A COMPARISON
OF MULTIPLE-CHOICE TEST
RESPONSE MODES AND SCORING METHODS

by

Nina J. Thayer

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
EDUCATIONAL RESEARCH AND EVALUATION

APPROVED:

Robert B. Frary, Chairman

Lawrence H. Cross

Dennis E. Hinkle

Jerard F. Kehoe

Edwin P. Martin

John A. McLaughlin

May 19, 1982
Blacksburg, Virginia

ACKNOWLEDGEMENTS

I am indebted to Robert B. Frary for his counsel, advice, encouragement and continued support throughout this research.

To Dr. Dennis Hinkle and Dr. Lawrence Cross, who have provided advice throughout my graduate program, I am sincerely and deeply grateful.

To Dr. John McLaughlin and Dr. Edwin Martin, I owe special thanks for their continued encouragement and support.

To Dr. Jerard Kehoe, my thanks for taking a risk.

Especially, I thank my daughters , , and , for their courage, encouragement and support.

In particular, I thank my mother,
, for her endurance and belief; and my dad,
, for encouraging curiosity.

To IFO, more thanks than words can express.

DEDICATION

To and From

TABLE OF CONTENTS

LIST OF TABLES

Table page

CHAPTER I

Background of the Study

In the early 1900's, the use of objectively-scored, group-administered tests began to replace essay tests and oral examinations. Multiple-choice testing filled the need to administer quickly and score accurately large numbers of such examinations. As their use in schools, industry, civil and military services became widespread, some aspects of multiple-choice testing began to concern measurement experts and the public. Of particular interest was the fact that examinees who were totally ignorant of the content of some test items could gain points by guessing randomly among item choices, answering a proportion of those test items correctly.

Various scoring methods were devised in an attempt to discourage such random guessing. The conventional correction-for-guessing yielded mixed results in curbing the guessing behaviors, for which it had been devised (Rowley and Traub, 1977). Additionally, Horst (1932) had suggested, and Pressey (1950) and Davis (1959) had observed, that examinees gain points on a test by making use of partial information, the practice of eliminating some distractors from consideration before guessing among the remaining choices. However, partial information was not well understood and the presence of the resulting score gain was considered to be as undesirable as that from random guessing. Meanwhile, however, some measurement specialists and educators were using multiple-choice response/scoring alternatives, methods which permit or require the use of multiple-marks per test item. These methods credit an examinee for using partial information. Alternative methods were used by Pressey (1950) and Coombs (1953). Pressey used Answer-Until-Correct (AUC) and asked examinees to continue to punch item choices from a two-layered board until the correct choice (indicated by a red dot) was revealed. In contrast, Coombs gave examinees credit for marking item distractors and, in addition to the credit, imposed a penalty for marking the correct choice. These and various other methods give examinees a response/scoring incentive to use partial information when responding to test items.

Need for the Study

The response/scoring methods mentioned above have been used, studied, and evaluated in a variety of settings. A major concern of this investigator has been that the results from these studies have been quite mixed with regard to the reliability and validity estimates obtained from test scores (see Chapter II, Review of Literature). Some studies made use of tests which were not intended for course grading purposes. Examinees, aware of that, had no particular motivation to do well. However, examinee item responses, test scores and the resulting estimates of reliability and validity might have been quite different had test scores counted toward a course grade (Kogan and Wallach, 1964).

The research on alternative response/scoring methods also has ignored the differences in their schemes for crediting partial information and/or penalizing misinformation, the elimination of the correct choice from those item choices among which an examinee selects or guesses. When several response/scoring methods were examined analytically by Frary (1980), differences were noted to exist among the methods in the manner by which each credits or penalizes the examinee for comparable levels of information, partial information, total ignorance, or misinformation on a test item. These differences could have an effect on the reliability and validity estimates obtained from the scores; however, research has not been undertaken to compare such estimates across several alternative response/scoring methods.

Research Questions

There is a need to compare a variety of response/scoring methods in an academic environment with identical multiple-choice tests that will be used for grading purposes. Essentially, such a comparison could provide information regarding the effect of these alternative methods on:

1. Examinee item responses,

2. Resulting test scores, and

3. Estimates of reliability and validity calculated from the scores.

Description of the Study

The study results from the concerns discussed above and reports the comparison of seven different response/scoring methods used in an academic setting with students enrolled in a required undergraduate art appreciation course at a medium-sized state university. One response/scoring method was assigned randomly to each of the seven sections of 35 to 40 students. After taking identical pretests scored on the basis of the number of correct responses, each section took two subsequent tests using the response/scoring method previously assigned. All test scores counted toward course grades, and examinees were so informed. Responses were used to obtain estimates of internal consistency reliability and scores were used to estimate validity with the previous quarter's grade-point average and Scholastic Aptitude Test subscores.

An Evaluation Questionnaire, administered in each section, obtained self-report information about examinee studying habits, responding behaviors and testing preferences and experience. Observations of examinee behavior, subjective in nature, helped explain or confirm various empirical findings from this study or from the literature.

CHAPTER II

REVIEW OF THE LITERATURE

The review is limited to studies concerning the response modes and scoring methods related to this study and concentrates on their potential for providing insight into examinee responding, guessing and test-taking behaviors.

Response modes are defined by the instructions the examinees receive concerning the manner by which they are to select item choices. The more familiar methods allow only one choice to be selected per test item; alternative methods permit multiple marks per item. Elimination formats have examinees mark only item distractors and, when feedback is provided, the examinees are informed when they inappropriately mark the correct choice. Inclusion formats ask examinees to select those item choices necessary to feel sure that the correct choice is included among those marked. When feedback is provided, it is inevitable that the examinee discover the correct choice.

Scoring methods are the rules by which credits and/or penalties are awarded to the choice(s) made by an examinee in response to a test item. Response modes and scoring rules have been referred to in this study as response/scoring methods due to the interdependency of the scoring rule and the response mode being used.

Number Right (NR)

The NR response/scoring method asks examinees to mark the correct choice to an item and to mark every item. NR scoring credits each correctly marked item with one point and gives each incorrectly marked or omitted item a zero. Items receiving multiple marks are considered to be incorrect.

The NR method encourages examinee guessing behavior for all those examinees who do not know the correct choice to a test item. This is especially true for ill-prepared examinees who have a greater need and opportunity to guess (Frary, 1969). Because an item guessed correctly receives one point and an item guessed incorrectly receives no penalty, guessing pays with NR and, ethically, should be encouraged in the directions. As Davis (1964) and Slakter (1976) have noted, examinees who hesitate to guess are at an obvious scoring disadvantage compared to examinees who, by guessing, acquire credits.

Corrected for Guessing (CG)

A review of CG serves as a background against which to consider some of the studies which have been conducted since its inception. Numerous studies on test score reliability and validity included CG as one of the methods used for comparison purposes. Although not used as a response/scoring method in this study, CG does have implications for one method devised by Arnold and Arnold (1970) which was used and is discussed later in this chapter.

The CG response/scoring method asks the examinee to mark the correct choice on items where the examinee knows the correct choice or can eliminate at least one of the distractors from consideration. The scoring awards one point for each correctly-marked item and imposes a "correction" of -1/(n-1) points for each incorrectly-marked n-choice item.

CG was originally grounded on the "all or none" assumption that examinee responses are based either on complete knowledge of the correct choice or on total ignorance (Davis, 1964; De Finetti, 1965). This assumption implies that all incorrect items result from the random guessing of the totally ignorant examinee. Other levels of examinee information are not assumed to be measured by CG.

The correction for guessing has yielded mixed results in curbing guessing. This may be due, in part, to differences in guessing behaviors among examinees. Gritten and Johnson (1941) considered that the correction for guessing was a "...correction for individual differences in confidence...." to guess. Slakter (1968) found that examinees who were hesitant to guess when NR was used, became even more cautious about guessing when CG was used. Also, Cross and Frary (1977) found that a substantial proportion of examinees continued to guess randomly, contrary to test instructions which they admittedly had understood. This study also found that, when asked to mark their "best guess" on items previously omitted under CG, examinees marked more of these items correctly than predicted by chance.

The "all or none" assumption originally underlying the CG method has not been taken seriously for many years, for it has been shown that examinees can and do respond to a test item from any of a number of levels of information. (Pressey, 1950; Coombs, Milholland and Womer, 1956; Dressel and Schmid, 1953; Arnold and Arnold, 1970.) Davis (1964) suggested that an examinee may respond to a multiple-choice test item from any of the following levels of information regarding the item:

1. Sufficient knowledge to identify the correct choice.

2. Partial knowledge that permits elimination of one or more of the incorrect choices, followed by guessing among all of the remaining choices....

3. Guessing among all of the choices after considering the item as a whole.

4. Partial misinformation that leads to the elimination of one or more choices, including the correct choice, followed by guessing among all of the remaining choices.

5. Sufficient misinformation to identify as correct one of the incorrect choices.

These levels of examinee information were examined by Davis initially in 1959, using data generated by Mead and Smith in 1957. The Mead and Smith data consisted of 148 true-false items to which 100 examinees also had been asked to indicate whether their responses represented "certainty," "doubt," or "pure guess." By counting which of the above responses were accorded to items marked correctly, incorrectly, or omitted, Davis partitioned the 148,000 responding instances into his five levels of information and verified that partial-information made a contribution to examinee test-scores which probably was offset by misinformation.

Arnold and Arnold (AA-)

As devised by Arnold and Arnold (1970), AA- used an Elimination format. With AA-, each item is credited with the expected CG score as inferred from the level of information represented by an examinee's item choices (see Chapter III, Theoretical Considerations). However, if the examinee marks the correct choice, a uniform deduction of 1/(n-1) points is imposed (on an n-choice item) in place of any credit. This deduction is identical to the CG "correction."

Arnold and Arnold used their method in a university class in elementary statistics, asking examinees first to respond by marking item distractors and then to circle their best choice for the answer. Arnold and Arnold thereby obtained AA- item scores that provided insight into the levels of examinee information, and also acquired NR item scores. Both the AA- and the NR scores were used to rank the examinees. Arnold and Arnold found that those examinees who scored highly using NR methods also scored highly using AA-. This finding also was true for the examinees who scored poorly. However, the middle score range showed a significant difference in ranks for the examinees, a finding which was attributed to examinee use of partial information under the AA- method. Arnold and Arnold did not report estimates of reliability or validity for the test scores obtained in their study.

Coombs (CBS)

Reported by Coombs in 1953, CBS uses an Elimination format and credits an item one point for each distractor marked. In addition to any credits awarded, CBS imposes a penalty of (n-1) points (on an n-choice item) if the correct choice is marked.

Cross (1973) used CBS scoring as one of the methods by which to study examinee guessing. A series of three algebra tests was administered to twelve sections of eleventh grade students. For each test, the examinees first were instructed to respond using CBS methods and then to use the NR method, encouraging guessing, on those same items. Guessing was assumed to be very limited under the CBS method due to the penalty for marking the answer. Cross found that matched-half reliability estimates for score-sets became higher as the scores became more guessing-free, a condition observed from the number of choices among which the examinee was inferred to have guessed when both CBS and NR item marks were compared.

Lowry (1975) used a dual-responding method on six tests administered to three college biology classes. Examinees first used CBS responding and then were directed to mark their best guess to each item. Lowry's examinees thereby provided both CBS and NR scores. Results indicated that examinees gained by more than chance expectation when forced to guess. This outcome concurs with Slakter (1968) and with Cross and Frary (1977).

Coombs, Milholland and Womer (1956) conducted an empirical study of their method. Three tests, each consisting of 40 5-choice items, were administered to a total of 855 high school students who had been divided into three groups. A testing plan was developed such that each group received one test using CG methods, one test using CBS, and one test using a method that required examinees to rank item-choices according to their assurance in recognizing each choice as a distractor. The results indicated that--of the three methods used--CBS had the effect on test-score reliability of increasing the length of the test by as much as 20 percent, or from 40 items to 48 items.

The CBS method was used by Kohler (1971), who compared the reliability and validity coefficients estimated from CBS scores, CG scores, and scores from a method devised by Dressel (1953). Kohler found no significant differences among the estimates of reliability or of validity calculated from the test scores for each method.

Collet (1971) compared CBS and CG reliability and validity estimates obtained from a single administration of a test, scored three ways. The third method used by Collet in administering 100 5-choice items (taken from parallel forms of the Henmon-Nelson test of Mental Maturity) weighted each item choice according to the number of examinees choosing that choice (empirical choice weighting). Collet found that the reliability and validity estimates were superior for CBS scores.

Dressel (DRL)

Developed by Dressel and Schmid in 1953, DRL uses an Inclusion format and awards an item one point if the correct choice is marked and, in addition, imposes a 1/(n-1) point penalty (on an n-item choice) for every item distractor also marked. This method has been referred to as a response-complement to CBS.

Dressel and Schmid believed that a multiple-choice item "...may stimulate a rather involved and extended thought process on the part of the student." It was this belief that led them to consider a responding alternative which would permit a greater discrimination among examinees than was possible using NR or CG.

In 1953, Dressel and Schmid used their method to administer a two-part testing program as a final examination for five sections of 90 college students each. During the first hour of the exam, all students were given the same NR multiple-choice test. During the second hour, one group continued to use NR while the other four groups used DRL. Scores obtained using DRL were found to be significantly more reliable than those obtained from the same measure using NR.

Although special directions were provided for the DRL tests, the authors found that the examinees were unaccustomed to altering their responding methods in order to use multiple marks and needed instruction in learning how to indicate the extent of their knowledge when using partial information.

Answer-Until-Correct (AUC)

When adapted to provide the examinee with immediate feedback regarding item performance, the Inclusion format is termed AUC (Pressey, 1950). To provide this feedback, AUC can use a punchboard (Angell, 1949) which has choices with removable holes that reveal the correct answer with a colored dot. An erasable version of this punchboard covers the item choices with erasable ink shields under which letters or numbers indicate when the appropriate choice is erased by the examinee. Latent-image ditto masters make use of image-clearing pens which, when used by the student, chemically reveal the correct choice.

