-15, $\underline{W}$ (123,123) = 1.24, $\underline{p} > .05$; and for Item Position 6-10,16-20, $\underline{W}(123,123) = .923$, $\underline{p} > .05$. Thus, the magnitude of the reliabilities for Test I were found to be equal.

For Test II, the reliability of the alternate-choice items in Item Position 1-5,11-15 was shown to be significantly greater, $\underline{W}$ (123,123) = 1.499, $\underline{p} < .025$, than the reliability of the content-equivalent true-false items. The magnitude of the reliabilities of the content-equivalent alternate-choice and true-false items in Item Position 6-10,16-20 were found to be equal, $\underline{W}(123,123) = .802$, $\underline{p} > .05$. Given these results for Test I and Test II, Hypothesis 3 is only partially supported.

The last operational hypothesis to be tested stated:

H_4: The criterion-related validity of the alternate-choice items, as defined by the product moment correlation between the alternate-choice total scores and the criterion related weighted score, will be greater than the criterion related validity of the true-false items as defined by the product moment correlation between true-false total scores and the criterion total weighted score. The correlations were transformed to $\underline{z}_r$ scores and the $\underline{z}$-test statistic for two independent correlations was used to test for differences.

The course grade for each student was based on a weighted accumulated score of all quizzes and tests. This accumulated score was adjusted by removing the weighted scores of all alternate-choice and true-false items. Correlations were then computed between the total

alternate-choice score for each Item Position and the adjusted score, and the total true-false score for each Item Position and the adjusted score. These correlation coefficients (r) are shown in Table 18.

Each correlation was transformed to a z_r score and a z-test for two independent samples was performed for content-equivalent items in each Item Position. For Test I, and items in Item Position 1-5,11-15, $z = 1.55$, $p < .067$; and for items in Item Position 6-10,16-20, $z = 1.318$, $p < .097$. In both cases, the alternate-choice and the true-false item correlations with the course grade criterion were not significantly different at the .05 level.

For Test II, the criterion-related correlation of the alternate-choice items in Item Position 1-5,11-15 was significantly greater than that of the true-false items, $z = 2.56$, $p < .006$. The criterion-related correlation of the true-false items in Item Position 6-10,16-20 was significantly greater than that of the alternate-choice items, $z = -2.56$, $p < .006$. Given the mixed results for Test I and Test II, Hypothesis 4 was only partially supported.

Part II

The operational hypotheses in Part II were both stated in null form:

H_1: Agreement between two departmental colleagues' judgements as to the better version of an alternate-choice item will be better than chance (50 percent).

H_2: Agreement between two departmental colleagues' judgements as to the better version of a true-false item will be no better than chance (50 percent).

The two instructors who judged the better version of each alternate-choice item (AC_{ci} and AC_{ic}) and the better version of the true-false item (TF_t and TF_f), expressed a great deal of frustration concerning the completion of the task. Both found judging the alternate-choice item versions more exasperating than judging the true-false versions. One judge noted on the recording sheet that: "I have completed this task but I have no confidence that, confronted with the same task again, I would make the same choice". The other judge stated verbally that the task was a piece of "nonsense" and that he was certain that he would not be able to produce the same judgments if he were to redo the task--which he stated that he would not do.

The percent of agreement between the judges in their choice of the best alternate-choice items was 55 percent, only 5 percent greater than that expected by chance. There was also a 55 percent agreement for the true-false items. It can be concluded from these results that the agreement of two departmental colleagues' judgements as to the better item form of each alternate-choice and each true-false item is no better than chance. Hypothesis 1 and Hypothesis 2 were accepted.

PART III

The same statistical tests used in Part I of this study were used in Part III. The alpha level for all hypotheses was set at .05. All hypotheses in Part III were stated in null form.

The first operational hypothesis to be tested stated:

H_1: The mean score of the AC_{ci} items will not differ from the mean score of the AC_{ic} items when tested by the Latin square F-test.

The means, standard deviations, and other item statistics for each

alternate-choice Item Form for each Item Position within the Latin Square design are shown in Table 21; the results of the F-tests are shown in Table 22. The mean difficulty of the AC_{ci} and AC_{ic} items were found to be equal, $F(1,202) = 3.03$, $p > .05$. Given these results, Hypothesis 1 could not be rejected.

Although item position was not of primary interest in this part of the study, the results deserve discussion. The mean score of items in Item Position 6-10,16-20 were found to be significantly higher than the mean of items in Item Position 1-5,11-15. These results suggest, as they did in Part I, that there is an interaction effect between item position and item content on student performance on this examination.

The second operational hypothesis to be tested stated:

H_2: The item-total point-biserial (r_{pbis}) correlations of the AC_{ci} items will not differ from the item-total point-biserial (r_{pbis}) correlations of the AC_{ci} items when tested by the sign test.

A point-biserial correlation (r_{pbis}) coefficient was computed for the total alternate-choice score and each AC_{ci} and each AC_{ic} item on the test. These values are found in Appendix E. The r_{pbis} of each of these 20 content-equivalent AC_{ci} and AC_{ic} items was placed side by side and a sign test used to test for differences in discrimination.

Table 21

Latin Square Design and Item Statistics for Part III

Item Position	Item Forms ACCI	ACIC
A	Group I M = 6.43 SD = 1.75 $\overline{r}_{pbis}$ = .333 r_{tt} = .375 r = .692	Group II M = 6.94 SD = 1.67 $\overline{r}_{pbis}$ = .337 r_{tt} = .271 r = .542
B	Group II M = 7.45 SD = 1.62 $\overline{r}_{pbis}$ = .328 r_{tt} = .402 r = .649	Group I M = 7.43 SD = 1.66 $\overline{r}_{pbis}$ = .357 r_{tt} = .453 r = .548

Note: N = 204 for each cell
r_{tt} = KR20 reliablity
r = correlation between each Item Form and course grade criterion

Table 22

Latin Square Analysis of Part III

Source	df	MS	F	p
Between-Subjects	101	3.59	0.00	
AB (b)	1	3.57	.99	n.s.
error (b)	100	3.59		
Within-Subjects	102	2.31	2.28	
A (Item Form)[a]	1	3.06	3.03	n.s.
B (Item position)	1	29.06	28.77	.001
AB (w)	1	0.00	0.00	n.s.
error (w)	202	1.01		
Total	203			

[a]Item Form AC_{ci} vs AC_{ic}

The results of the sign test was $\underline{T} = 12$, $\underline{p} > .05$, where $\underline{T}$ is the number of times the $\underline{r}_{pbis}$ of the AC_{ic} item was greater than the $\underline{r}_{pbis}$ of the content-equivalent AC_{ci} item. Given these results, it can be concluded that there is no significant difference between the item discriminations of the AC_{ci} and AC_{ic} items. Hypothesis 2 could not be rejected. Note that the average $\underline{r}_{pbis}$ ($\underline{r}_{pbis}$) for each cell in the Latin square design is shown in Table 21.

The third operational hypothesis to be tested stated:

H_3: KR-20 reliability coefficient of the AC_{ci} items will not differ from the KR-20 reliability coefficient of the AC_{ic} items. The Feldt Test for Equality of two KR-20 Reliabilities was used to test this hypothesis.

The KR-20 reliability coefficients ($\underline{r}_{tt}$) for the 5 AC_{ci} items and the 5 AC_{ic} items were computed for each cell in the Latin Square design. These KR-20 reliability coefficients are shown in Table 21. The reliabilities of the content equivalent items for each Item Position were tested for equality by use of the Feldt Test for Equality of Two KR-20 Reliabilities (Feldt, 1969). For Item Position 1-5,11-15, $\underline{W}(44,45) = 1.06$, $\underline{p} > .05$; and for Item Position 6-10,16-20, $\underline{W}(44,45) = 1.09$, $\underline{p} > .05$. Given these results, it can be concluded that there is no significant difference between the reliabilities of these two item forms. Hypothesis 3 could not be rejected.

The last operational hypothesis to be tested stated:

H_4: The criterion related validity of the AC_{ci} items, as defined by the product moment correlation between the AC_{ci} total scores and the criterion total weighted scores (with the scores of all AC_{ci} and AC_{ic} items removed), will not differ from the criterion related validity of the Ac_{ic} items, as defined by the product moment correlation between Ac_{ic} total scores and the criterion total weighted scores (with the scores of all AC_{ci} and AC_{ic} items removed). The correlations were transfered to $\underline{z}_r$ scores and the $\underline{z}$-test statistics for two independent correlations was used to test this hypothesis.

The weighted accumulated score for each student was adjusted by removing the weighted scores of all AC_{ci} and AC_{ic} items. For each Item Position, correlations were computed between the total AC_{ci} score and the adjusted score, and between the total AC_{ic} score and the adjusted score. These correlation coefficients ($\underline{r}$) are shown in Table 21.

Differences in the correlations of AC_{ci} and AC_{ic} items were tested by transforming them to $\underline{z}_r$ scores and, for each Item Position, performing a $\underline{z}$ test for two independent samples. For items in Position 1-5,11-15, $\underline{z} = 1.17$, $\underline{p} > .05$, and for items in Item Position 6-10,16-20, $\underline{z} = .78$, $\underline{p} > .05$. Given these results it can be concluded that there are no significant differences between the criterion related validities. Hypothesis 4 could not be rejected.

CHAPTER V

SUMMARY AND CONCLUSIONS

Ebel (1982) proposed the use of the "alternate-choice" item as a replacement for the true-false item. The alternate-choice item he described was a modified two-choice multiple-choice item in which the two responses were included within the stem of the item.

A search of the literature showed, other than for Ebel's study, a dearth of recent research studies on the comparison of two-choice or alternate-choice and true-false items. The results of two studies conducted in the 1920's showed conflicting outcomes for the reliabilities of the two-choice multiple-choice and true-false items, but results that were consistent with those of Ebel's concerning the less difficult nature and greater predictive validity of the two-choice multiple-choice/alternate-choice item.

The purpose of the present study was three-fold: 1) to compare the difficulty level, discrimination level, reliability, and criterion-related validity of the alternate-choice item form and the content-equivalent true-false form; 2) to investigate the practicability of judging whether the alternate-choice version AC_{ci} or AC_{ic} is the better form of the item, and whether the true-false version TF_t or TF_f is the better form of this item; and 3) to examine the effects of the alternate-choice item with the correct answer given first (AC_{ci}) and the distractor given first (AC_{ic}) on difficulty, discrimination, reliability, and criterion-related validity.

This study was conducted in three parts. Each part corresponded to each purpose of this study. In this chapter, a summary, a discussion of the findings, and conclusions are presented for each part. Limitations of the study, and suggestions for future research are presented last.

Part I

The instruments used in Part I of this study were a midterm (Test I) and a final examination (Test II). Both tests were administered to lower division college students in a natural science course that emphasized genetics and reproduction. Each test consisted of Form A and Form B. Each form of Test I and each form of Test II contained 10 alternate-choice items, 10 true-false items, and respectively, 22 and 65 four- and five-choice multiple-choice items or key-type multiple-choice items. The alternate-choice and true-false items on Form B were content equivalent of the true-false and alternate-choice items on Form A.

From a pool of 400 items, the senior instructor of the course selected 65 items for conversion to alternate-choice and true-false items for Test I and 100 items for conversion for Test II. For each item he indicated the correct answer and the distractor he judged the most reasonable answer given by an uninformed student.

Items were converted from alternate-choice to true-false form by randomly eliminating either the correct response or the distractor, within the parameter than 60 percent of the items would be false. Each item in its multiple-choice, its alternate-choice, and its true-false form was submitted to two measurement experts to be judged for equivalence of content. All items were judged equivalent.

Forms A and B were distributed to students in a regular alternating sequence to discourage the copying of answers and to obtain randomly equivalent groups. There were 247 students who took both Test I and Test II. There was one student randomly eliminated to produce a balanced Latin square design. Both tests were considered power tests.

During the initial exploration of the data, a repeated measures analysis indicated that students performed differently on Test I and Test II. As a result it was decided to treat Test I and Test II as independent substudies within Part I.

The results of the Latin square analysis *F*-tests indicated that for both Test I and Test II, the alternate-choice items were significantly easier to answer than the content-equivalent true-false items. These results are consistent with those found by Ruch and Stoddard (1925), Charles (1926), and Ebel (1982).

The consistent findings that the alternate-choice item is less difficult than the true-false item can most likely be attributed to the additional piece of information present in the alternate-choice item that is not present in the true-false item. This information probably acts as a focusing mechanism to assist the student in determining more precisely what information the item writer is seeking.

There was an additional finding of an interaction effect of item position and item content on item difficulty for both Test I and Test II. This finding suggests a need for control of these variables in future research of this type.

The results of the sign test comparing the $\underline{r}_{pbis}$ correlations of the content-equivalent alternate-choice items with those of the true-false items for both Test I and Test II indicated that there were no differences in the discrimination ability of the two item forms. In his study, Ebel (1982) reported that the alternate-choice items had higher discrimination ability than the true-false items. One reason for these differing results may be that Ebel did not control for content-equivalence of items in his study, therefore the alternate-choice items in this study may have been easier in content than his true-false items. It also must be noted that Ebel did not conduct statistical tests on his discrimination data. It is possible that the difference between the $\underline{D}$ values of .28 for the true-false items and of .30 for the alternate-choice items were not statistically significant. There is also a probability that, in this study, the error variance introduced by guessing or item ambiguity may have obscured any real differences in discrimination of these item forms. The examination of the means of the true-false items (see Table 18) show two of the four means to be just slightly above the guessing level (5.48 and 5.48), and one mean to be below the guessing level (4.94).

When the reliabilities of the alternate-choice and true-false items in each item position in each test was tested for equality by the Feldt Test for Equality of Two KR-20 Reliabilities, only the alternate-choice and true-false items in Item Position 1-5,11-15 in Test II were found to differ significantly. The reliability for the alternate-choice items in this item position was .333, the reliability for the content equivalent true-false items was .000.

To more directly compare the magnitude of the reliabilities of this study with those found in previous studies, the reliabilities adjusted to 100 items by the Spearman-Brown formula reported in Tables 1,2,3 and 18 are shown in Table 23. The average reliability of the alternate-choice items in this study compare favorably to those found by Ruch and Stoddard (1925), Charles (1926), and Ebel (1982).