AUC has been used with each kind of feedback answer sheet and, when feedback is provided, it is inevitable that the examinee discover the correct choice to a test item. Pressey (1950) used the Angell punchboard for about 30 years to improve learning. Kaess and Zeaman (1960) also used the punchboard and found that knowledge of item performance on hard items differentially affected examinees. Once an initial distractor had been punched on a hard item, some examinees were noted to punch subsequent item choices unintentionally, by mistake.

The erasable version was used by Gilman and Ferry (1972) to administer a 66-item, 5-choice test. AUC scores were found to yield significantly higher split-half reliability estimates than a set of NR scores which were inferred from the single-erasure responses on the same AUC exam. The authors deduced that the increase in reliability probably was due to an apparent lengthening of the test, since each of an examinee's responses could have been considered to be a new decision representing, effectively, an increase in test length.

Hanna (1974) made use of AUC in eleven test administrations and then proceeded to correlate the examinee's AUC scores with inferred NR scores obtained from the AUC single erasure responses. Reliability estimates for the two sets of scores were nearly identical, or so close as to question the added effort of having administered and scored multiple-mark feedback tests.

The latent-image ditto transfer was used by Evans and Misfeldt (1974), who (possibly fallaciously) attributed impressive gains in reliability for AUC scores over NR scores to an effective tripling of the range of possible scores and to an increase in test score variance. For each examinee, a score had been obtained for each method, using separate administrations of parallel forms of their test with directions appropriate to the method used.

Cross (CRS)

An Elimination format providing feedback, CRS was devised by Cross (see Cross and Thayer, 1980; Frary, 1980). CRS awards each unerased item choice one point and each of an item's erased distractors two points. However, if the correct answer is erased, all item credit is lost for distractors already erased. Since the examinee is provided immediate feedback regarding item performance, the examinee who erases the correct choice, either through misinformation or guessing, will lose points by erasing further choices. CRS encourages examinees to guess once (on a 5-choice item) if no item choices are recognized as distractors, since the potential for gain is greater than for loss (see Chapter III, Theoretical Considerations).

Using latent-image forms, Cross did an empirical study of this method in 1980 with an examination that made use of NR response/scoring for one half of the test and CRS for the other half. Cross found that matched-half reliability estimates were not significantly different between methods.

Summary

Each of the studies reviewed made use of one or more multiple mark response/scoring methods. The results of these studies were mixed with regard to estimates of reliability and validity. Some studies yielded no significant difference in reliability estimates between response/scoring methods; other studies attributed significant differences to the use of partial information, to apparent increases in test length, or to severe penalties imposed by some methods. Several explanations emerge for the mixed findings:

1. Multiple-mark methods were novel when used. Examinees were not accustomed to considering the extent of their knowledge, nor to having a test score depend on such a skill.

2. In most of the studies reviewed, use of multiple marks had an effect on examinee item responding and on the resulting estimates of test score reliability and validity.

3. In some studies reviewed, tests and the resulting scores were not used for grading purposes; hence, examinees were under no particular motivation nor obligation to do well. Their responding, guessing, and test-taking behaviors in such situations probably were not indicative of the behaviors expected had the tests been intended for grading.

4. Awareness of the response/scoring methods to be used in the various studies perhaps prompted examinees to study differently for multiple mark tests than for traditional NR methods, thereby influencing the resulting item responses, the test scores and subsequent estimates of reliability and validity.

CHAPTER III

THEORETICAL CONSIDERATIONS

Overview

Response modes are defined by the instructions the examinees receive concerning the manner by which they are to select item choices. Scoring methods are the rules by which credits and/or penalties are awarded to the choice(s) made by an examinee in response to a test item. Due to the interdependency of the scoring rule and the response mode being used, response modes and scoring methods are referred to in this study as response/scoring methods, and they differ considerably in a number of ways:

1. Responding: the number of responses (choices) permitted per item, and the type of format by which examinees are to respond.

2. Level of Information: the level of information from which an examinee may respond.

3. Scoring: the item credit and/or penalty awarded an examinee for the level of information inferred from the number of item choices marked or erased.

4. Guessing: the extent to which the above conditions encourage guessing and/or require an examinee to respond from a true level of information regarding test items.

These four topics are reviewed in the following sections.

Responding

The more familiar methods allow only one choice to be selected per test item; alternative methods permit or require multiple marks per item and vary in the format by which examinees are directed to respond. Inclusion formats ask examinees to mark as many item choices as are believed necessary to insure that the correct choice is marked. Elimination formats instruct examinees to mark or erase only item distractors. The examinee receiving feedback using either an Inclusion or Elimination format will discover when the correct choice is appropriately or erroneously erased.

Levels of Information

By permitting multiple marks per test item, some methods provide a scoring incentive to the examinee to respond from any of a number of levels of information. The informed examinee, for example, knows the correct choice to a multiple-choice test item and marks the item appropriately. However, the partially-informed examinee does not know the item's correct choice but can identify one or more of the item's distractors. The totally ignorant examinee has no information about the item by which to discriminate one choice from another and is termed uninformed. However, if an item's correct choice is believed to be a distractor or a distractor is believed to be correct, the examinee is said to be misinformed.

Davis' five levels of information (see Chapter II, Review of Literature) are reviewed in Table 1.

Tables 2 through 18 extend Davis' levels further, discuss possible guessing strategies--given the number of item choices from which an examinee could guess and, for a hypothetical 5-choice test item having one correct choice (R) and four distractors (W), cite the extent of information for each level. The levels of information are lettered according to the information status, with "i" representing information and partial information, and "m" representing misinformation. Additionally, each level is ranked with a number which represents the extent of information, 1 being the most information and 5 the least or none.

These tables are helpful in providing insight into the variety and complexity of examinee responding, especially since the item scores which result from that responding are used to calculate descriptive statistics and estimates of test score reliability and validity.

TABLE 1

Levels of Information (Based on Davis)

INFORMATION LEVELS
Sufficient knowledge to identify the correct choice.
Partial knowledge that permits elimination of one or more of the incorrect choices, followed by guessing among all of the remaining choices.
Guessing among all of the choices after considering the item as a whole.
Partial misinformation that leads to the elimination of one or more choices, followed by guessing among all of the remaining choices.
Sufficient misinformation to identify as correct one of the incorrect choices.

TABLE 2

Level i1--Information

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i1	Examinee is informed. The Right choice and the four Wrongs are known. Appropriate choice(s) are marked or erased.	
	INCLUSION OR ELIMINATION FORMATS:	RWWWW kkkkk
Potential Guessing:	None is needed.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 3

Level i2--Partial Information, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i2	Examinee is partially-informed. Three choices are known to be Wrong and are not marked nor erased. The Right choice is NOT known.	
	INCLUSION FORMATS:	RWWWW --kkk
Potential Guessing:	(a) marking methods: Between the two remaining choices, which include the Right choice, guessing consists of marking only one rather than both choices. (b) feedback methods: One or both choices will be erased sequentially until the Right choice is found.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 4

Level i2--Partial Information, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
i2	Examinee is partially-informed. Three choices are known to be Wrong, and are marked or erased. The Right choice is NOT known.
	ELIMINATION FORMATS: RWWWW --mmm
Potential Guessing:	(a) marking methods: One of the two remaining choices may be marked as a guess. (b) feedback methods: A further erasure constitutes guessing between the two choices that remain, including the Right choice.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 5

Level i3--Partial Information, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i3	Examinee is partially-informed. Two Wrong choices are known and are not marked nor erased. The Right choice is NOT known.	
	INCLUSION FORMATS:	RWWWW ---kk
Potential Guessing:	(a) marking methods: Among the three choices which remain, including the Right choice, a guess consists of marking one or two choices rather than all three. (b) feedback methods: One to three of the choices will be erased sequentially until the Right choice is found.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 6

Level i3--Partial Information, Elimination

LEVELS	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
i3	Examinee is partially-informed. Two Wrong choices are known and are marked or erased. The Right choice is NOT known. ELIMINATION FORMATS: RWWWW ---mm
Potential Guessing:	(a) marking methods: One or two choices may be marked as a guess among the three choices that remain. (b) feedback methods: One or two further erasures constitute guesses among the three remaining choices.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known   ---=unknown

TABLE 7

Level i4--Partial Information, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i4	Examinee is partially-informed. One Wrong choice is known and is not marked or erased. The Right choice is NOT known.	
	INCLUSION FORMATS:	RWWWW ----k
Potential Guessing:	(a) marking methods: Among four choices which remain, including the Right choice, guessing consists of marking one to three choices rather than all four. (b) feedback methods: One to four of the choices that remain will be erased sequentially until the Right choice is inevitably found.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 8

Level i4--Partial Information, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i4	Examinee is partially-informed. One Wrong choice is known and is marked or erased. The Right choice is NOT known.	
	ELIMINATION FORMATS:	RWWWW ----m
Potential Guessing:	(a) marking methods: One to three choices may be marked as a guess among the four remaining choices. (b) feedback methods: One to three further erasures constitute a guess among the four remaining choices.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 9

Level i5--Total Ignorance, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i5	Examinee is uninformed, or totally ignorant regarding item. No choices are known.	
	INCLUSION FORMATS:	RWWWW -----
Potential Guessing:	(a) marking methods: Among the item's five choices, including the Right choice, any mark constitutes guessing. (b) feedback methods: The examinee guesses among all five choices until the Right choice is inevitably erased.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

TABLE 10

Level i5--Total Ignorance, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
i5	Examinee is uninformed, or totally ignorant regarding item. No choices are known.	
	ELIMINATION FORMATS:	RWWWW -----
Potential Guessing:	(a) marking methods: Any mark constitutes a guess. (b) feedback methods: Any erasure constitutes a guess.[1]	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown

[1] However, note that under CRS the first guess is encouraged.

TABLE 11

Level m4--Misinformation, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
m4	Examinee is misinformed. The Right choice is believed to be Wrong and is NOT marked nor initially erased. No other choices are known.	
	INCLUSION FORMATS:	RWWWW X----
Potential Guessing:	(a) marking methods: Among four of the item's choices, NOT including the Right choice, guessing consists of marking fewer than four choices. (b) feedback methods: None; the Right choice will be erased last, since it is the one choice believed to be Wrong.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 12

Level m4--Misinformation, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m4	Examinee is misinformed. The Right choice is believed to be Wrong and is marked or erased first. No other choices are known. ELIMINATION FORMATS: RWWWW m----
Potential Guessing:	(a) marking methods: Any further mark constitutes a guess. (b) feedback methods: None; any further erasure would reduce score.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 13

Level m3--Misinformation, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m3	Examinee is partially-misinformed. One Wrong choice is known and is not marked or erased. The Right choice, believed to be Wrong, is NOT marked nor initially erased. INCLUSION FORMATS: RWWWW / X--k-
Potential Guessing:	(a) marking methods: Among three remaining choices, NOT including the Right choice, marking fewer than three choices constitutes guessing. (b) feedback methods: All three of the remaining choices will be erased. Then guessing will occur between the Right choice and the correctly identified Wrong choice.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 14

Level m3--Misinformation, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m3	Examinee is partially-misinformed. The Right choice, believed to be Wrong, is marked, as is the one Wrong known to be Wrong. With feedback, the sequence of choice erasure will determine whether one or both of these two choices are erased.

ELIMINATION FORMATS: RWWWW
m--m-

Potential Guessing:

(a) marking methods:
One to two of the choices that remain may be marked.

(b) feedback methods:
The Right choice, believed to be wrong, may be erased first or second in sequence. Once the choice is erased, no Right choice is erased, no further erasures will be made.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 15

Level m2--Misinformation, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES	
m2	Examinee is partially-misinformed. Two Wrong choices known are neither marked nor erased. The Right choice, believed to be a Wrong, is not marked nor initially erased.	
	INCLUSION FORMATS:	RWWWW X--kk
Potential Guessing:	(a) marking methods: Between the two choices that remain, NOT including the Right choice, one choice constitutes guessing. (b) feedback methods: Both of the remaining choices choices will be erased. Then guessing will determine the sequence of erasures among the other three, including the answer.	

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 16

Level m2--Misinformation, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m2	Examinee is partially-misinformed. Two Wrong choices are known and are marked, as is the Right choice, believed to be Wrong. With feedback, the sequence of erasure of the Right choice determines whether or not all three choices are erased. ELIMINATION FORMATS: RWWWW m--mm
Potential Guessing:	(a) marking methods: Between the two choices that remain, one mark constitutes guessing. (b) feedback methods: The sequence of choice erasure will determine how many choices are erased prior to erasure of the Right choice.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 17

Level m1--Misinformation, Inclusion

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m1	Examinee is partially-misinformed. A Wrong choice is believed to be Right and is marked or erased first. The Right choice and the remaining Wrongs are believed to be Wrong if only by default.

INCLUSION FORMATS: RWWWW
Xkkmk

Potential Guessing:

(a) marking methods:
None; the examinee believes a Wrong choice is Right and marks only that choice.

(b) feedback methods:
The choice erased as being Right will be a Wrong, and the examinee will continue erasing until the RIGHT is erased.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known ---=unknown
X=eliminated from consideration

TABLE 18

Level m1--Misinformation, Elimination

LEVEL	INFORMATION STATUS AND POSSIBLE GUESSING STRATEGIES
m1	Examinee is partially-misinformed. A Wrong choice is believed Right and is not marked. The Right choice and the remaining Wrongs are believed Wrong, if only by default, and are marked. With feedback, sequence of erasure determines the number of choices erased before the Right choice.

ELIMINATION FORMATS: RWWWW
mmmXm

Potential
Guessing:

(a) marking methods:
None:
The Right choice (and remaining Wrongs) are marked if only by default.

(b) feedback methods:
A Wrong choice is believed to be Right, and is NOT erased. Three Wrongs are known to be Wrongs, the Right is believed to be Wrong. THESE choices are erased sequentially in the order of the examinee's belief in the "Wrongness" of the choices until the Right choice is revealed.