However, this is much less the case for the true-false items. If the low reliabilities of these items are due to ambiguity and/or guessing, then there should be an increase in error variance but not in the true variance of the test, and it should follow that reliability would be reduced. Given the means that reflect near guessing levels for these items, this appears to be the case in this study.

The last comparison made was of the criterion-related validity of the alternate-choice and true-false items. The criterion-related validity was defined as the Pearson product moment correlation between the alternate-choice total score and the total weighted score for the course. The student's final grade was based on this total weighted score.

For Test I, both the alternate-choice and the true-false scores were found to be equally correlated with the course grade criterion. For Test II, the results were mixed. The alternate-choice items in Item Position 1-5,11-15 were more highly correlated with the criterion than the true-false items. This result is not surprising given the zero reliability of the true-false items in this item position. For items in Item Position 6-10,16-20, the true-false items were more highly correlated with the criterion than were the alternate-choice items.

Table 23

Reliabilities adjusted to 100 items by the Spearman-Brown formula

Test	AC 100 Item $\underline{r}_{tt}$	TF 100 Item $\underline{r}_{tt}$
Test I		
Item Position 1-5,11-15	.876	.788
Item Position 6-10,16-20	.272	.555
Test II		
Item Position 1-5,11-15	.859	.000
Item Position 6-10,16-20	.831	.896
Ebel (1982)	.890	.780
Charles (1926)	.646	.751
Ruch and Stoddard (1925)	.749	.714

The conclusions drawn from the results in Part I of this study were:

1) The alternate-choice item form was less difficult than the true-false item form.
2) There was evidence of an interaction effect of item position and item content on item difficulty.
3) The alternate-choice item form and the true-false item form do not differ in discrimination ability, and do not consistently differ in reliability or in their relationship to the final score for the course upon which grades are based.

Part II

Part II of this study was concerned with the practicability of judging the better item version of the alternate-choice and true-false item, and with the amount of agreement between judges on choosing the better version. The senior instructor of the natural science course used in this study and a departmental collaborator, both of whom taught the natural science course for many years, were asked to be the judges.

Each alternate-choice item used in Test I and Test II was converted to two version, one with the correct answer presented first (AC_{ci}), and the other with the incorrect answer presented first (AC_{ic}). Each true-false item used on these two tests were also converted to two versions, one was the true version (TF_t) and the other was the false version (TF_f) of the item.

Each judge was asked to choose the better version of each item; that is, the version that would, in his estimation, simultaneously maximize the chances of a correct answer being made by a student who knows the material, and an incorrect answer being made by an uninformed

student. The percent of agreement between the judges as to the better version of each alternate-choice item was 55 percent, and the agreement as to the better version of each true-false item was also 55 percent. From the frustrations expressed by the judges, it is unlikely that a test writer would seriously attempt a task similar to this one. The results in Part III of this study indicated that the two alternate-choice versions functioned very much the same on a test. The resultant lack of agreement between the judges as to the better form suggest that the use of a table of random numbers to choose the better version could produce a timely written test while keeping the test writer more even-tempered throughout the test development process. It should be noted, however, that because the false version of the true-false item is a better discriminating item than the true version, it is important for the practitioner to consider including more false than true versions of items on a true-false test.

Part III

The instrument used in this part of the study was a final examination (Test III) that was administered to lower division college students in a natural science course that emphasized genetics and reproduction. This course was taught by the same senior instructor who taught the natural science course in Part I. Test III contained 20 alternate-choice items randomly selected from those in Tests I and II, plus 22 multiple-choice or key-type items that were selected from the item pool. There were 10 AC_{ci} items randomly assigned to Form A, and 10 of their AC_{ic} content-equivalent version assigned to Form B. There were 10 AC_{ic} items randomly assigned to Form A and their content-equivalent AC_{ci} versions to form B.

These two forms were administered to 102 students using the same procedure as that used in Part I. Students had two hours to complete Test III (66 items), therefore Test III was also considered a power test. A Latin square design was used for data analysis.

The results of each statistical test indicated that no differences existed between these two item versions on difficulty, reliability, criterion-related validity, or discrimination. Thus, unlike the false version of the true-false item form, which is more discriminating than the true version, the AC_{ic} item version does not discriminate better than the AC_{ci} item version. As in Part I of this study, a significant interaction effect of item position and item content on item difficulty was found for these items.

The conclusions drawn from the results in Part III of this study were:

1) The alternate-choice item with the correct answer presented first and the alternate-choice item with the incorrect answer presented first did not differ in difficulty, discrimination, reliability, or criterion-related validity.
2) There was evidence that an interaction effect on item difficulty exists between item position and item content.

Limitations of the Study

The judgmental method was not used to convert items from alternate-choice form to true-false form. Instead, a table of random numbers was used to decide whether the correct answer or distractor was to be eliminated form the alternate-choice item. Given the results of the item form judgment task it is retrospectively doubted that the performance of the true-false item forms would have been different if

the judgmental method had been used. However, it is possible that this random conversion method may have affected the true-false item performance in this study. Further research which would include a random conversion procedure for developing true-false item from multiple-choice type item forms would be helpful in clarifying this issue.

Suggestions for Further Research

The following recommendations for further research on item forms are made:

1. That there be replication of Part I using items that are less difficult in nature
2. That there be a replication of Part III of this study to test for systematic differences in the AC_{ci} and AC_{ic} item versions.
3. That future investigations involving the conversion of item forms from a multiple-choice type form to true-false form should include a random selection method to determine the true or false form of the item.
4. That future investigations take into account the element of guessing which was not examined this this study. Guessing may be less on the alternate-choice item forms than on the true-false item form because of the additional piece of information supplied in the alternate-choice item. The extent of guessing on these two item forms needs to be studied using a three parameter latent trait analysis model.

5. That an investigation be made into the efficiency of Ebel's alternate-choice item form. Because of its compactness, this form may have greater efficiency than the two-choice multiple-

APPENDICES

APPENDIX A

MULTIPLE-CHOICE, ALTERNATE-CHOICE, AND TRUE-FALSE FORMS OF EXPERIMENTAL ITEMS

Multiple-choice, Alternate-choice, and True-false Forms of Experimental Items

MIDTERM EXAM

TEST 1

The next three items are based on the following information: In snapdragons tallness (T) is dominant over dwarfness (t), while red flower color is due to a gene (R) and white to its allele (r). The heterozygous condition results in pink flower color. A dwarf homozygous red snapdragon is crossed with a plant homozygous for tallness and white flowers.

1. MC form
What is the genotype and phenotype of the F_1's?

A) ttRr, dwarf and pink
B) ttrr, dwarf and white
C) TtRr, tall and red
* D) TrRr, tall and pink
E) None of these

1. AC form
The genotype and phenotype of the F_1's are a) TtRr, tall and pink b) TtRr, tall and red.

1. TF form
The genotype and phenotype of the F_1's are TtRr, tall and pink.

2. MC form
If two plants of the genotypes ttRr and TtRR are crossed and no mutations occur, what are the chances that they will produce a dwarf white plant?

A) 1/2
B) 1/4
C) 3/16
D) 1/16
* E) 0

2. AC form
If two plants of the genotypes ttRr and TtRR are crossed and no mutations occur, the chances are a) 0 b) 1/16 that they will produce a dwarf white plant.

2. TF form
If two plants of the genotypes ttRr and TtRR are crossed and no mutations occur, the chances are 1/16 that they will produce a dwarf white plant.

3. MC form
A plant which is heterozygous for tallness and red flowers is self-pollinated. What is the probability that the offspring will be short and white?

A) 9/16
B) 3/16
C) 3/9
* D) 1/16
E) 0

3. AC form
A plant which is heterozygous for tallness and red flowers is self-pollinated. The probability is a) 0 b) 1/16 that the offspring will be short and white.

3. TF form
A plant which is heterozygous for tallness and red flowers is self-pollinated. The probability is 0 that the offspring will be short and white.

4. MC form
A son is born whose father is normal but whose grandfather, on his mother's side was hemophilic. What are the chances that he, too, would bear this trait?

A) 75%
* B) 50%
C) 10%
D) 1/16
E) 0%

4. AC form
A son is born whose father is normal but whose grandfather, on his mother's side was a hemophilic. The chances that he, too, would bear this trait is a) 50% b) 75%

4. TF form
A son is born whose father is normal but whose grandfather, on his mother's side was hemophilic. The chances that he, too, would bear this trait is 75%.

5. MC form
In a cross between individuals heterozygous for two traits, the expected number of homozygous recessive individuals is

A) 9/16
B) 1/2
C) 1/4
D) 3/16
* E) 1/16

5. AC form
In a cross between individuals heterozygous for two traits, the expected number of homozygous recessive individuals is
a) 1/16 b) 1/4

5. TF form
In a cross between individuals heterozygous for two traits, the expected number of homozygous recessive individuals is 1/16.

6. MC form
Dr. Corcos works with a little plant, Arabidopsis thaliana, whose chromosome number is 2n=10. In such a plant the number of possible combinations of paternal and maternal chromosomes is

A) 64
* B) 32
C) 16
D) 8
E) 4

6. AC form
Dr. Corcos works with a little plant, Arabidopsis thaliana, whose chromosome number is 2n=10. In such a plant the number of possible combinations of paternal and maternal chromosomes is a) 8
b) 32

6. TF form
Dr. Corcos works with a little plant, Arabidopsis thaliana, whose chromosome number is 2n=10. In such a plant the number of possible combinations of paternal and maternal chromosomes is 32.

7. MC form
A streak of white in an otherwise colored head of hair is known as white forelock. It is due to a dominant gene. If a woman with a white forelock marries a normal man and their first child is normal, her genotype is

A) AA
* B) Aa
C) aa

7. AC form
A streak of white in an otherwise colored head of hair is known as white forelock. It is due to a dominant gene. If a woman with a white forelock marries a normal man and their first child is normal, her genotype is a) aa b) Aa .

7. TF form
A streak of white in an otherwise colored head of hair is known as white forelock. It is due to a dominant gene. If a woman with a white forelock marries a normal man and their first child is normal, her genotype is aa.

8. MC form
What are the chances that the second child of the marriage above has a white forelock?

A) 3/4
* B) 1/2
C) 1/4
D) 0

8. AC form
The chances that the second child of the marriage above has a white forlock is a) 1/2 b) 1/4 .

8. TF form
The chances that the second child of the marriage above has a white forelock is 1/2.

9. MC form
Some organisms have sex chromosomes of the XO, XX type, in which males have one X chromosome, females two; in other organisms the female has two X chromosomes, the male an X and a Y; instill other organisms such as birds, the female has XY and the male XX. Some animals and plants can even be male in one situation and female in another, when conditions are favorable; others are hermaphroditic. All this would seem to indicate that sex determination is

A) ultimately a hereditary decision prescribed by the male.
B) by the sex chromosomes only.
C) wholly random and unpredictable in any case
* D) not always entirely determined by the karyotype.

9 AC form
Some organisms have sex chromosomes of the XO, XX type, in which males have one X chromosome, females two; in other organisms the female has two X chromosomes, the male an X and a Y; in still other organisms such as birds, the female has XY and the male XX. Some animals and plants can even be male in one situation and female in another, when conditions are favorable; others are hermaphroditic. All this would seem to indicate that sex determination is a) not always entirely determined by the karyotype b) by the sex chromosomes only.

9. TF form:
Some organisms have sex chromosomes of the XO, XX type, in which males have one X chromosome, females two; in other organisms the female has two X chromosomes, the male an X and a Y; in still other organisms such as birds, the female has XY and the male XX. Some animals and plants can even be male in one situation and female in another, when conditions are favorable; others are hermaphroditic. All this would seem to indicate that sex determination is by the sex chromosomes only.

10. MC form
Few genes are Y linked. The reason for this is probably that

A) the Y chromosome is largely homologous with the X.
B) both sexes possess the Y chromosomes.
C) both sexes possess two X chromosomes.
* D) the Y chromosome occurs only in one sex and is small.

10. AC form
Few genes are Y linked. The reason for this is probably that the Y chromosome a) is largely homologous with the X b) occurs only in one sex and is small.

10. TF form
Few genes are Y linked. The reason for this is probably that the Y chromosome is largely homologous with the X.

11. MC form
The "drumstick" chromosome often found in the female nuclei of white blood cells

* A) indicates the sex of the person involved.
B) represents an inactivated Y-chromosome.
C) could occur in a person with the XYY syndrome.
D) is characteristic of the cri-du-chat disorder
E) none of the above

11. AC form
The "drumstick" chromosome often found in the female nuclei of white blood cells a) represents an inactivated Y-chromosome b) indicates the sex of the person involved.

11. TF form
The "drumstick" chromosome often found in the female nuclei of white blood cells indicates the sex of the person involved.

12. MC form
In guinea pigs black is dominant over white. A cross between a heterozygous black and a white guinea pig would give a ratio of

A) about 3 blacks to 1 white
B) all black
* C) about 1 black to 1 white
D) about 3 whites to 1 black
E) about 1 black to 2 grey to 1 white

12. AC form
In guinea pigs black is dominant over white. A cross between a heterozygous black and a white guinea pig would give a ratio of a) all black b) about 1 black to 1 white.

12. TF form
In guinea pigs black is dominant over white. A cross between a heterozygous black and a white puinea pig would give a ratio of about 1 black to 1 white.

13. MC form
A form of Vitamin D resistant rickets, known a hypophatemia, is inherited as a sex-linked dominant trait. If a male with hypophatemia marries a normal female, which of the following predictions concerning the potential progeny would be true?

A) All their sons would inherit the disease.
* B) All their daughters would inherit the disease.
C) None of their sons would inherit the disease.
D) None of their daughters would inherit the disease.
E) Both b and c are true.

13. AC form
A form of Vitamin D resistant rickets, known a hypophatemia, is inherited as a sex-linked dominant trait. If a male with hypophatemia marries a normal female, all their a) sons b) daughters would inherit the disease.

13. TF form
A form of Vitamin D resistant rickets, known a hypophatemia, is inherited as a sex-linked dominant trait. If a male with hypophatemia marries a normal female, all their sons would inherit the disease.

14. MC form
When children do not express a trait unless at least one parent expresses it, it is an indication that the gene involved is

A) x-linked dominant
* B) autosomal dominant
C) autosomal recessive
D) polygenetically inherited
E) skipping a generation

14. AC form
When children do not express a trait unless at least one parent expresses it, it is an indication that the gene involved is a) skipping a generation b) autosomal dominant.

14. TF form
When children do not express a trait unless at least one parent expresses it, it is an indication that the gene involved is skipping a generation.