R=Right (correct) W=Wrong (distractor)
m=marked or erased
k=known X=removed from consideration

Scoring

As shown in Tables 2 to Table 18, an examinee can respond to a test item from different levels of information, depending on the method used. By permitting multiple marks per test item, many response/scoring methods provide a scoring incentive to the examinee to respond from these levels of information. However, methods which encourage or provide a scoring payoff for examinee guessing do not require the examinee to respond from a self-perceived true level of information regarding ANY test item. The extent to which each examinee responds to test items from a true level of information is expected to influence coefficients of reliability and validity estimated from the resulting scores, since item scores which are comprised primarily of true score and measurement error are expected to attain higher estimates of internal consistency reliability than item scores which, in addition, contain a guessing component (Frary, 1969).

For each of the response/scoring methods used in this study, the responding, scoring and requirements placed on the examinee to respond from a true level of information regarding test items are discussed, and the observed item scores are tabled (see Tables 19-25) with the corresponding expected item scores (raw and standardized) as derived from the probabilities which an examinee has of choosing the appropriate choice(s) from those among which he guesses.

<u>Number Right</u> (<u>NR</u>)

Number-Right (NR) responding asks the examinee to mark the correct choice to an item and to mark every item.

Observed NR scoring (see Table 19) gives each correctly-marked item a raw score of one point, and each incorrectly-marked item a raw score of zero. Items omitted or having multiple marks receive a score of zero. The total score of an examinee is the count of the number of test items answered correctly.

Credit is awarded to a correctly-marked item whether marked through partially-informed or uninformed guessing, or through knowledge of the correct choice. All levels of misinformation result in incorrectly-marked, zero-scored items, since the misinformed examinee either guesses among a set of item choices which does not include the correct choice, or mistakenly believes a distractor is the answer. This examinee cannot possibly choose the correct choice under NR. NR scoring is uniform for all incorrectly-marked

TABLE 19

NR Raw and Expected Item Scores

Observed NR Raw Scores

	RWWWW		
	m----	=	1 point
Omit =	-----	=	0 points
	---m-	=	0 points

NR Expected Item Scores
For the Levels of Information i1 to m4

LEVEL	RAW (NR)[2]	STDZD[1]
i1	1.00	1.00
i2	0.50	0.38
i3	0.34	0.16
i4	0.25	0.06
i5	0.20	0.00
m4	0.00	-0.25
m3	0.00	-0.25
m2	0.00	-0.25
m1	0.00	-0.25

[1] Linear transformation to assign 1.00 to Level i1 and 0.00 to Level i5.

[2] Assumes average gains from guessing whenever answer not known.

items, whether marked through misinformation or through the guessing of a partially-informed or an uninformed examinee.

Since the NR method promotes examinee guessing behavior for all examinees who do not know the answer to a test item, the method makes no requirement on the examinee to reveal a true level of information regarding any test item. In the absence of scoring or admonitory restraints on guessing, scores obtained from NR tests do not reflect the various levels of information from which an examinee may respond and are comprised of true score, error and guessing components.

Corrected-for-Guessing (CG)

Corrected-for-guessing (CG) responding asks the examinee to mark the correct choice on items where the examinee knows either the correct choice or can eliminate from consideration at least one of the wrong choices. Examinees are discouraged from guessing randomly among all of an item's choices.

Observed CG scoring awards one point for each correctly-marked item and imposes a correction of $-1/(n-1)$ points for each incorrectly-marked n-choice item. Only one mark per item is permitted, and CG observed item scores differ from expected item scores, shown in Table 20.

TABLE 20

CG Raw and Expected Item Scores

Observed CG Raw Scores

	RWWWW	
	m----	= 1 point
Omit	-----	= 0 points
	---m-	=-.25 points

CG Expected Item Scores For the Levels of Information i1 to m4

LEVEL	RAW&STDZD[1] (CG)[2]
i1	1.00
i2	0.38
i3	0.16
i4	0.06
i5	0.00
m4	-0.25
m3	-0.25
m2	-0.25
m1	-0.25

[1] Linear transformation to assign 1.00 to Level i1 and 0.00 to Level i5.

[2] Assuming average gain from guessing where appropriate.

Arnold and Arnold (AA-)

The Arnold and Arnold (AA-) method uses an Elimination format which was adapted to an Inclusion format (AA+) for this study. For both of these methods, an item is awarded the expected CG item score for the level of information which can be inferred from the number and type of item choices marked.

Guessing provides neither an advantage nor a penalty for the partially-informed nor the uninformed examinee, since such gains, when averaged, yield the same expected item score that would have been received prior to guessing. Even so, examinees may be tempted to guess and, except for such guessing, the Arnold methods encourage examinees to respond from a true level of information regarding a test item and credit that responding according to the level of information revealed by the choices marked. The misinformed examinee, unaware of the misinformed nature of the choices marked, responds from a true level of misinformation; however, the uniformly low penalty for misinformation--imposed regardless of the number of choices marked--does not reflect different levels of misinformation in the scoring. Hence, indices of reliability and validity are not derived from scores representing these different levels, and are expected to be higher than those for NR but lower than those obtained from methods which vary the scoring for each level of misinformation.

NR expected item scores have been shown to be linear transformations of CG expected item scores; therefore, the AA+, AA-, NR, and CG methods yield metrically comparable expected item scores, after standardization. AA+ and AA- scoring for each level of information is presented in Table 21.

Coombs (CBS)

The CBS method uses an Elimination format where the examinee is to mark item distractors. CBS scoring assigns an item one point for each distractor marked and, in addition, imposes a penalty of -(n-1) points (on an n-choice item) if the correct choice is marked. For a 5-choice item, this becomes a 4-point penalty for marking the correct choice, and this penalty is subtracted from any credits already earned for the item. Responding and scoring for the levels of information are presented in Table 22.

Zero is the CBS item score for total ignorance, while CBS penalizes misinformation whenever the correct choice is marked. Since examinees can accumulate both positive and

TABLE 21

AA Raw and Expected Item Scores

Observed AA Raw Scores

AA+				AA-
RWWWW				RWWWW
m----	=	1.00	=	-mmmm
mm---	=	0.38	=	--mmm
mmm--	=	0.16	=	---mm
mmmm-	=	0.06	=	----m
mmmmm	=	0.00	=	-----
-mmmm	=	-0.25	=	m----
--mmm	=	-0.25	=	mm---
---mm	=	-0.25	=	mmm--
----m	=	-0.25	=	mmmm-

AA Expected Item Scores For the Levels of Information i1 to m4

LEVEL	RAW & STDZD[1] (AA- and AA+)
i1	1.00
i2	0.38
i3	0.16
i4	0.06
i5	0.00
m4	-0.25
m3	-0.25
m2	-0.25
m1	-0.25

[1] linear transformation to assign 1.00 to Level i1 and 0.00 to Level i5.

TABLE 22

CBS Raw and Expected Item Scores

Observed CBS Raw Scores

RWWWW		RWWWW	
-mmmm	= 4 points	m----	= -4 points
--mmm	= 3 points	mm---	= -3 points
---mm	= 2 points	mmm--	= -2 points
----m	= 1 point	mmmm-	= -1 points

RWWWW

Omit = ----- = 0 points

CBS Expected Item Scores For the Levels of Information i1 to m4

LEVEL	RAW	STDZD[1]
	(CBS)	
i1	4.00	1.00
i2	3.00	0.75
i3	2.00	0.50
i4	1.00	0.25
i5	0.00	0.00
m4	-4.00	-1.00
m3	-3.00	-0.75
m2	-2.00	-0.50
m1	-1.00	-0.25

[1] Linear transformation to assign 0.00 to Level i1 and 0.00 to Level i5.

negative points to a test item, depending on the number of choices marked, examinees consider carefully the extent of their knowledge regarding test items before they respond, lest they acquire a penalty for marking the right choice. Such an examinee is not likely to guess; rather, the examinee is expected to respond from a true level of information. Estimates of reliability and validity obtained for scores from a CBS test are expected to be comprised of more true score component and have less variance due to guessing than the estimates obtained from NR, AA-, or AA+, which do not provide insight into the different levels of misinformation.

Dressel (DRL)

Dressel (DRL) uses an Inclusion format which has examinees mark only the choices believed necessary to include the correct choice.

DRL scoring awards an item one point if the correct choice is marked but removes 1/n-1 point on an n-choice item for each item distractor also marked. DRL scoring is CBS scoring divided by the constant four and has been referred to as the response-complement to CBS. Responding and scoring for the levels of information are presented in Table 23. For this study, DRL scoring was multiplied by the constant four in order to compare results with CBS scores.

TABLE 23

DRL Raw and Expected Item Scores

Observed DRL Raw Scores[0]

RWWWW				RWWWW			
m----	=	4	points	-mmmm	=	-4	points
mm---		3	points	--mmm	=	-3	points
mmm--	=	2	points	---mm	=	-2	points
mmmm-	=	1	point	----m	=	-1	point

	RWWWW	
Omit =	-----	= 0 points

DRL Expected Item Scores
For the Levels of Information i1 to m4

LEVEL	RAW	STDZD[1]
	(DRL)	
i1	4.00	1.00
i2	3.00	0.75
i3	2.00	0.50
i4	1.00	0.25
i5	0.00	0.00
m4	-4.00	-1.00
m3	-3.00	-0.75
m2	-2.00	-0.50
m1	-1.00	-0.25

[0] Original DRL Scoring multiplied by the constant 4 for comparison with CBS scores.

[1] Linear transformation to assign 0.00 to Level i5 and 1.00 to Level i1.

With DRL, zero is the item-score for total ignorance. Misinformation is penalized according to the number of incorrect choices marked when the correct choice is not marked.

The misinformed examinee is unaware of the misinformed nature of his responding and may respond from a true level of misinformation regarding test items. Since DRL requires the examinees to consider the extent of their knowledge before responding, DRL--like CBS--is expected to result in item and test scores comprised primarily of true score and measurement errors and should yield reliability and validity indices higher than those obtained from methods making a lesser requirement on the examinee for a true report.

Answer-Until-Correct (AUC)

Using a test form which has erasable answer shields, AUC asks examinees to erase item choices until a letter, keyed to the correct choice, is erased. AUC awards each unerased choice one point.

With feedback methods, the sequence of choice erasure can influence the number of erasures made by the examinee before erasing the correct choice, but it is inevitable that the correct choice is erased. For example, the misinformed

TABLE 24

AUC Raw and Expected Item Scores

Observed AUC Raw Item Scores

RWWWW	RWWWW
m---- = 4 points	mmm-- = 2 points
mm--- = 3 points	mmmm- = 1 point

RWWWW

Omit = ----- = 0 points

or = mmmmm

AUC Expected Item Scores
For Levels of Information i1 to m4

LEVEL	RAW	STDZD[1]
	(AUC)[2]	
i1	4.00	1.00
i2	3.50	0.75
i3	3.00	0.50
i4	2.50	0.25
i5	2.00	0.00
m4	0.00	-1.00
m3	0.50	-0.75
m2	1.00	-0.50
m1	1.50	-0.25

[1] Linear transformation to assign 1.00 to Level i1 and 0.00 to Level i5.

[2] Assumes average gains from guessing where appropriate.

examinee first will erase the entire set of wrong choices mistakenly believed to contain the correct choice. Thereafter, the credit received will depend on the sequence of choice erasing from among those item choices this examinee now believes contain the correct choice.

Since erasing sequence can influence the number of erasures made by an examinee, it is not possible to determine whether an erasure is made from information, partial information, total ignorance, or misinformation. Although the responding directions place the requirement on the examinees to respond from a true level of information, these erasing and subsequent scoring procedures do not incorporate item scores for actuals levels of information into the computations of test or item statistics.

Cross (CRS)

CRS, an elimination method which uses erasable forms, asks examinees to erase only item distractors. CRS awards each unerased item-choice one point, and each of the item's erased distractors two points. However, if the correct choice is erased, all item credit for distractors already erased is lost (see Table 25). Compared with other methods, the CRS method reverses the direction of the penalty for the levels of misinformation. If this were not so, an examinee who mistakenly erases the correct choice on the first erasure would profit by making further erasures.

The uninformed CRS examinee, totally ignorant regarding an item's choices, is encouraged to guess on one choice, since there are four chances in five of erasing a distractor with only one erasure, thereby gaining a credit of six points and only one chance in five of erasing the correct choice for a credit of four points.

For all levels of misinformation except m4, the sequence of choice erasure can influence the number of choices erased prior to the correct choice. The misinformed examinee at level m4 mistakenly believes that the correct choice is incorrect, KNOWS no other choices, and will inevitably erase the correct choice first and stop there in order to retain the item's remaining 4 points. This level of examinee misinformation is readily observable. However, this same erasure could be made by the totally ignorant examinee who, following directions to guess just once, follows directions! It seems reasonable that both of these examinees should receive the same score of four points since neither has any information regarding the remaining item choices.

The CRS method requires partially-informed examinees to consider erasures carefully lest credit already obtained be

TABLE 25

CRS Raw and Expected Item Scores

Observed CRS Raw Scores

RWWWW		RWWWW	
-mmmm	= 9 points	m----	= 4 points
--mmm	= 8 points	mm---	= 3 points
---mm	= 7 points	mmm--	= 2 points
----m	= 6 points	mmmm-	= 1 point

RWWWW

Omit = ----- = 0 points

CRS Expected Item Scores
For Levels of Information i1 to m4

LEVEL	RAW	STDZD[1]
	(CRS)[2]	
i1	9.00	1.00
i2	8.00	0.71
i3	7.00	0.41
i4	6.00	0.12
i5	5.60	0.00
m4	4.00	-0.47
m3	3.50	-0.62
m2	3.00	-0.76
m1	2.50	-0.91

[1] Linear transformation to assign 1.00 to level i1. and 0.00 to level i5.

[2] Assumes average gains from guessing where appropriate.

lost. Hence, guessing with the CRS method is hardly worth considering, intentionally, except at level i5. The CRS scoring incorporates the levels of information of the examinees into computations of reliability and validity estimates. However, it does credit misinformation in the reverse of both CBS and DRL and, for that reason, is not metrically comparable to either of these methods.

Scoring Overview

As shown in Table 26, the response/scoring methods differ considerably in the Expected raw and standardized item scores for the different levels of information.

Scoring penalties may inhibit the guessing behavior of some guessing-prone examinees, but it is doubtful whether penalties influence the responses of misinformed examinees, who mistakenly believe they are responding appropriately.

Although alternative methods may provide scoring incentives to an examinee to respond to test items from a true level of information, the methods vary in their provision for incorporating the scores for these levels into the computation of estimates of reliability or validity.

TABLE 26

Expected Raw and Standardized[1] Item Scores

(based on Frary, 1980)

For the Levels of Information i1 to m4

LEVEL	RAW	STDZ	RAW	STDZ	RAW	STDZ	RAW		STDZ
	(CBS/DRL)		(AUC)[2]		(CRS)[2]		(NR)[2]	(AA)	(NR/AA)
i1	4.00	1.00	4.00	1.00	9.00	1.00	1.00	1.00	1.00
i2	3.00	0.75	3.50	0.75	8.00	0.71	0.50	0.38	0.38
i3	2.00	0.50	3.00	0.50	7.00	0.41	0.33	0.16	0.16
i4	1.00	0.25	2.50	0.25	6.00	0.12	0.25	0.06	0.06
	0.00	0.00	2.00	0.00	5.60	0.00	0.20	0.00	0.00
m4	-4.00	-1.00	0.00	-1.00	4.00	-0.47	0.00	-0.25	-0.25
m3	-3.00	-0.75	0.50	-0.75	3.50	-0.62	0.00	-0.25	-0.25
m2	-2.00	-0.50	1.00	-0.50	3.00	-0.76	0.00	-0.25	-0.25
m1	-1.00	-0.25	1.50	-0.25	2.50	-0.91	0.00	-0.25	-0.25

[1] Linear transformation to assign 1.00 to level i1 and 0.00 to level i5.

[2] Assumes average gains from guessing where appropriate.

CHAPTER IV

METHODOLOGY

Seven different response/scoring methods were used to administer two of three multiple-choice tests given to students enrolled in seven sections of an art appreciation course at a medium-sized state university. Six sections of the art appreciation course each had approximately 38 students, while the seventh had 24 students; the overall total was 253 students.

The art appreciation course was required for all of the degree programs at the university at some time during each student's four-year academic course of study. The placement of any particular student into any particular section has been assumed to have been relatively uninfluenced by factors other than scheduling on the part of the Registrar or on the time at which particular students were permitted to register and a substantial degree of randomization in assignment of students into sections is asserted. Additionally, as intact groups, each section was assigned randomly to one of the response/scoring methods used in this study: NR, AA-, AA+, CBS, DRL, AUC, and CRS.

Personal Data

Approval to gain access to student records was acquired from the individual students, each of whom signed a release form later filed with the Registrar of the university. University records contained the previous quarter's grade point average, high school rank and grade point average, and Scholastic Aptitide Test subscores.

Test Construction and Development

All course sections (GRP) used the same text, Artforms (Preble, 1978). Text-related multiple-choice quizzes were developed, administered, and used for grading purposes. In order to coordinate the content of the test material, group and individual meetings were held with the professors teaching the art appreciation course. The professors were asked for potential test items and for preferred testing dates. Readings covering quiz material were assigned to the classes and, where possible, quizzes on the this material were administered prior to class discussion of the content.

Items were constructed after a thorough review of the text content and were devised to provide equally likely selections from which totally ignorant examinees would have difficulty choosing. For each quiz, professors added additional test items of their own choosing, using true-false, matching, fill-in the blank and slide/picture identifications.

In the event that sections did not use the same test on the same day, test items were relocated within the body of the test and choices were reordered within items to discourage an exchange of examinee information regarding item or choice sequencing.

Quizzes One through Three

A 5-choice 25-item pretest was developed from the text. This first quiz, administered to all seven sections between the second and third weeks in the quarter, used the NR response/scoring method and provided an initial indication of examinee performance for each section.

A second multiple-choice test of 18 5-choice items, constructed using the same principles as the first, was administered to each section between the fourth and fifth weeks of the quarter. Although this quiz served as an examinee training exercise in the use of the response/scoring method randomly assigned to each section, the quiz results counted toward course grades and examinees were so informed.

A third quiz, administered between the seventh and eighth weeks of the quarter, was comprised of 25 5-choice items, 20 of which were common to all sections. Scores from this quiz, which was administered to each section using the same response/scoring method that had been used for the second quiz, also counted toward course grades.

Quiz Administration

For each response/scoring method, including NR, an instruction sheet was provided to every examinee, and explained the method's responding and scoring principles (See Appendix A). Each instruction sheet had two or more examples for the examinee to try for practice. The instructions also were written on the chalkboard and were explained verbally. Pencils were provided for those sections using machine scorable answer sheets, as well as for those methods needing clean erasers.

The quizzes were completed within the 50-minute class period or as soon as the examinee was finished. None of the examinees needed longer than the class period and most took only 25 minutes. The author and the professor were present during each test administration to answer response/scoring method questions.

Data Analysis

Descriptive statistics were generated for all of the variables in the study. For the initial 25-item NR quiz, response sheets were optically scanned and their magnetic tape images were computer analyzed to produce scores and conventional item analysis. Data included means, standard deviations, the number of examinees per section, and KR20 internal consistency reliability; however, Cronbach's alpha estimated reliability for multiple mark quizzes.

In all sections using multiple mark methods, Quizzes Two and Three were scored by hand using scoring keys that noted only the correct choice. Additionally, for each response/scoring method quiz, a simple count was kept of the numbers of items to which the various number of marks had been made.

An analysis of variance (ANOVA), using the Statistical Analysis System (SAS, 1979), was performed on the scores for each of the criterion variables to determine whether the sections differed significantly with regard to criterion performance.

Pearson correlation coefficients were computed among scores on all of the variables and quizzes to provide validity estimates. Each correlation was tested under the null hypothesis rho=0. Additionally, each possible pair of estimates was tested for significant differences (using the Z transformation), under the null hypothesis rho1=rho2 (see Sokal and Rohlf, 1969).

KR20 or Cronbach's alpha measures of internal consistency reliability were computed for the items from each quiz. For Quiz Two and and Quiz Three, all possible pairs of estimates were tested for significant differences, using Feldt's test (1969).

Response/Scoring Method Evaluation

Examinee studying and responding behaviors and test-taking strategies were inventoried for each response/scoring method on an objectively-scored self-report evaluation questionnaire (see Appendix B) which covered such topics as test studying habits, testing method preferences, previous testing experience, and test-taking strategies. Classroom observations of examinee behavior were recorded in a journal for each section.

CHAPTER V

FINDINGS

A total of 253 students enrolled in the seven sections of Art Appreciation provided scores on Quizzes One through Three. Access to their university records was obtained for 203 of the 253 students. Some examinees returned the approval form beyond the Registrar's deadline; others knew that they had no SAT score nor GPA in their records. Only 23 of the 253 students preferred to keep their records confidential.

For this study, quiz data from all students were retained for calculations of descriptive statistics. The use of some statistical tests and procedures necessitated the deletion of those students whose data were incomplete. University records did not contain GPA's for most transfer students, although their records often provided both VSAT and MSAT subscores. Therefore, the 164 students who provided VSAT and MSAT data were not necessarily the same students providing GPA's.

Criterion Variables

GPA. Student grade point averages (GPA) from the previous quarter were obtained from university records. These were available for 195 of the 203 students. GPA provided an index of student achievement in the existing university environment and was used as a criterion related variable for scores obtained from Quizzes One through Three. Descriptive statistics for GPA are presented in Table 27.

A one-way analysis of variance (ANOVA) was performed on the GPA data, using the Statistical Analysis System (SAS). The resulting F value of 1.61 was not signficant (p=.145).

VSAT-MSAT. Both verbal and math Scholastic Aptitude Test subscores were available for 164 students. These subscores were used as criterion variables with scores obtained from Quizzes One through Three. Descriptive statistics for VSAT and MSAT are presented in Table 28.

A one-way ANOVA performed on the VSAT data resulted in an F value of 1.28, which was not significant (p=0.269). An F value of 1.85, resulting from the ANOVA on the MSAT data (P=0.0933), was not significant.

TABLE 27

Descriptive Statistics: GPA

Section	N	Mean	Std.Dev.
CBS	33	2.36	0.72
DRL	30	2.55	0.70
AUC	24	2.55	0.61
CRS	29	2.55	0.62
NR	32	2.55	0.76
AA-	33	2.39	0.58
AA+	14	2.01	0.75

ANOVA F=1.61 (p=.145)

TABLE 28

Descriptive Statistics: VSAT-MSAT

Section	N	Mean	Std.Dev.
		VSAT	
		MSAT	
CBS	27	446.3	26.78
		483.7	22.83
DRL	23	396.9	20.21
		449.6	21.38
AUC	20	432.5	20.33
		461.5	22.29
CRS	26	405.8	22.83
		451.9	25.33
NR	27	424.1	21.54
		426.8	19.19
AA-	27	414.1	25.86
		439.6	19.10
AA+	14	412.1	33.65
		435.0	16.54

ANOVA Fvsat=1.28 (p=.269)
ANOVA Fmsat=1.85 (p=.093)

Quiz One

Quiz One (Q1) consisted of 25 5-choice items administered using an NR scoring rule.

Descriptive Statistics. Quiz One descriptive statistics for each section are presented in Table 29. The CRS section obtained the highest and the NR section the lowest Q1 means. Both of these sections took Quiz One several days after the other five sections had had their quizzes scored and returned. Examinees did have time to exchange information regarding Quiz One. However, since these two sections attained the extremes of the Quiz One means, it seems unlikely that an exchange of information influenced examinee performance, for that information selectively would have had to favor one section while it hindered the other. Therefore it has been concluded that examinee benefits from exchanges of information probably had little or no influence on the mean performance observed in either the NR or the CRS sections.

ANOVA. A one-way ANOVA performed on the Q1 scores resulted in a significant F value of 9.11 (p=.0001). The Duncan Multiple Range Test, modified by Kramer (1956), identified that the NR and CRS sections differed significantly from all other sections (see Table 29).

TABLE 29

Descriptive Statistics: Q1

Section	N	Mean	St.Dev.
CBS	37	16.08*+	3.76
DRL	39	15.64*	4.11
AUC	35	16.14*+	4.53
CRS	41	19.17	3.19
NR	38	12.97	4.64
AA-	39	17.77 +	3.56
AA+	24	16.04*+	3.53

*,+ Section means which were not significantly different (p=.05) using the Modified Duncan Multiple Range Test (see Kramer, 1956; SAS 1979).

KR20 Reliability. For the Quiz One scores, KR20 reliability was estimated at the same time that the quizzes were machine-scored. The estimates obtained for sections CBS, DRL and AUC; and for sections AA+ and AA-, resulted from the pooling of the Quiz One test forms (item data are no longer available by which to calculate individual section estimates). The resulting KR20 reliability estimates for each section are presented in Table 30.

Both the AA+ and the AA- sections were taught by the same professor. The pooling of these sets of tests for scoring and computational purposes increased the sample size from which KR20 was estimated without adding the confounding effects of teacher differences.

Validity. Quiz One validity estimates obtained for GPA, VSAT, and MSAT are presented in Table 31. Each Pearson product-moment validity estimate was tested under the null hypothesis, rho=0. Probability levels attained are given in the body of Table 31.

GPA: Of the 195 students contributing GPA's to this study, 189 also took Quiz One. All possible pairs of GPA estimates were tested (using the Z transformation), under the null hypothesis rho1=rho2 (see Sokal and Rohlf, 1969). Each pairs of section estimates which was not significantly

TABLE 30

KR20 Reliability: Q1

Section	N	KR20	Sm
CBS	37	.735	4.056
DRL	39		
AUC	35		
CBS	41	.631	1.803
NR	38	.762	2.164
AA-	39	.704	2.018
AA+	24		

Sections whose Quiz One tests were processed together:

Set One: CBS and DRL; and AUC
Set Two: AA+ and AA-

TABLE 31

Validity Estimates: Q1

Section	N	GPA	N	VSAT	MSAT
CBS*	32	.331	26	.295	.186
		p=.0644		p=.1440	=.3635
DRL* **	30	.563	23	.292	.171
		p=.0012		p=.1767	=.4347
AUC **	24	.798	20	-.333	-.004
		p=.0001		p=.1514	=.9867
CRS*	29	.243	26	.204	-.055
		p=.2040		p=.3190	=.7886
NR * **	27	.582	27	.201	.282
		p=.0014		p=.3144	=.1537
AA-*	33	.471	27	.402	.247
		p=.0056		p=.0376	=.2127
AA+*	14	.343	14	-.208	.459
		p=.2297		p=.4757	=.0990

p: attained under Ho: rho=0
*, **: n.s.d. per GPA pair,
under Ho: rho1=rho2.

different (p=.05 per pair) is indicated by a star and/or a double-star in Table 31.

VSAT-MSAT: Quiz One scores were available for 164 of the 167 examinees who contributed the VSAT-MSAT data. The Pearson product-moment correlations are presented in Table 31. The probability level attained for each estimate is given in the body of the table. Since the estimates fluctuated considerably from one section to another, VSAT and MSAT scores were plotted, by section, with the Quiz One scores. The resulting rectangular or circular plots indicated that little if any relationship existed between these variables and the Quiz One scores.

Quiz Two: Q2

Quiz Two had 18 items and was used as practice in the response/scoring method to be used for Quiz Three. Examinees were aware that the quiz was to be used for grading purposes.

Descriptive Statistics. Descriptive data for Quiz Two are presented in Table 32.

Sections CRS, CBS, and NR took Quiz Two several days after the other sections, and examinees had time to exchange item and choice sequencing information. As a precaution

TABLE 32

Descriptive Statistics: Q2

Section	N	Mean	St.Dev.
CBS	37	38.59	14.93
DRL	39	32.10	17.92
AUC	35	60.60	6.82
CRS	41	128.07	18.85
NR	38	11.35	3.19
AA-	39	12.29	3.14
AA+	24	13.81	3.71

against such an exchange, the items and choices for Quiz Two were relocated within the body of the test, as had been done previously for Quiz One.