15. MC form
If, in testing a genetic hypothesis you found a Chi-square of zero, you should

A) reject the hypothesis.
B) accept the hypothesis.
* C) redo the experiment.
D) discard Mendelian genetics
E) discard the Chi-square method

15. AC form
If, in testing a genetic hypothesis you found a Chi-square of zero, you should a) accept the hypothesis b) redo the experiment.

15. TF form
If, in testing a genetic hypothesis you found a Chi-square of zero, you should redo the experiment.

16. MC form
If the somatic cells of a male were found to contain a Barr body in each of their nuclei, what would be the most likely genetic constitution of the individual?

A) XO
B) XX
C) XYY
* D) XXY
E) XXX

16. AC form
If the somatic cells of a male were found to contain a Barr body in each of their nuclei, the most likely genetic constitution of the individual is a) XYY b) XXY .

16. TF form
If the somatic cells of a male were found to contain a Barr body in each of their nuclei, the most likely genetic constitution of the individual is XYY.

17. MC form
The theory of inheritance during Mendel's time was known as "blending". If this theory were correct, the outcome of a cross between a black animal and a white animal would produce offspring of what color?

A) Black
B) White
C) Spotted
* D) Gray
E) Impossible to tell

17. AC form
The theory of inheritance during Mendel's time was known as "blending". If this theory were correct, the outcome of a cross between a black animal and a white animal would produce a) Spotted b) Gray offspring.

17. TF form
The current theory of inheritance during Mendel's time was known as "blending". If this theory were correct, the outcome of a cross between a black animal and a white animal would produce Spotted offspring.

18. MC form
The relative distance between linked genes may be determined by

A) cell fusion experiments
* B) crossing over frequencies
C) epistasis
D) pleiotropism
E) a and b above

18. AC form
The relative distance between linked genes may be determined by a) crossing over frequencies b) cell fusion experiments.

18. TF form
The relative distance between linked genes may be determined by cell fusion experiments.

19. MC form
How many possible combinations of gametes could be produced by one individual in a trihybrid cross?

A) 3
B) 6
* C) 8
D) 16
E) 64

19. AC form
The possible combinations of gametes that could be produced by one individual in a trihybrid cross is a) 6 b) 8 .

19. TF form
The possible combinations of gametes that could be produced by one individual in a trihybrid is 8.

20. MC form
If you flip 3 coins, the probability of getting 2 heads and 1 tail is

A) 1/6
B) 1/8
C) 3/9
* D) 3/8
E) none of these

20 AC form
If you flip 3 coins, the probability of getting 2 heads and 1 tail is a) 1/8 b) 3/8

20. TF form
If you flip 3 coins, the probability of getting 2 heads and 1 tail is 3/8.

FINAL EXAM QUESTIONS

TEST II

1. MC form
The main activity of science is to

A. observe nature.
* B. make and test theories.
C. debate issues with organized religion.
D. create machines which will improve human society.
E. support political systems.

1. AC form
The main activity of science is to a) observe nature
b) make and test theories .

1. TF form
The main activity of science is to make and test theories.

2. MC form
In scientific methodology, prediction means nearly the same as

A. interpretation of data.
B. generalization from empirical observation.
* C. expectancy.
D. experimentation.
E. none of the above.

2. AC form
In scientific methodology, prediction means nearly the same as
a) expectancy b) interpretation of data .

2. TF form
In scientific methodology, prediction means nearly the same as interpretation of data.

3. MC form
The lowest level of explanation is

A. a theory.
* B. an hypothesis.
C. a fact.
D. an assumption.

3. AC form
The lowest level of explanation is a) an hypothesis b) an assumption .

3. TF form
The lowest level of explanation is an assumption.

4. MC form
Which of Mendel's procedures differed from those of his predecessors and contributed most to his success?

A. He kept breeding records.
B. He observed distinct inherited traits.
C. He observed many characteristics for each trait.
* D. He quantitatively (statistically) analyzed his data.
E. He used one of the few organisms which can be grown in a laboratory.

4. AC form
Mendel's procedure that differed from those of his predecessors and contributed most to his success was that a) he observed distinct inherited traits b) he quantitatively (statistically) analayzed his data .

4. TF form
Mendel's procedure that differed from those of his predecessors and contributed most to his success was that he observed distinct inherited traits.

5. MC form
In radishes, long and round are alleles, as are red and white. In a cross between a long, red variety, and a round, white variety the F_1 is oval and purple. How many different phenotypes you would expect to find in the F_2?

A. 16
* B. 9
C. 4
D. 3
E. 2

5. AC form
In radishes, long and round are alleles, as are red and white. In a cross between a long, red variety, and a round, white variety the F_1 is oval and purple. The different phenotypes you would expect to find in the F_2 is a) 4 b) 9 .

5. TF form
In radishes, long and round are alleles, as are red and white. In a cross between a long, red variety, and a round, white variety the F_1 is oval and purple. The different phenotypes you would expect to find in the F_2 is 4.

6. MC form
A characteristic of a dominant trait is that

* A. the trait never skips a generation.
B. the genotype can be determined directly from the phenotype.
C. the phenotype cannot be read from the genotype.
D. the homozygote for the train can be distinguished from the heterozygote.
E. more than one above.

6. AC form
A characteristic of a dominant trait is that a) the trait never skips a generation b) the genotype can be determined directly from the phenotype .

6. TF form
A characteristic of a dominant trait is that the genotype can be determined directly from the phenotype.

7. MC form
Two black female mice are mated to a brown male. In several litters, Female I produced 9 blacks and 7 browns, Female II produced 57 blacks. Assuming black to be dominant over brown, what are the respective genotypes of the Female I, Female II, and the male?

* A. Bb, BB, bb
B. BB, Bb, bb
C. Bb, bb, BB
D. bb, Bb, BB
E. BB, BB, bb

7. AC form
Two black female mice are mated to a brown male. In several litters, Female I produced 9 blacks and 7 browns, Female II produced 57 blacks. Assuming black to be dominant over brown, the respective genotypes of the Female I, Female II, and the male are a) BB, Bb, bb b) Bb, BB, bb .

7. TF form
Two black female mice are mated to a brown male. In several litters, Female I produced 9 blacks and 7 browns, Female II produced 57 blacks. Assuming black to be dominant over brown, the respective genotypes of the Female I, Female II, and the male are Bb, BB, bb.

8. MC form
A woman who has Turner syndrome is found to have hemophilia; yet neither of her parents have the disease. She

A. got the defective gene from her father.
* B. got the defective gene from her mother.
C. could have gotten the defective gene from either parent.
D. could not have gotten the defective gene from either parent.
E. must be adopted since hemophilia is due to a dominant gene.

8. AC form
A woman who has Turner syndrome is found to have hemophilia; yet neither of her parents have the disease. She a) got the defective gene from her mother b) could have gotten the defective gene from either parent.

8. TF form
A woman who has Turner syndrome is found to have hemophilia; yet neither of her parents have the disease. She got the defective gene from her mother.

9. MC form
The extra Y chromosome of the XYY male was thought for some time to cause

A. stunted and stocky build in affected males.
B. above average intelligence.
C. above average strength.
D. sterility in such males.
* E. aggressive and antisocial behavior.

9. AC form
The extra Y chromosome of the XYY male was thought for some time to cause a) sterility in such males b) aggressive and antisocial behavior .

9. TF form
The extra Y chromosome of the XYY male was thought for some time to cause sterility in such males.

10. MC form
What process probably occurs during meiosis to produce an XXY individual?

A. Segregation
B. Crossing over
* C. Nondisjunction
D. Random assortment
E. None of these

10. AC form
The process that probably occurs during meiosis to produce an XXY individual is a) crossing over b) nondisjunction .

10. TF form
The process that probably occurs during meiosis to produce an XXY individual is crossing over.

11. MC form
Sex chromatin found in body cells and called Barr bodies have a relationship with the number of X chromosomes present in a given individual's body cells. If a given male had a sex chromosome composition of XXXXY, the number of Barr bodies observable in somatic tissue cells would be

A. five
B. four
* C. three
D. two
E. none of these

11. AC form
Sex chromatin found in body cells and called Barr bodies have a relationship with the number of X chromosomes present in a given individual's body cells. If a given male had sex chromosome composition of XXXXY, a) three b) four Barr bodies would be observable in somatic tissue cells.

11. TF form
Sex chromatin found in body cells and called Barr bodies have a relationship with the number of X chromosomes present in a given individual's body cells. If a given male had sex chromosome composition of XXXXY, three Barr bodies would be observable in somatic tissue cells.

12. MC form
More than one man was responsible for proposing the one-gene, one-enzyme hypothesis. Those responsible were

A. Watson and Crick.
B. Mendel and Morgan.
C. Lysenko and Lamarck.
* D. Beadle and Tatum.
E. None of these.

12. AC form
The men responsible for proposing the one-gene, one-enzyme hypothesis were a) Beadle and Tatum b) Watson and Crick.

12. TF form
The men responsible for proposing the one-gene, one-enzyme hypothesis were Watson and Crick.

13. MC form
When cells of certain bacteria are grown on glucose they do not produce beta-galactosidase (an enzyme which is important in breaking down lactose). However, when the same cells are placed in lactose they begin to make beta-galactosidase almost immediately. The results of this experiment support the hypothesis that

A. DNA is a genetic material.
B. genes are influenced by the environment
C. not all the genes are operative all the time.
* D. B and C
E. A, B, and C

13. AC form
When cells of certain bacteria are grown on glucose they do not produce beta-galactosidase (an enzyme which is important in breaking down lactose). However, when the same cells are placed in lactose they begin to make beta-galactosidase almost immediately. The results of this experiment support the hypothesis that a) genes are influenced by the environment b) DNA is the genetic material .

13. TF form
When cells of certain bacteria are grown on glucose they do not produce beta-galactosidase (an enzyme which is important in breaking down lactose). However, when the same cells are placed in lactose they begin to make beta-galactosidase almost immediately. The results of this experiment support the hypothesis that genes are influenced by the environment.

14. Watson-Crick base pairing requires that the adenine content of one strand of DNA equals the

A. thymine content of the complementary strand
B. thymine content of the same strand
C. adenine content of the complementary strand
D. uracil content of the complementary strand
E. guanine content of the complementary strand

14. Watson-Crick base pairing requires that the adenine content of one strand of DNA equals the thymine content of the a) complementary b) same strand.

14. Watson-Crick base pairing requires that the adenine content of one strand of DNA equals the thymine content of the same strand.

15. MC form
A nucleotide consists of either a purine or pyrimidine, a five-carbon sugar and a

A. carbohydrate.
B. amino acid.
C. peptide.
* D. phosphate group.
E. sulfate group.

15. AC form
A nucleotide consists of either a purine or pyrimidine, a five-carbon sugar and a) a phosphate group b) an amino acid .

15. TF form
A nucleotide consists of either a purine or pyrimidine, a five-carbon sugar and a phosphate group.

16. MC form
Amino-acids are carried to ribosomes by

A. messenger RNA.
* B. transfer RNA.
C. proteins.
D. cytoplasmic DNA.
E. nuclear DNA.

16. AC form
Amino-acids are carried to ribosomes by a) messenger RNA b) transfer RNA.

16. TF form
Amino-acids are carried to ribosomes by messenger RNA.

17. MC form
According to the genetic code, a gene responsible for the formation of a protein of 200 amino-acid subunits should have

A. 200 nucleotides.
B. 400 nucleotides.
* C. 600 nucleotides.
D. 800 nucleotides.
E. Dr Corcos and Dr Marinez, do you think we are math majors?

17. AC form
According to the genetic code, a gene responsible for the formation of a protein of 200 amino-acid subunits should have a) 600 b) 200 nucleotides .

17. TF form
According to the genetic code, a gene responsible for the formation of a protein of 200 amino-acid subunits should have 200 nucleotides.

18. MC form
In a DNA molecule, one strand contains the following sequence of bases A-G-A-T-C. Which of the following represents the complementary sequence on the other strand?

A. C-C-T-A-G
B. A-G-A-T-C
* C. T-C-T-A-G
D. U-C-U-A-G
E. None of these

18. AC form
In a DNA molecule, one strand contains the following sequence of bases A-G-A-T-C. The complementary sequence on the other strand is a) U-C-U-A-G b) T-C-T-A-G .

18. TF form
In a DNA molecule, one strand contains the following sequence of bases A-G-A-T-C. The complementary sequence on the other strand is T-C-T-A-G .

19. MC form
Mongoloid idiocy or Down's Syndrome is a consequence of abnormalities in:

A. sex-linked heredity
B. sex-influenced heredity
C. the number of sex chromosomes.
* D. the number of autosomal chromosomes.
E. None of these

19. AC form
Mongoloid idiocy or Down's Syndrome is a consequence of abnormalities in the number of a) sex b) autosomal chromosomes .

19. TF form
Mongoloid idiocy or Down's Syndrome is a consequence of abnormalities in the number of sex chromosomes.

20. MC Form
In recombinant DNA research, an alien gene is

A. treated with tRNA.
B. incorporated into a bacterial plasmid.
C. combined with a repressor substance.
D. mixed with histone proteins.
* E. injected into the host cell with a sex pilus.

20. AC form
In recombinant DNA research, an alien gene is a) incorporated into a bacterial plasmid b) injected into the host cell with a sex pilus.

20. In recombinant DNA research, an alien gene is incorporated into a bacterial plasmid.

APPENDIX B

UNIVERSITY COMMITTEE ON RESEARCH INVOLVING HUMAN SUBJECTS (UCRIHS) APPROVAL LETTER

MICHIGAN STATE UNIVERSITY

UNIVERSITY COMMITTEE ON RESEARCH INVOLVING
HUMAN SUBJECTS (UCRIHS)
238 ADMINISTRATION BUILDING
(517) 355-2186

EAST LANSING • MICHIGAN • 48824

December 12, 1983

Ms. Nancy A. Maihoff
424 Administration Building

Dear Ms. Maihoff:

Subject: Proposal Entitled, "A Comparison of Alternate-Choice and True-False Item Form Used in Classroom Examinations"

UCRIHS review of the above referenced project has now been completed. I am pleased to advise that the rights and welfare of the human subjects appear to be adequately protected and the Committee, therefore, approved this project at its meeting on December 5, 1983.

You are reminded that UCRIHS approval is valid for one calendar year. If you plan to continue this project beyond one year, please make provisions for obtaining appropriate UCRIHS approval prior to December 5, 1984.