Reliability. Cronbach's Alpha estimates of internal consistency reliability are presented in Table 33.

Although both of the Inclusion format methods (DRL and AA+) attained higher reliability estimates than their Elimination counterparts, no pair of estimates was significantly different (p=.05) using Feldt's test (1969).

Validity. Quiz Two correlations with GPA, VSAT, and MSAT are presented in Table 34. Probability levels attained, under the null hypothesis rho=0, are given in the body of the table.

GPA: The CRS and the CBS sections, both Elimination formats, obtained the highest GPA validity estimates (.697 and .549, respectively). However, when all possible pairs of estimates were tested (using the Z transformation), under the null hypothesis rho1=rho2, there was no significant difference between any of the possible pairs of estimates.

VSAT-MSAT: The estimates for both VSAT and MSAT fluctuated considerably and the rectangular or circular plots indicated the absence of a relationship between these variables and quiz scores.

TABLE 33

Reliability: Q2

Section	N	KR20	Sm
CBS	37	.673	8.54
DRL	39	.752	8.92
AUC	35	.689	3.80
CRS	41	.696	10.39
NR	38	.664	1.85
AA-	39	.692	1.74
AA+	24	.722	1.96

No pair of estimates was significantly different, p=.05, using Feldt's Test (1969).

Quiz Three: Q3

Each of the seven sections of Art Appreciation used the response/scoring method for Quiz Three which had been used for Quiz Two. Quiz Three had 25 items, 20 of which were common to all seven sections. Only the results from these 20 items have been reported in this section.

Descriptive Statistics. Descriptive data for Quiz Three are presented in Table 35.

Reliability. Cronbach's Alpha estimates of internal consistency reliability for Quiz Three responses are presented in Table 36. Paired estimates were tested for significant differences (p=.05 per pair), using Feldt's test (1969). Results indicated that estimates for the CBS, CRS and NR sections each were significantly different from the DRL, AUC, AA- and AA+ sections, respectively; although CBS, CRS and NR were not significantly different from each other. Similarly, paired estimates between the DRL, AUC, AA- and AA+ sections were not significantly different, respectively, from each other.

TABLE 34

Validity Estimates: Q2

Section	N	GPA[1]	N	VSAT	MSAT
CBS	33	.549	26	.116	-.054
		p=.0009		p=.5635	=.7875
DRL	30	.452	23	.345	.111
		p=.0122		p=.1070	=.6130
AUC	24	.507	20	-.388	-.141
		p=.0114		p=.0909	=.5533
CRS	29	.697	26	.115	.205
		p=.0001		p=.5745	=.3142
NR	27	.606	27	.182	.168
		p=.0008		p=.3646	=.4013
AA-	33	.488	27	.066	.232
		p=.0040		p=.7376	=.2122
AA+	14	.484	14	.189	.517
		p=.0795		p=.5164	=.0586

p: attained under Ho: rho=0

[1] Under Ho: rho1=rho2, there was no significant difference (p=.05) between any of the possible pairs of GPA estimates.

TABLE 35

Descriptive Statistics: Q3

Section	N	Mean	St. Dev.
CBS	37	30.73	14.16
DRL	39	33.72	17.39
AUC	35	66.31	7.69
CRS	41	145.02	16.84
NR	38	11.28	2.84
AA-	39	12.51	3.91
AA+	24	12.11	4.34

TABLE 36

Reliability: Q3

Section	N	KR20	Sm
CBS	37	.518*	9.83
DRL	39	.730 +	9.04
AUC	35	.716 +	4.09
CRS	41	.599*	11.18
NR	38	.590*	1.81
AA-	39	.691 +	2.17
AA+	24	.779 +	2.04

*, + Quiz Three Estimates which were not significantly different (p=.05 per pair), using Feldt's Test (1969).

Validity: Q2-Q3. Each section of Art Appreciation took Quiz Two and Quiz Three using the same response/scoring method. The primary difference between the quizzes was specific content. Items had been constructed in the same manner and the tests were administered in the same way for both quizzes. The correlations obtained between Quiz Two and Quiz Three scores are presented in Table 37. For each estimate tested, under the null hypothesis rho=0, the probability level attained is given in the body of the table. Additionally, each possible pair of estimates was tested (using the Z transformation), under the null hypothesis rho1=rho2 (see Sokal and Rholf, 1969). Each pair of estimates not significantly different (p=.05 per pair) is indicated by a star and/or a double star.

Validity. Scores from Quiz Three were correlated with GPA, VSAT and MSAT scores. The results are presented in Table 38. Probability levels attained, under the null hypothesis rho=0, are given in the body of the table. Each possible pair of estimates was tested (using the Z transformation), under the null hypothesis rho1=rho2. Each pair of estimates which was not significantly different (p=.05 per pair) is indicated by a star and/or double star. The AUC and the CRS estimates were significantly different (p<.05).

TABLE 37

Validity Estimates: Q2-Q3

Section	N	Validity
CBS * **	37	.648 p=.0001
DRL *	39	.430 p=.0062
AUC **	35	.729 p=.0001
CRS *	41	.456 p=.0054
NR *	38	.466 p=.0032
AA-	39	.317 p=.0487
AA+ *	24	.467 p=.0213

p: attained under Ho: rho=0
*, **: n.s.d. per GPA pair,
under Ho: rho1=rho2.

TABLE 38

Validity Estimates: Q3

Section	N	GPA	N	VSAT	MSAT
CBS* **	33	.562	26	.193	.056
		p=.0007		p=.3359	=.7817
DRL* **	30	.383	23	.268	.356
		p=.0367		p=.2161	=.0956
AUC **	24	.222	20	.070	.214
		p=.2950		p=.7690	=.3655
CRS*	29	.680	26	.328	.559
		p=.0001		p=.1020	=.0030
NR * **	27	.383	27	.405	.032
		p=.0484		p=.0358	=.8704
AA-* **	33	.603	27	-.096	.480
		p=.1223		p=.7434	=.0823
AA+* **	14	.438	14	-.208	.065
		p=.0108		p=.2982	=.7482

p: attained under Ho: rho=0
*, **: n.s.d. per pair, under Ho: rho1=rho2

GPA: The most stable validity estimate for Quiz Three scores across the seven sections, GPA again was the highest for the Elimination format methods, CBS, CRS and AA-.

VSAT-MSAT: The estimates for both of these variables again fluctuated among sections, and the rectangular plots confirmed the absence of a relationship between these variables and quiz scores.

Item-Mark Totals

A form of item information recorded was a simple count of the number of items on which the various number of marks had been made; item mark totals for each response/scoring method quiz represent the total number of response instances possible for the number of examinees in the section marking the number of items in the quiz. For example, the 37 CBS examinees each responded to the 18 items of Quiz Two, and yielded a total of 666 possible item responding instances (see Table 43). Of these 666 instances, 280 represented items where the 4 marks appropriately did NOT include the correct choice; 83 did. There were 168 instances of items having 3 choices marked. For 126 of these, the correct choice was not among those marked, while for 42 it was.

For each response/scoring method, the number of item choices marked was tabulated for each quiz. The results are discussed by response/scoring method and are presented in Tables 43 through 48, which are extensions of Tables 19 through 25 (see Chapter III).

NR. Since the NR method does not advise against guessing, the instances identifying the correct choice or marking a distractor could have resulted from guessing (see Table 39).

AA. The Arnold methods give a uniform, low penalty to ALL inappropriately marked items. Differences between types of inappropriate marks were not observable from the item scores. Therefore, in order to obtain a count of the number of item marks for these types of marks, it was necessary to hand-tabulate examinee answer sheets. These tabulations, presented in parentheses in Table 40, show that examinee responding in the Arnold sections was very similar to that of the examinees in the CBS and DRL sections (see Tables 45 and 46). However, both CBS and DRL scoring incorporate individual item scores for each type of inappropriate mark, while the Arnold methods do not.

TABLE 39

NR Quiz Two and Quiz Three Item Totals

QUIZ TWO

	RWWWW	
431	m----	= 1 point
5	-----	= 0 points
248	---m-	= 0 points
684	Total	

QUIZ THREE

	RWWWW	
427	m----	= 1 point
11	-----	= 0 points
322	---m-	= 0 points
760	Total	

TABLE 40

AA Quiz Two and Quiz Three Item Totals

QUIZ TWO

		AA+	Points	AA-		
		RWWWW		RWWWW		
	300	m----	= 1.00 =	-mmmm	473	
	29	mm---	= 0.38 =	--mmm	74	
	4	mmm--	= 0.16 =	---mm	24	
	2	mmmm-	= 0.06 =	----m	7	
	14	mmmmm	= 0.00 =	-----	9	
	(1)	-mmmm	=-0.25 =	m----	(4)	
83	(9)	--mmm	=-0.25 =	mm---	(11)	115
	(28)	---mm	=-0.25 =	mmm--	(36)	
	(45)	----m	=-0.25 =	mmmm-	(64)	
	432	Total		Total	702	

QUIZ THREE

		AA+	Points	AA-		
		RWWWW		RWWWW		
	292	m----	= 1.00 =	-mmmm	510	
	23	mm---	= 0.38 =	--mmm	41	
	7	mmm--	= 0.16 =	---mm	21	
	4	mmmm-	= 0.06 =	----m	2	
	23	mmmmm	= 0.00 =	-----	21	
	(3)	-mmmm	=-0.25 =	m----	(5)	
131	(14)	--mmm	=-0.25 =	mm---	(21)	185
	(41)	---mm	=-0.25 =	mmm--	(57)	
	(73)	----m	=-0.25 =	mmmm-	(102)	
	480	Total		Total	780	

TABLE 41

CBS Quiz Two and Quiz Three Item Totals

QUIZ TWO

	RWWWW			
280	-mmmm	=	4	points
126	--mmm	=	3	points
66	---mm	=	2	points
21	----m	=	1	point
30	-----	=	0	points
2	m----	=	-4	points
22	mm---	=	-3	points
42	mmm--	=	-2	points
83	mmmm-	=	-1	point

666	Total			

QUIZ THREE

	RWWWW			
231	-mmmm	=	4	points
141	--mmm	=	3	points
67	---mm	=	2	points
32	----m	=	1	point
31	-----	=	0	points
4	m----	=	-4	points
22	mm---	=	-3	points
82	mmm--	=	-2	points
130	mmmm-	=	-1	point

740	Total			

CBS. The CBS examinees who marked the correct choice (R) either were misinformed or had made an unlucky guess. Tables 12, 14, 16 and 18 note that the misinformed examinee believes the correct choice is a distractor and is unaware of the misinformed basis for the mark. However, totally ignorant examinees who mark the correct choice are guessing, without regard for sequence of marking. For CBS examinees, marking additional distractors reduced the penalty and some examinees were observed using the strategy of marking 4 choices on a number of items.

DRL. The DRL examinee who did not include the correct choice among those marked either believed that it was a distractor or made an unlucky guess. Tables 11, 13, 15 and 17 note that the correct choice is believed to be incorrect by a misinformed examinee who, instead, marks a distractor believed to be correct. Additional choices marked were insurance or guesses, since choices known to be distractors were NOT marked.

AUC. Since with AUC, the correct choice inevitably was erased, each item total (by score) contains instances of guessing, misinformation and even erasing mistakes.

TABLE 42

DRL Quiz Two and Quiz Three Item Totals

QUIZ TWO

	RWWWW			
280	m----	=	4	points
118	mm---		3	points
47	mmm--	=	2	points
25	mmmm-	=	1	point
20	mmmmm	=	0	points
2	-mmmm	=	-4	points
22	--mmm	=	-3	points
70	---mm	=	-2	points
118	----m	=	-1	point

702	Total			

QUIZ THREE

	RWWWW			
292	m----	=	4	points
130	mm---		3	points
60	mmm--	=	2	points
33	mmmm-	=	1	point
46	mmmmm	=	0	points
2	-mmmm	=	-4	points
18	--mmm	=	-3	points
65	---mm	=	-2	points
134	----m	=	-1	point

780	Total			

TABLE 43

AUC Quiz Two and Quiz Three Item Totals

QUIZ TWO

	RWWWW	
410	m----	= 4 points
111	mm---	= 3 points
58	mmm--	= 2 points
32	mmmm-	= 1 point
19	-----	= 0 points

630	Total	

QUIZ THREE

	RWWWW	
435	m----	= 4 points
123	mm---	= 3 points
83	mmm--	= 2 points
46	mmmm-	= 1 point
13	-----	= 0 points

700	Total	

TABLE 44

CRS Quiz Two and Quiz Three Item Totals

QUIZ TWO

	RWWWW	
383	-mmmm	= 9 points
114	--mmm	= 8 points
39	---mm	= 7 points
10	----m	= 6 points
19	-----	= 5 points
49	m----	= 4 points
51	mm---	= 3 points
37	mmm--	= 2 points
36	mmmm-	= 1 point
738	Total	

QUIZ THREE

	RWWWW	
470	-mmmm	= 9 points
122	--mmm	= 8 points
26	---mm	= 7 points
9	----m	= 6 points
8	-----	= 5 points
48	m----	= 4 points
49	mm---	= 3 points
43	mmm--	= 2 points
45	mmmm-	= 1 point
820	Total	

CRS. The CRS examinees who erased the correct choice either believed that it was a distractor or made an unlucky guess. The CRS item count, by score, showed that feedback with Elimination responding let the examinee know immediately when the correct choice was erased; and the scoring inhibited further erasing.

The number of CRS instances of marks inappropriately including the correct choice among those marked was distributed very evenly across the four score-levels representing the values 4, 3, 2, and 1. However, the CBS and DRL item mark totals for each of the four negative score values show increasing numbers of marks as the negative scores becomes less severe (see Table 41).

Evaluation Questionnaire

After the last scheduled quiz and before the final exam, examinees were given an eight-question evaluation questionnaire (see Appendix B). The results are tabled by question.

Testing Experience: Questions 1a and 1b (see Table 45).