Any changes in procedures involving human subjects must be reviewed by the UCRIHS prior to initiation of the change. UCRIHS must also be notified promptly of any problems (unexpected side effects, complaints, etc.) involving human subjects during the course of the work.

Thank you for bringing this project to our attention. If we can be of any future help, please do not hesitate to let us know.

Sincerely,

Henry E. Bredeck
Chairman, UCRIHS

HEB/jms

cc: Mehrens

APPENDIX C

EXAM INSTRUCTION SHEET

Exam Instruction Sheet

A. CORCOS
D. MARINEZ

NATURAL SCIENCE 115
FINAL EXAM

FALL 1983

FORM B

THERE ARE 85 ITEMS ON 15 PAGES OF THIS EXAMINATION. BE SURE YOU HAVE ALL OF THEM.

1. Check which form of the exam you have. The answer sheet should be BLUE if you have FORM A, and BROWN if you have FORM B. Raise your hand if this is not the case.

2. Print your last name and first initial, and your student number on the answer sheet, then darken the corresponding circles.

3. Each item is worth one point. Select the one best answer for each item.

4. Note that there are four-and five-choice multiple-choice items, alternate-choice (two-choice) items, and true-false items. Be sure to mark the appropriate space on the answer sheet by darkening the circle corresponding to the answer you select.

5. Do any calculations or scribbling on the last page of this exam booklet. If you mark answers on the test be sure you transfer the answers to the answer sheet.

6. If you make stray marks on the answer sheet or fail to erase completely the answer you wish to change, your response will not be counted.

7. Keep the marked part of your answer sheet covered at all times.

8. Your score on this examination will be the number of answers you marked correctly. Try to answer each item, but do not spend too much time on any one item.

9. If you have any questions, ask your instructor now, before starting the examination.

Good luck.

Happy Holiday Season
Feliz Navidad y Prospero Ano Nuevo
Joyeux Noel et Bonne Annee

APPENDIX D

ITEM JUDGMENT INSTRUCTION AND RECORDING SHEETS

Item Judgement Instruction and Recording Sheets

ALTERNATE-CHOICE ITEM FORM
JUDGMENT TASK

SECOND MIDTERM EXAM

Directions:

The following pages contain 20 alternate-choice items asked on a Natural Science 115 Midterm Examination. Each item is written in two forms: with the correct answer placed first, and as the incorrect answer placed first.

Your task is to choose the form for each item that will, in your judgment, simultaneously maximize the chance of a correct answer from a student knowing the material and the chance of an incorrect answer from an uninformed student.

The process suggested is to review all 20 items and select the form that you judge to be the best form of the item.

Transfer your choice of item form to the enclosed RECORDING SHEET, by writing the number of each item in either the CORRECT ANSWER FIRST column or in the INCORRECT ANSWER FIRST column.

Thank you very much.

RECORDING SHEET

ALTERNATE-CHOICE BEST ITEM FORM

CORRECT ANSWER FIRST	INCORRECT ANSWER FIRST
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

Note: Please add more lines to column if you need them.

ALTERNATE-CHOICE ITEM FORM
JUDGMENT TASK

FINAL EXAM

Directions:

The following pages contain 20 alternate-choice items asked on a Natural Science 115 Final Examination. Each item is written in two forms: with the correct answer placed first, and as the incorrect answer place first.

Your task is to choose the form for each item that will, in your judgment, simultaneously maximize the chance of a correct answer from a student knowing the material and the chance of an incorrect answer from an uninformed student.

The process suggested is to review all 20 items and select the form that you judge to be the best form of the item.

Transfer your choice of item form to the enclosed RECORDING SHEET, by writing the number of each item in either the CORRECT ANSWER FIRST column or in the INCORRECT ANSWER FIRST column.

Thank you very much.

RECORDING SHEET

ALTERNATE-CHOICE BEST ITEM FORM

CORRECT ANSWER FIRST	INCORRECT ANSWER FIRST
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

Note: Please add more lines to a column if you need them.

TRUE OR FALSE ITEM FORM
JUDGMENT TASK

SECOND MIDTERM EXAM

Directions:

The following pages contain 20 items asked on a Natural Science 115 Midterm Examination. Each item is written in two forms: as a true statement, and as a false statement.

Your task is to choose the form for each item that will, in your judgment, simultaneously maximize the chance of a correct answer from a student knowing the material and the chance of an incorrect answer from an uninformed student.

An additional constraint within which you must work is to ultimately identify only 8 true items and 12 false items from the total 20 items.

The process suggested is to review all 20 items and select those that you judge to be best as a true or as a false item. Then determine how many items you put in each category. If the number of items selected exceeds 8 true forms, then review again your choices so that only 8 are identified.

Transfer your choice of item form to the enclosed RECORDING SHEET, by writing the number of each item in either the TRUE FORM BEST column or in the FALSE FORM BEST column.

Thank you very much.

RECORDING SHEET

TRUE/FALSE BEST ITEM FORM

TRUE FORM BEST	FALSE FORM BEST
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
	_____
	_____
	_____
	_____

TRUE OR FALSE ITEM FORM
JUDGMENT TASK

FINAL EXAM

Directions:

The following pages contain 20 items asked on a Natural Science 115 Final Examination. Each item is written in two forms: as a true statement, and as a false statement.

Your task is to choose the form for each item that will, in your judgment, simultaneously maximize the chance of a correct answer from a student knowing the material and the chance of an incorrect answer from an uninformed student.

An additional constraint within which you must work is to ultimately identify only 8 true items and 12 false items from the total 20 items.

The process suggested is to review all 20 items and select those that you judge to be best as a true or as a false item. Then determine how many items you put in each category. If the number of items selected exceeds 8 true forms, then review again your choices so that only 8 are identified.

Transfer you choice of item form to the enclosed RECORDING SHEET, by writing the number of each item in either the TRUE FORM BEST column or in the FALSE FORM BEST column.

Thank you very much.

RECORDING SHEET

TRUE/FALSE BEST ITEM FORM

TRUE FORM BEST	FALSE FORM BEST
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
	_____
	_____
	_____
	_____

APPENDIX E
POINT-BISERIAL CORRELATIONS OF EXPERIMENTAL ITEMS- TESTS I, II, AND III

Table 24

Point-biserial Correlations of Experimental Items - Tests I, II, and III

TEST I			
Item	$\underline{r}$ pbis	Item	$\underline{r}$ pbis
Group I		Group II	
AC_1	.1778	TF_1	.4148
AC_2	.6591	TF_2	.5495
AC_3	.3655	TF_3	.1234
AC_4	.3579	TF_4	.3569
AC_5	.4628	TF_5	.3703
AC_{11}	.5232	TF_{11}	.3846
AC_{12}	.3955	TF_{12}	.4033
AC_{13}	.4336	TF_{13}	.3446
AC_{14}	.3534	TF_{14}	.3728
AC_{15}	.1940	TF_{15}	.3491
Group II		Group I	
TF_6	.1223	AC_6	.2918
TF_7	.3803	AC_7	.2585
TF_8	.2184	AC_8	.3494
TF_9	.2760	AC_9	.3058
TF_{10}	.4005	AC_{10}	.2905
TF_{16}	.2729	AC_{16}	.3857
TF_{17}	.2424	AC_{17}	.4188
TF_{18}	.3373	AC_{18}	.2817
TF_{19}	.4846	AC_{19}	.3692
TF_{20}	.5242	AC_{20}	.3016

TEST II			
Item	$\underline{r}$ pbis	Item	$\underline{r}$ pbis
Group I		Group II	
AC_1	.2896	TF_1	.3030
AC_2	.4132	TF_2	.4147
AC_3	.4232	TF_3	.3665
AC_4	.4726	TF_4	.3715
AC_5	.4027	TF_5	.2299
AC_{11}	.4039	TF_{11}	.3474
AC_{12}	.4256	TF_{12}	.3521
AC_{13}	.3615	TF_{13}	.1574
AC_{14}	.2906	TF_{14}	.3732
AC_{15}	.3171	TF_{15}	.2102
Group II		Group I	
TF_6	.5012	AC_6	.4474
TF_7	.3163	AC_7	.1915
TF_8	.4329	AC_8	.4661
TF_9	.4264	AC_9	.4226
TF_{10}	.4921	AC_{10}	.2986
TF_{16}	.4423	AC_{16}	.4791
TF_{17}	.2701	AC_{17}	.4745
TF_{18}	.3121	AC_{18}	.3484
TF_{19}	.4085	AC_{19}	.3402
TF_{20}	.5071	AC_{20}	.1820

TEST III

Item	$\underline{r}$ pbis	Item	$\underline{r}$ pbis
	Group I		Group II
AC_{31ci}	.0748	AC_{31ic}	.1535
AC_{32ci}	.4141	AC_{32ic}	.0957
AC_{34ci}	.2318	AC_{34ic}	.3236
AC_{37ci}	.2548	AC_{37ic}	.4973
AC_{39ci}	.1298	AC_{39ic}	.2551
AC_{40ci}	.1986	AC_{40ic}	.2499
AC_{41ci}	.6497	AC_{41ic}	.4733
AC_{42ci}	.3757	AC_{42ic}	.6174
AC_{43ci}	.5076	AC_{43ic}	.1933
AC_{45ci}	.4948	AC_{45ic}	.3934
	Group II		Group I
AC_{33ci}	.2549	AC_{33ic}	.4899
AC_{35ci}	.2911	AC_{35ic}	.4878
AC_{36ci}	.2985	AC_{36ic}	.4537
AC_{38ci}	.3834	AC_{38ic}	.3439
AC_{44ci}	.3706	AC_{44ic}	.4090
AC_{46ci}	.2357	AC_{46ic}	.2907
AC_{47ci}	.4721	AC_{47ic}	.2868
AC_{48ci}	.3827	AC_{48ic}	.2662
AC_{49ci}	.4778	AC_{49ic}	.3235
AC_{50ci}	.1170	AC_{50ic}	.2155

LIST OF REFERENCES

REFERENCES

Ahmann, J.S., & Glock, M.D. (1967). Evaluating pupil growth (3rd ed. revised). Boston: Allyn and Bacon.

Alpert, R., & Haber, R.N. (1960). Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 61, 207-215.

Brenner, M.H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. Journal of Applied Psychology, 48, 98-100.

Brown, F.G. (1970). Principles of educational and psychological testing. Hinsdale: Dryden.

Burmester, M.A., & Olson, L.A. (1966). Comparison of item statistics for items in multiple-choice and in alternative-response form. Science Education, 50, 467-470.

Carrier, N.A., & Jewell, D.O. (1966). Efficiency in measuring the effect of anxiety upon academic performance. Journal of Educational Psychology, 57, 23-26.

Carter, H. (1942). How reliable are the common measures of difficulty and validity of objective test items? Journal of Psychology, 13, 31-39.

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

Cochran, W.G., & Cox, G.M. (1957). Experimental designs (2nd ed.). New York: Wiley.

Cronbach, L.J. (1970). Essentials of Psychological Testing. New York: Harper and Row.

Davis, F.B. (1951). Item selection techniques. In E.F. Lindquist (Ed.), Educational measurement (pp. 266-328). Washington D.C.: American Council on Education.

Ebel, R.L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R.L. (1982). Proposed solutions to two problems of test construction. Journal of Educational Measurement, 19, 267-278.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373.

Frisbie, D.A. (1971). Comparative reliabilities and validities of true-false and multiple-choice tests. Unpublished doctoral dissertation, Michigan State University.

Gibbons, C.C. (1940). The predictive value of the most valid items of an examination. Journal of Educational Psychology, 31, 616-621.

Glass, G.V., & Stanley, J.C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N.E. (1965). Measurement and Evaluation in Teaching. New York: Macmillan.

Gronlund, N.E. (1977). Constructing achievement tests (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.

Huck, S.W., & Bowers, N.D. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 9, 105-111.

Klosner, N.C., & Gellman, E.K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. Educational and Psychological Measurement, 33, 413-418.

Lindquist, E.F. (1956). Design and analysis of experiments. Boston: Houghton Mifflin.

Loree, M.R. (1948). A study of a technique for improving tests. Unpublished doctoral dissertation, University of Chicago.

Mandler, G., & Sarason, S.B. (1952). A study of anxiety and learning. Journal of Abnormal and Social Psychology, 47, 166-173.

Marso, R.N. (1970). Test item arrangement, testing time, and performance. Journal of Educational Measurement, 7, 113-118.

McKeachie, W.J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. Journal of Abnormal and Social Psychology, 50, 93-98.

Monk, J.J., & Stallings, W.M. (1970). Effects of item order on test scores. Journal of Educational Research, 63, 463-465.

Munz, D.C., & Smouse, A.D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. Journal of Educational Psychology, 59, 370-374.

Oosterhoff, A.C., & Glasnapp, D.R. (1974). Comparative reliabilities and difficulties of the multiple-choice and true-false formats. Journal of Experimental Education, 42, 62-64.

Owens, R.E., Hanna, G.S., & Coppedge, F.L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. Journal of Educational Measurement, 7, 87-90.

Plake, B.S. (1980). Item arrangement and knowledge of arrangement on test scores. Journal of Experimental Education, 49, 56-58.

Ruch, G.M., & Charles, J.W. (1928). A comparison of five types of objective tests in elementary psychology. Journal of Applied Psychology, 12, 398-403.

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

Smith, K. (1958). An investigation of the use of "double-choice" items in testing achievement. Journal of Educational Research, 51, 387-389.

Smouse, A.D., & Munz, D.C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. Journal of Psychology, 68, 181-184.

Stanley, J.C. (1961). Studying status versus manipulating variables. In R.O. Collier & Elam, S.M. (Eds.). Research design and analysis. Bloomington IN: Phi Delta Kappan.

Williams, B.J., & Ebel, R.L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. The 14th Yearbook of the National Council on Measurements Used in Education (pp. 63-65). East Lansing, MI: Michigan State University.

Zuckerman, M. (1960). The development of an affect adjective check list for the measurement of anxiety. Journal of Consulting Psychology, 24, 457-462.

GENERAL REFERENCES

Mehrens, W.A., & Lehmann, I.J. (1978). Measurement and evaluation in education and psychology (2nd ed.). New York: Holt, Rinehart and Winston.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., Bent, D.H. (1975). Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill.

Raj, D. (1972). Design of sample surveys. NY: McGraw-Hill.