Multiple choice testing methods were familiar to 206 of the 232 examinees completing the questionnaire. Methods using multiple marks and rewarding partial information with variable credit had been used previously by 12 examinees.

TABLE 45

Evaluation Question 1

1a "Have you ever had a multiple-choice quiz prior to those given in this course?"

N	Section	Yes	No
40	NR	35	5
21	AA+	19	2
31	AA-	28	3
33	CBS	28	5
33	DRL	29	4
37	AUC	31	6
37	CRS	36	1
232	Total	206	26

1b "Have you ever had a multiple-choice quiz which made use of partial information?"

N	Section	Yes	No
40	NR	4	36
17	AA+	1	16
32	AA-	2	30
32	CBS	2	30
32	DRL	0	32
34	AUC	1	33
35	CRS	2	33
222	Total	12	210

TABLE 46

Evaluation Question 2

2 "Given a choice among multiple-choice methods, knowing what you know now, which would you prefer to use?"

		Methods	
N	Section	NR	Multiple-Mark
22	AA+	9	13
33	AA-	12	21
33	CBS	11	22
33	DRL	14	19
36	AUC	5	31
37	CRS	14	23
194	Total	65	129

Testing Preferences: Question Two (see Table 46).

Given a choice to use NR methods or multiple-mark methods, 31 (85%) of 36 AUC examinees would prefer AUC, while only 19 (57%) of the 33 DRL examinees would prefer DRL. A total of 129 examinees would prefer multiple-mark methods.

Studying: Questions 3A and 3B (see Table 47).

Of the 154 examinees stating that they had studied in the same way for a multiple-mark quiz as they would have for an NR quiz; 123 stated that they studied less than they would have for an NR quiz.

Study Habits: Question 4 (see Table 48).

Only 35 examinees indicated that concepts were the main objective of their studying. The majority emphasized facts or general information.

Responding: Questions 5 and 6 (see Table 49).

Examinees were in agreement regarding the use of multiple item marks to obtain partial item credit. AUC examinees were the most favorable toward multiple marking. Regarding the process of marking, however, 122 examinees were con-

TABLE 47

Evaluation Question 3

3a "When you knew one of the quizzes was going to make use of partial information, did you study:"

N	Section	Diff.than for NR Quiz	Same as for NR Quiz
19	AA+	3	16
29	AA-	3	26
29	CBS	3	26
30	DRL	7	23
37	AUC	5	32
36	CRS	5	31
180	Total	26	154

3b "When you knew one of the quizzes was going to make use of partial information, did you study:"

N	Section	More Than For an NR Quiz	Less Than For an NR Quiz	Same As For an NR Quiz
17	AA+	1	15	1
23	AA-	2	21	0
23	CBS	1	21	1
29	DRL	2	20	7
29	AUC	1	21	7
31	CRS	4	25	2
152	Total	11	123	18

TABLE 48

Evaluation Question 4

4 "When you knew one of the quizzes was going to permit use of partial information and scoring, did you emphasize in your studying:"

Section	Facts	Concepts	General
AA+	12	4	6
AA-	23	3	8
CBS	22	4	8
DRL	11	1	20
AUC	20	6	16
CRS	18	17	16
Total	106	35	74

TABLE 49

Evaluation Questions 5 and 6

<u>Rating Scale</u>

1=Strongly Agree 3=Disagree
2=Agree 4=Strongly Disagree

5 "It was good to know that more than one try could be made on a question and still receive credit."

N	Section	1	2	3	4	omit
22	AA+	2	12	5	3	0
33	AA-	5	21	2	4	1
37	CBS	5	21	3	3	5
33	DRL	7	23	1	1	1
37	AUC	18	16	2	0	1
37	CRS	9	22	4	2	0
199	Total	46	115	17	13	8

6 "I was concerned whether the marks I made were made appropriately (according to the directions)."

N	Section	1	2	3	4	Omit
22	AA+	2	12	8	0	0
33	AA-	6	15	10	1	1
34	CBS	7	13	11	1	2
33	DRL	0	24	8	0	1
37	AUC	3	13	12	6	3
37	CRS	4	23	7	0	3
196	Total	22	100	56	8	10

cerned that the marks they had made might have been made inappropriately as mistakes.

Effects: Questions 7 and 8 (see Table 50).

Examinee opinions were mixed regarding effects of the methods and directions when used the first time (for Quiz One); 165 agreed that the methods and the directions were easier to use the second time, with Quiz Three.

Observations

A journal of classroom observations kept for each section of Art Appreciation notes that in all sections the rooms were overheated and dry. Since this study was conducted during the winter quarter, the only cool ventilation available was from windows.

For all sections using multiple mark methods, examinee quizzes were scored by hand. The amount of time needed to score each examinee's quiz was between 10 and 15 minutes. Although test-scoring keys were developed for each section and helped in noting whether or not the correct choice had been marked, the amount of time required for scoring the quizzes was excessive.

TABLE 50

Evaluation Questions 7 and 8

Rating Scale
1=Strongly Agree 3=Disagree
2=Agree 4=Strongly Disagree

7 "The methods and directions needed more explanation and caused more anxiety the first time they were used than they were worth."

N	Section	1	2	3	4	omit
22	AA+	6	6	9	1	0
33	AA-	7	13	11	1	1
33	CBS	6	15	10	1	1
33	DRL	5	14	11	2	1
35	AUC	7	7	16	5	0
37	CRS	14	9	12	2	0
193	Total	45	64	69	12	3

8 "The methods and directions were easier to follow the second time they were used."

N	Section	1	2	3	4	omit
22	AA+	5	14	2	1	0
32	AA-	7	20	3	2	0
32	CBS	7	20	2	3	0
33	DRL	2	25	5	0	1
37	AUC	13	20	1	2	1
36	CRS	11	21	2	2	0
192	Total	45	120	15	10	2

Additional observations, presented by section, are followed by specific comments made by examinees in each section, who were encouraged to contribute comments.

<u>NUMBER RIGHT</u> (<u>NR</u>)

<u>Time</u>: <u>10:00 MWF</u>

There was no apparent anxiety or distress with the methods. Examinees were the most vocal of all examinees about the preparation of the quizzes being done by an "outsider," until they were informed that their teacher, a novice, had provided information for the quizzes. Specific comments follow:

1. This type of quiz will probably work well. It just takes getting used to. (This examinee had not used multiple choice methods before.)

2. The test--itself--is a good idea, but I feel it would have been to our benefit to have our own teacher write the test. When another person writes the test, they don't know the material that was emphasized in class.

3. These tests (multiple choice) were unfair and stupid. I had an A average in this class until I took these tests.

4. I think multiple choice tests are a big waste of time, and these quizzes should not have counted toward our course grade.

5. I haven't used this method before (multiple choice), and I'm glad I was introduced to it. But, I like the old methods better.

ARNOLD AND ARNOLD (AA-)

Time: 1 p.m. MWF

These examinees were curious about this research and some examinees suggested alternatives for the scoring. This group was communicative and positive in their attitude.

Many examinees were observed marking four choices to numerous quiz items. Examinees using such a strategy would get either the maximum credit or the uniform, low penalty. Specific comments follow:

1. I believe the system should have different scoring: 4 wrongs=5 points; 3 wrongs=2.5 points; 2 wrongs=1.5 points; 1 wrong=0.5 points; 0 wrong=0.0 points. Under the present scoring, I have a tendency not to use the system as much.
2. These quizzes were basic information, but the method was a hassle.

ARNOLD AND ARNOLD (AA+)

Time: 3 p.m. MWF

This section met at a time many students found undesirable. There were frequent questions regarding the quiz and responding. Examinees were openly hostile and questioned the purpose of tests and of the methods. Only about half of the examinees voluntarily agreed to permit access to their records. The other students refused. Specific comments follow:

1. I feel these were stupid and wasted my time. This way of grading has lowered my grade at least one letter grade. I did much better when the teacher gave these tests orally than this totally ridiculous way of doing it.

2. These tests only pulled my grade down. We didn't even cover the material in class.

3. I feel this way of scoring affected my grade negatively. This system allows the student to be more lax. It doesn't reinforce you as it should.

4. These tests really don't show what someone knows..only how well they are able to guess--I know from experience.

5. Multiple choice was more confusing only due to the fact that the amount of information needed to know was greatly increased.

<u>COOMBS</u> (<u>CBS</u>)

<u>Time</u>: <u>10:30</u> <u>a.m.</u> <u>T</u> <u>TH</u>

The CBS examinees appeared to understand the method. There was little frustration or apprehension evident and examinees expressed interest in the scoring.

Some responding strategies emerged on the quizzes. Examinees were observed marking four choices to many items. Such a practice would give the maximum credit of four points for appropriately marked distractors and the minimum penalty if the correct choice were among those marked. Specific comments follow:

1. I think that this method is much better; it helps in eliminating answers. You have a chance to consider all the answers and by taking more time, it is easier to find a correct answer.

2. Credit for partial information is helpful - nothing is more frustrating than taking a test and getting down to two possible answers, then deciding which one. However thinking in terms of "wrongs" takes practice. This may throw off results.

3. I like it and think it gives you a better chance to get points.

4. I think these tests are as hard as others because you tend not to be as sure of yourself in answering and thus you get confused.

5. It takes too long to fill in all the dots. I don't know any strategy in trying to put down what I am sure of or not sure of.

6. It would not be practical for SAT or ACT scoring.

7. The grading style is better than the regular type, for it gives the student the ability to let you know they even partially know what was on a test but couldn't remember everything.

8. A good method, but confusing after so many years of the opposite method.

9. It is somewhat difficult to reorient one's thinking, but probably offers a fairer chance to the average student.

10. Is appropriate to this type of class in which some students are not familiar with information. Not suitable for a chem or business class.

11. I wouldn't want to grade it! Besides, it makes me tense...I don't like it on an art test..maybe on some other kind. Art information is too frustrating.

12. I don't think it makes much difference, but it IS time consuming.

DRESSEL (DRL)

Time: 9 a.m. MWF

The DRL examinees were not apprehensive nor exceedingly cautious during the quizzes. There were not many questions regarding the method or the answer sheet. The examinees, however, appeared to use standard NR responding, prompting the author to reiterate the DRL directions. However, it became evident that the examinees were using a strategy and were aware that one mark per item would give the maximum credit if the correct choice were marked or the minimum penalty if it were not. Specific comments follow:

1. At times this was confusing, but I would like to see this instituted for the partial information. Very interesting.
2. I prefer either right or wrong, for it is more confusing to have to decide between several different answers. I prefer to make one choice.
3. The tests would have worked better if you had taught the class.

ANSWER-UNTIL-CORRECT (AUC)

Time: 8 a.m. MWF

There was little anxiety or apprehension observed with regard to the method. However, there was physical effort needed in order to erase item choices. A fresh eraser, supplied to each examinee, helped alleviate part of the problem caused by dry ink-shields covering the choice letters. Some choices did not erase well and it was decided that partly erased choices would count as unerased if the letter did not show through. Specific comments follow:

1. Partial credit was great!
2. If you knew something about the item in question, but not the answer, you still got credit for what you knew.
3. The tests only pulled my grade average down.
4. In my opinion, these tests were unfair. There was not adequate time to study for them. They should not have counted in our grade.

<u>CROSS</u> (<u>CRS</u>)

<u>Time</u>: <u>1:30 p.m. T TH</u>

The CRS method required examinees to erase a maximum of four choices per item. This required effort, and strained the patience of many examinees who became frustrated at dry ink-shields covering choice letters. Additionally, examinees were visibly distressed, as evidenced by verbal outbursts in some cases, when an item's correct choice was erased first or second. Some examinees verbally expressed frustration that they could not devise any logical responding strategy. Specific comments follow:

1. The tests were made to SEEM more difficult than they were.

2. You don't learn as much - and, it has an effect on you if you erase a RIGHT at first..it affects you the rest of the test.

3. Very different. It's like gambling--you worry more about the gambling than being right or wrong. It's okay after the first time, though.

4. I was scared after erasing a Right answer on the first try.

Overview. Examinee comments from all of the sections were reviewed for similarity of topic. Seven topics resulted. Presented below is a count of the number of examinees, by section, making a comment in reference to one of the topics.

1. Favor method using Partial Information:

   CBS: 4 DRL: 1 AUC: 2

2. Resent methods requiring Partial Information:

   CBS: 1 DRL: 1 CRS: 1 AA+: 2

3. Believe the methods adversely affect responding:

   CBS: 4 CRS: 3

4. Prefer that scores not count toward grades:

   NR: 1 AUC: 1

5. Blame method for lowering pre-existing grades:

   NR: 1 AA-: 1 AA+: 3 AUC: 1

6. Believe that method takes too much time:

   AA+: 1 CBS: 2

7. Observe that grading quizzes takes time:

<u>CBS</u>: <u>2</u>

CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS

The purpose of this study was to compare the results of identical tests which were administered, using seven different response/scoring methods, to students enrolled in sections of the same course. A limitation of this study was the difficulty in coordinating test items and dates for the seven intact sections. Additionally, the section randomly selected to serve as the NR control group was taught by a novice teacher who was being advised by the professor for the AA- and AA+ sections. Except for such limitations, the following conclusions are drawn and recommendations made:

1. <u>Conclusion</u>: <u>Item Responding</u>. In response to the Evaluation Questionnaire, several examinees expressed a lack of familiarity with multiple-mark methods and with the use of partial information. These examinees commented that it was difficult to re-orient their thinking to use partial information after so many years of seeking the correct choice to test items.

Recommendation: Examinees need considerable training and practice in gauging the extent of their information (Dressel, 1953) before using partial information. It is recommended that multiple mark responding methods be accompanied by training exercises and explanations of partial information, especially prior to use with tests intended for grading.

2. Conclusion: Item Responding. On the Evaluation Questionnaire, several examinees indicated a lack of experience with multiple-choice tests. Additionally, verbal comments by many examinees suggested that they were apprehensive about using machine-scorable answer forms. The erasable answer cards also posed some problems in that the ink-shields covering item choice letters often were too dry to erase properly.

Recommendation: Test instruction or training should include emphasis on the use of the answer form being used, so that marks made by examinees are intentional rather than mistakes. Erasable answer cards should be preserved in foil until use to prevent the ink-shields from drying out.

3. Conclusion: Item Responding. Examinee use of the erasable answer card with the CRS method caused erasing-fatigue, since the examinee had to erase a maximum of four distractors for each test item.

Recommendations: The CRS method should be used with a less strenuous answer form such as latent-image so that item responses reflect levels of examinee information and do not incorporate fatigue.

4. Conclusion: Item Responding. Item mark totals provide instructors with feedback regarding examinee performance for each different level of information. This feedback can help instructors identify areas of content needing revision, or test items needing rewording.

Recommendations: Where there is concern regarding the amount of time needed to score a multiple-mark test, student proctors from the class can be used. Computerized methods for multiple-mark testing can eliminate hand-scoring entirely.

5. Conclusion: Reliability Estimates. Although there were no significant differences between any of the possible pairs of Quiz Two estimates, internal consistency reliability was higher for the Inclusion for methods (DRL and AA+) than for Elimination methods. It was more concluded that examinee responding perhaps was consistent for methods most similar to the familiar NR. Examinees in the elimination format sections occasionally may have reverted their thinking, through habit, to seeking the correct choice. Such a practice immediately would be penalized by the Elimination format and would have yielded an item penalty resulting from a mistake in responding rather than from misinformation or guessing.

For Quiz Three, reliability estimates for the Inclusion methods DRL, AA+ and AUC were significantly different than their Elimination counterparts (p=.05). The Elimination format, confusing to examinees, perhaps resulted in examinee responding mistakes that depressed internal consistency reliability. Additionally, responding strategies used by many examinees perhaps contributed to guessing variance.

Without further study, there is no reason to conclude that any one method or format is any more or less reliable than another. The differences in reliability estimates were not sufficient to warrant the considerable effort expended in administering and hand-scoring multiple-mark tests.

Recommendation: Examinees need considerable training in the use of Elimination format methods before test results should be expected to reflect true levels of information. Further evidence of responding mistakes and the use of strategies should be gathered in order to identify the presence of these components in item and test scores.

6. Conclusion: Validity Estimates. The sections did not differ significantly in GPA, VSAT or MSAT performance, based on the ANOVA performed on each variable. Quiz Two and Three estimates of validity with GPA were higher for the sections using Elimination format methods (CBS, CRS and AA-) and were lower for the Inclusion sections (DRL, AUC, and AA+), although only the AUC and CRS sections were significantly different (p=.05).

The higher Cronbach's alpha estimates of internal consistency reliability obtained for the Inclusion methods should be expected to yield lower validity estimates from correlations with a heterogeneous variable such as GPA.

Without further study on the topic, there was no reason to conclude that Elimination methods were inherently any more or less valid than Inclusion methods. Although it had been expected that item scores from the different response/scoring methods would have an effect on correlations with such variables as VSAT and MSAT, the correlations fluctuated erratically, and the rectangular plots of these variables with scores from each Quiz confirmed that little if any relationship existed between Quiz Scores and these variables. The differences in validity estimates were not sufficient to justify the effort of administering and hand-scoring multiple-mark quizzes.

Recommendation: The use of other criterion-related variables is suggested for future studies of response/scoring alternatives. These variables should include measures of risk-taking, test-wiseness and locus of control.

Discussion

Observed multiple-choice item and test scores were used to calculate estimates of reliability and validity for quiz scores. Under the different response/scoring methods used, many outcomes were possible for the items of a quiz and, thereby, for the total scores. For example, some methods exacted a uniform penalty across all levels of misinformation while other methods imposed a penalty which depended upon the level of misinformation observed. Thus, the total-score of an examinee who was misinformed on some items could have been very different from the score of an examinee who demonstrated a different level of misinformation on the same items, or from the score of an examinee who was totally ignorant regarding those same items. Similarly, an examinee who was misinformed on certain test items could have received item-scores with one method which could have differed considerably from the item scores received with another method.

The different response/scoring methods provide incentives to the examinee to respond to an item from various levels of information. However, these methods vary considerably in the extent to which item scores for these levels of information are incorporated into the computation of total scores and, subsequently, estimates of test score validity. The

CBS, DRL and CRS methods permit the examinee to respond to an item from all the levels previously discussed. Additionally, these methods impose penalties for inappropriate marks or erasures, thereby providing an effective control on guessing. Item and test scores obtained from these methods incorporate all the levels of information in their scoring schemes. Reliability should be higher for these methods than for methods which do not. Higher reliability estimates could be expected to result in higher validity estimates, regardless of the utility of the variables concerned. For this study, GPA, VSAT and MSAT were used and, although not significantly different among sections (see Table 27 and Table 28), the influence of scoring method on subsequent validity estimates with these variables was of interest. The different response/scoring methods were considered for their impact on item responses, item scores and on resulting estimates of reliability and validity.

For this study, seven different response/scoring methods were compared in an academic setting with students enrolled in a required undergraduate art appreciation course at a medium-sized state university. The number of correct responses, based on a single choice per item, constituted the score on a preliminary test in each section of 35 to 40 students. Then, a distinct response/scoring method was randomly assigned to each section for two additional tests. One section continued to use number-right scoring and six sections used methods requiring or permitting multiple marks per item. All test scores counted toward course grades and examinees were so informed. Responses were used to determine estimates of internal consistency reliability and of validity with the previous quarter's grade-point average and subscores of the Scholastic Aptitude Test. An Evaluation Questionnaire, administered in each section, obtained for each response/scoring method self-report information about examinee study habits, testing preferences and experience, and responding behaviors. Observations of examinee behavior, subjective in nature, helped explain or confirm various empirical findings from this study or from the literature.

Based on the findings from this study, it was concluded that:

1. observed differences in estimates of reliability or validity were not sufficient to justify the effort expended in administering and hand-scoring multiple-mark tests.

2. examinees experience substantial difficulty becoming familiar with response/scoring methods which permit multiple marks. These methods require more than casual explanation and practice before examinees become adept in their use.

3. Item mark totals from methods scoring all the levels of information can provide more feedback to instructors regarding the performance of examinees on specific test items than could be acquired from the NR method. This feedback can help the

instructor identify areas of content needing revision or test items needing rewording. When the examinees receive feedback, they have the opportunity to observe, score and learn from their own item performance, Pressey's initial consideration in 1950.

BIBLIOGRAPHY

Angell, G. W. The effect of immediate knowledge of quiz results on final exam scores in freshman chemistry. Journal of Educational Measurement, 1949, 42, 391-394.

Arnold, J. C. and Arnold, P. L. On scoring multiple choice exams allowing for partial knowledge. Journal of Experimental Education. 1970, 39(1), 8-13.

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T. Statistical Analysis System: User's Guide. Raleigh, NC: SAS Institute, 1979.

Collet, L. S. Elimination scoring: An empirical evaluation. Journal of Educational Measurement, 1971, 8(3), 209-214.

Coombs, C. H. On the use of objective examinations. Educational and Psychological Measurement, 1953, 13, 308-310.

Coombs, C. H., Milholland, J. E., and Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cross, L. H. and N. J. Thayer. A new method for administering and scoring multiple-choice tests: Theoretical and empirical considerations. Paper presented at the Annual NCME Convention, Boston. April 8-10, 1980.

Cross, L. H. An investigation of a scoring procedure designed to eliminate score variance due to guessing on multiple-choice tests. Unpublished doctoral dissertation, University of Pennsylvania, 1973.

Cross, L. H., and Frary, R. B. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14(4), 313-321.

Davis, F. B. Use of correction for chance success in test scoring., Journal of Educational Research, 1959, 52, 279-280.

Davis, F. B. Educational Measurements and Their Interpretation. Belmont, California. Wadsworth, 1964.

Davis, F. B. A note on the correction for chance success. The Journal of Experimental Education, 1969, 35, 42-47.

De Finetti, B. Methods for discriminating partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18 87-123.

Dressel, P. L., and Schmid, P. Some modifications of the multiple-choice test item. Educational and Psychological Measurement, 1953, 13, 574-595.

Duncan, D. B. Multiple range and multiple F tests. Biometrika, 1955, 11, 1-142.

Evans, R. M., and Misfeldt, K. Effect of self-scoring procedures on test reliability. Perceptual and Motor Skills, 1974, 38, 1248.

Educational Testing Service. About Your PSAT/NMSQT Scores, 1978.

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

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. Educational and Psychological Measurement, 1969, 29, 359-365.

Frary, R. B., Cross, L. H., and Lowry, S. R. Random guessing, correction for guessing, and reliability of multiple choice test scores. Journal of Experimental Education, 1975, 11-15.

Frary, R. B. The effect of misinformation, partial information, and guessing on expectred multiple-choice item scores. Applied Psychological Measurement, 1980, 4(1), 29-90.

Gilman, D. A., and Ferry, P. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 1972, 9(3), 205-207.

Gritten, F. and Johnson, D. M. Individual differences in judging multiple choice questions. Journal of Educational Psychology, 1941, 32, 423-430.

Hanna, G. S. A study of reliability and validity effects of total and partial immediate feedback in multiple-choice testing. Journal of Educational Research, 1977, 14(1), 1-7.

Horst, P. The chance element in the multiple-choice item. Journal of General Psychology, 1932, VI, 209-211.

Kaess, W. and Zeaman, D. Positive and negative knowledge of results on a Pressey-Type punchboard. Journal of Experimental Psychology, 1960, 60(1), 12-17.

Kogan, N., and Wallach, M. A. Risk Taking: A Study in Cognition and Personality. New York: Holt, Rinehart and Winston, 1964.

Kramer, C. Y. Extension of multiple range tests to group means with unequal numbers of replication. Biometrics, 12, 1956, pp 307-310.

Kramer, Clyde Y. A First Course in Methods of Multivariate Statistics. Blacksburg: Clyde Y. Kramer, 1972.

Lindquist, E. F. (Ed.) Educational Measurement. Washington D. C. American Council on Education, 1951.

Lowry, S. R. The effect of luck and misinformation on the discrepancy between muitiple-choice test scores and true ability. Unpublished doctoral dissertation, 1975.

Mead, A. R., and Smith, B. M. Does the true-false scoring formula work? Some data on an old subject. Journal of Educational Research, 1957, 51, 47-53.

Preble, D. Artforms. New York: Harper & Row, Publishers, Inc. 1978.

Pressey , S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. Journal of Psychology. 1950, 29, 417-447.

Rowley, G. L., and Traub, R. E. Formula scoring, number right scoring, and test-taking strategy. Journal of Educational Measurement, 1977, 14, 15-22.

Sabers, D. L., and Feldt, L. S. An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. Journal of Educational Measurement, 1968, 5, 351-358.

Sax, G., and Collet, L. The effect of differing instructions and guessing formulas on reliability and validity. Educational Psychological Measurement, 1968, 28, 1127-1136.

Sheriffs, A. E., and Boomer, D. S. Who is penalized by the penalty for guessing? Journal of Educational Psychology, 1954, 45, 81-90,

Slakter, M. J. The penalty for not guessing. Journal of Educational Measurement. 1968, 5, 141-144.

Sokal, R. R., and Rohlf, F. J. Biometry. W. H. San Francisco: Freeman and Company, 1969.

Votaw, D. F. The effect of do-not-guess directions upon validity of true-false and multiple-choice tests. Journal of Educational Psychology, 1936, 27, 698-703.

Waters, L. K. Effect of perceived scoring formula on some aspects of test performance. Educational and Psychological Measurement, 1967, 27, 1005-1010.

Appendix A

QUIZ INSTRUCTIONS

NR Instructions

This is a multiple-choice test. Each question has only one correct choice, which you are to mark on the attached answer sheet. Your score on this test will be the number of questions you answer correctly--the Number Right. This particular test will be Machine Scored, so please use a #2 pencil, and erase answer-changes carefully. Print your name, I.D. or Social Security number where indicated on the answer sheet.

AA- Instructions

This is a multiple-choice quiz. Each question has five choices. For each question, only one choice is correct and four choices are incorrect.

Additionally, this quiz uses a machine scorable answer sheet. The questions each have CHOICES numbered 1, 2, 3, 4, and 5. This quiz also uses a different method of answering each question by permitting you to mark more than one choice per question. In particular, you are to mark choices which you KNOW to be INCORRECT. For instance, if you KNOW the correct choice to a question, you ALSO know the four incorrect choices, and can mark these. If you do NOT know the correct choice to a question, you can still get credit for what you DO know because you can darken as many of the INCORRECT choices as you know. Here are two simple examples. See how you do, using the answer sheet mark as many of the incorrect choices that you know: (These do NOT count on your score, of course.)

Example 1 For question #1 on the answer sheet, darken only the OBVIOUSLY incorrect choices to this question:

Who is buried in Grant's Tomb?

CHOICES

1. Grant 2. Lee
3. Jackson 4. Burnside
5. Turner

Example 2 For question #2 on the answer sheet, darken as many choices as you believe are incorrect:

Who designed the Statue of Liberty?

CHOICES

1. Le Corbusier 2. Joe Smith
3. Antoine Lavoisier 4. Auguste Bertholdi
5. Jacques Lartigue

For this quiz, you are to mark the choice(s) which you believe are INCORRECT and not INCLUDE the correct choice. It is NOT in your favor if the choice(s) you darken do INCLUDE the correct choice. The scoring rule illustrates this point:

AA- Scoring Rule

Number of Choices Marked (Does NOT include correct)	Points (Item Score)
4	1.00
3	0.38
2	0.16
1	0.06
0	0.00
(Includes correct choice)	
1 to 4	-0.25

As you can see, you can get credit for the choices you mark, but you also could--instead--get a penalty if the choices you mark INCLUDE the correct choice.

AA+ Instructions

This is a multiple-choice quiz. Each question has five choices. For each question, only one choice is correct and four choices are incorrect.

Additionally, this quiz uses a machine scorable answer sheet. The questions each have CHOICES numbered 1, 2, 3, 4, and 5. This quiz also uses a different method of answering each question by permitting you to mark more than one choice per question. For instance, if you KNOW the correct choice to a question you should darken--with pencil--ONLY that choice on the answer sheet. However, if you are unsure of the correct choice, you can darken as MANY choices to the item as you believe would surely INCLUDE the correct choice. Here are two simple examples. See how you do, using the answer sheet and as many marks as needed to include the correct choice. (These do NOT count on your score, of course.)

Example 1 For question #1 on the answer sheet, darken only the OBVIOUS right choice to this question:

Who is buried in Grant's Tomb?

CHOICES

1. Grant 2. Lee

3. Jackson 4. Burnside

5. Turner

Example 2 For question #2 on the answer sheet, darken as many choices as you believe would INCLUDE the correct choice to this question:

Who designed the Statue of Liberty?

CHOICES

1. Le Corbusier 2. Joe Smith

3. Antoine Lavoisier 4. Auguste Bertholdi

5. Jacques Lartique

For this quiz, you are to mark the choice(s) which you believe necessary to INCLUDE the correct choice. It is NOT in your fav or to darken ALL of the choices, but it is LESS in your favor if the choice(s) you darken do NOT include the correct choice. The scoring rule illustrates this point:

AA+ Scoring Rule

Number of Choices Marked (Includes correct choice)	Points (Item Score)
1	1.00
2	0.38
3	0.16
4	0.06
5	0.00
(Does NOT include correct)	
1 to 4	-0.25

As you can see, you can get credit for the choices you mark, but you also could--instead--get a penalty if the choices you mark do NOT include the correct choice.

CBS Instructions

This is a multiple-choice quiz. Each question has five choices. For each question, only one choice is correct and four choices are incorrect.

Additionally, this quiz uses a machine scorable answer sheet. The questions each have CHOICES numbered 1, 2, 3, 4, and 5. This quiz also uses a different method of answering each question by permitting you to mark more than one choice per question. In particular, you are to mark choices which you KNOW to be INCORRECT. For instance, if you KNOW the correct choice to a question, you ALSO know the four incorrect choices, and can mark these. If you do NOT know the correct choice to a question, you can still get credit for what you DO know because you can darken as many of the INCORRECT choices as you know. Here are two simple examples. See how you do, using the answer sheet mark as many of the incorrect choices that you know: (These do NOT count on your score, of course.)

Example 1 For question #1 on the answer sheet, darken only the OBVIOUSLY incorrect choices to this question:

Who is buried in Grant's Tomb?

CHOICES

1. Grant 2. Lee
3. Jackson 4. Burnside
5. Turner

Example 2 For question #2 on the answer sheet, darken as many choices as you believe are incorrect:

Who designed the Statue of Liberty?

CHOICES

1. Le Corbusier 2. Joe Smith
3. Antoine Lavoisier 4. Auguste Bertholdi
5. Jacques Lartigue

For this quiz, you are to mark the choice(s) which you believe are INCORRECT and not INCLUDE the correct choice. It is NOT in your favor if the choice(s) you darken DO INCLUDE the correct choice. The scoring rule illustrates this point:

CBS Scoring Rule

Number of Choices Marked	Points
(Does NOT include correct)	(Item Score)
4	4.00
3	3.00
2	2.00
1	1.00
0	0.00
(Includes correct choice)	
1	-4.00
2	-3.00
3	-2.00
4	-1.00

As you can see, you can get credit for the choices you mark, but you also could--instead--get a penalty if the choices you mark INCLUDE the correct choice.

DRL Instructions

This is a multiple-choice quiz. Each question has five choices. For each question, only one choice is correct and four choices are incorrect.

Additionally, this quiz uses a machine scorable answer sheet. The questions each have CHOICES numbered 1, 2, 3, 4, and 5. This quiz also uses a different method of answering each question by permitting you to mark more than one choice per question. For instance, if you KNOW the correct choice to a question you should darken--with pencil--ONLY that choice on the answer sheet. However, if you are unsure of the correct choice, you can darken as MANY choices to the item as you believe would surely INCLUDE the correct choice. Here are two simple examples. See how you do, using the answer sheet and as many marks as needed to include the correct choice. (These do NOT count on your score, of course.)

Example 1 For question #1 on the answer sheet, darken only the OBVIOUS right choice to this question:

Who is buried in Grant's Tomb?

CHOICES

1. Grant 2. Lee
3. Jackson 4. Burnside
5. Turner

Example 2 For question #2 on the answer sheet, darken as many choices as you believe would INCLUDE the correct choice to this question:

Who designed the Statue of Liberty?

CHOICES

1. Le Corbusier 2. Joe Smith
3. Antoine Lavoisier 4. Auguste Bertholdi
5. Jacques Lartigue

For this quiz, you are to mark the choice(s) which you believe necessary to INCLUDE the correct choice. It is NOT in your fav or to darken ALL of the choices, but it is LESS in your favor if the choice(s) you darken do NOT include the correct choice. The scoring rule illustrates this point:

DRL Scoring Rule

Number of Choices Marked	Points
(Includes correct)	(Item Score)
1	4.00
2	3.00
3	2.00
4	1.00
0 or 5	0.00
(Does NOT include correct)	
4	-4.00
3	-3.00
2	-2.00
1	-1.00

As you can see, you can get credit for the choices you mark,--but, you could--instead--get a penalty if the choices you mark do NOT include the correct choice.

AUC Instructions

This is a multiple-choice quiz. Each question has five choices. For each question, only one choice is correct and four choices are incorrect.

Additionally, this quiz uses a rather different type of answer sheet. Look at the answer sheet. The numbered items each have five circles of black, smudgy ink. These ink circles each have a yellow letter underneath. When the black ink is erased, the yellow letter will show through. Use the eraser to erase all of the black circles for question #1.

Under choices a, b, c, d, and e respectively. These same letters (I, T, H, E, and L) are under the choices of ALL the questions, but are in different order. To see this, use your eraser to erase ALL of the circles for question #2.

You should see the yellow letters L, E, I, T, and H under question #2 choices a, b, c, d, and e respectively.

In preparing this test, the correct choice for EACH question has been keyed to be with the letter "T"--when erased. Here is a simple example. For question #3 on the test form, erase only the correct choice to this question:

Example 1

Who is buried in Grant's Tomb?

CHOICES

1. Grant 2. Lee
3. Jackson 4. Burnside
5. Turner

Example 2

Who designed the Statue of Liberty?

CHOICES

1. Le Corbusier 2. Joe Smith
3. Antoine Lavoisier 4. Auguste Bertholdi
5. Jacques Lartigue

For this quiz, you are to erase the choice(s) until you find the correct choice. It is NOT in your favor to erase ALL of the choices.

AUC Scoring Rule

Number of Choices Erased Including Correct	Points
1	4.00
2	3.00
3	2.00
4	1.00
5	0.00

As you can see, you can score your test yourself by giving each item the points it deserves, based on the number of erasures you made to get to the correct choice.

CRS Instructions

For this quiz you will use a different kind of answer sheet and a different method of responding. The scoring is different from any you may have used before. The Answer Sheet is a Card:

> The answer card has room for 40 questions, each of which can have 5 choices which are labeled a, b, c, d, and e.
>
> Each choice is covered with black, erasable ink under which are yellow letters: L, T, H, E, and I.
>
> By using this card you will be able to SEE how you are performing. The quiz has been constructed in such a way that the correct choice is always keyed with the letter "H" and the incorrect choices are keyed with "L, T, E, and I."

The Responding Method:

For this quiz, you are to erase only those choices to a question which you believe are incorrect, taking care NOT to erase the correct choice.

The Scoring Method:

You will receive points, as follows, for EACH QUESTION:

1. Two (2) points for each INCORRECT choice erased AND
2. One (1) point for each choice left unerased. BUT...
3. If you erase the CORRECT choice ("H") you LOSE credit for ALL incorrect choices already erased!

To see how this works, suppose you definitely KNOW choices choices "a" and "b" are INCORRECT, but aren't sure about "c," "d," and "e." You erase the two Wrongs, to receive 7 points:

```
R W W W W
a b c d e
- I L - -
```

1+2+2+1+1= 7 points

Where R is for the correct, or Right choice, and the W's are for incorrect, or Wrong choices.

Now, suppose on that another question you know that "a," "b" and "c" are incorrect, and erase them. You aren't sure about "d" and "e" but decide to guess that "e" is also incorrect. Unfortunately, "e" turns out to be the correct answer and your score becomes 1 point:

$$\begin{array}{ccccc} \frac{R}{a} & \frac{W}{b} & \frac{W}{c} & \frac{W}{d} & \frac{W}{e} \\ L & T & E & & H \end{array}$$

0+0+0+1+0= 1 point

Of course, you would not erase "d" now, knowing that you would LOSE that 1 point credit!! If you had STOPPED erasing and NOT GUESSED, your score would have been 8 points:

$$\begin{array}{ccccc} \frac{R}{a} & \frac{W}{b} & \frac{W}{c} & \frac{W}{d} & \frac{W}{e} \\ & L & T & E & \end{array}$$

1+2+2+2+1= 8 points

Therefore, erase only choices you feel sure are incorrect. If you don't KNOW any choices, it is to your favor to guess, erasing just one choice.

Here are two simple examples. See how you do, using the answer card, erase only the incorrect choices.

<u>Example 1</u>: For question #1 on the answer card, erase only the OBVIOUSLY incorrect choices to this question:

Who is buried in Grant's Tomb?

<u>CHOICES</u>

1. Grant 2. Lee
3. Jackson 4. Burnside
5. Turner

<u>Example 2</u>: For question #2 on the answer card, erase only choices as you believe are incorrect:

Who designed the Statue of Liberty?

<u>CHOICES</u>

1. Le Corbusier 2. Joe Smith
3. Antoine Lavoisier 4. Auguste Bertholdi
5. Jacques Lartigue

For this quiz, you are to erase the choice(s) which you believe are INCORRECT and not INCLUDE the correct choice. It is NOT in your favor if the choice(s) you erase DO INCLUDE the correct choice. The scoring rule illustrates this point:

CRS Scoring Rule

Number of Choices Erased	Points
(Does NOT include correct)	(Item Score)
4	9.00
3	8.00
2	7.00
1	6.00
0	5.00
(Includes correct choice)	
1	4.00
2	3.00
3	2.00
4	1.00

As you can see, you can get credit for the choices you erase.

Appendix B

EVALUATION INSTRUMENT

Anonymous Opinions Please

1a. Have you ever had a multiple-choice quiz before those given in this course?

1. YES 2. NO

1b. Have you ever had a multiple-choice quiz which made use of partial information prior to this course?

3. YES 4. NO

2. Given a choice among multiple-choice methods, knowing what you know now--which would you prefer to use:

1. Number Right methods

2. Partial Information methods

3. When you knew one of the quizzes was going to make use of partial information, did you study: (Mark 1 choice in 'a' and 1 in 'b'.)

a. (1. differently than for a
(number-right quiz.
(2. about the same way as
for a number-right quiz.

b. (3. more than you would have
(for a number-right quiz.
(4. about the same amount as
(for a number-right quiz.
(5. less than you would have
(for a number-right quiz.

4. When you knew one of the quizzes was going to use partial information marking and scoring, did you did you emphasize in your studying: (Mark no more than 2 of the 3 choices.)

1. FACTS: to be able to recognize, for example names of artists and their works, or definitions of procedures, etc.

2. CONCEPTS: to be able to identify, for example, the rationale and events which led to Dadaism, or

to be able to distinguish a Dadaist from a Surrealist work of art.

4. 3. GENERALITIES: to be able to identify for example some of the wrong choices to test questions.

For items 5 to 8 below, mark one choice which best describes your opinion. When you were USING a method which permitted Partial Information responding and scoring do you:

<u>CHOICES</u>

1. strongly agree
2. agree
3. disagree
4. strongly disagree

5. It was nice to know that more than one try could be made on a question and that credit could still be received.

6. It was a concern wondering whether the marks I made were made appropriately, according to the directions.

7. The methods and directions needed more explanation and caused more anxiety the first time they were used than they were worth.

8. The methods and directions were easier to follow the second time they were used.

If you have any other comments you would care to make concerning these quizzes or the use of partial information response/scoring methods, please write these in the box at the upper right on the answer sheet or on another piece of paper. Please be anonymous and don't sign your name or id number.

The two page vita has been removed from the scanned document. Page 1 of 2

A COMPARISON OF MULTIPLE CHOICE TEST RESPONSE MODES AND SCORING METHODS

by

Nina J. Thayer

(ABSTRACT)

This study reports the comparison of seven different response/scoring methods used with multiple-choice tests in an academic setting with students enrolled in a required undergraduate art appreciation course at a medium-sized state university. The number of correct responses, based on a single choice per item, constituted the score on a preliminary test in each section of 35 to 40 students. Then, a distinct response/scoring method was randomly assigned to each section for two additional tests. One section continued to use number-right scoring and six sections used methods requiring or permitting multiple marks per item. All test scores counted toward course grades and examinees were so informed. Responses were used to determine estimates of

internal consistency reliability and of validity with the previous quarter's grade-point average and subscores of the Scholastic Aptitude Test. An Evaluation Questionnaire, administered in each section, obtained for each response/scoring method self-report information about examinee study habits, testing preferences and experience, and responding behaviors. Observations of examinee behavior, subjective in nature, helped explain or confirm various empirical findings from this study or from the literature.

Based on the findings from this study, it was concluded that:

1. observed differences in estimates of reliability or validity were not sufficient to justify the effort expended in administering and hand-scoring multiple-mark tests.

2. examinees experienced substantial difficulty becoming familiar with response/scoring methods which permit multiple marks. These methods require more than casual explanation and practice before examinees become adept in their use.

3. Item mark totals from methods scoring all the levels of information can provide more feedback to instructors regarding the performance of examinees on specific test items than could be acquired from the NR method. This feedback can help the instructor identify areas of content needing revision or test items needing rewording. When the examinees receive feedback, they have the opportunity to observe, score and learn from their own item performance, Pressey's initial consideration in 1950.