# AN INVESTIGATION OF THE RELATIONSHIPS OF TEST CHARACTERISTICS AND PERSONALITY VARIABLES TO PARTIAL INFORMATION AND MISINFORMATION IN MULTIPLE-CHOICE TEST SCORES

by

Mary Burnette Giles

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:

L. H. Cross, Chairman

R. B. Frary

T. M. Sherman

W. R. Pirie

L. M. Wolfle

August, 1979

Blacksburg, Virginia

ACKNOWLEDGEMENTS

I wish to thank the members of my committee for their help and support during all stages of this study. I am especially grateful to Dr. Lawrence H. Cross, the chairman of my committee, and Dr. Robert B. Frary for encouraging me to pursue this line of research. I appreciate the efforts of Dr. Walter R. Pirie, particularly in the early stages of my work, and the help of Dr. Lee M. Wolfle in interpreting the regression analyses. I also wish to note especially the understanding and assistance which I received from Dr. Thomas M. Sherman, for whom I worked during my graduate study and who provided me with access to the students from whom the data were gathered.

I wish also to acknowledge the assistance of with technical problems in scoring the attitude and personality instruments. Also important to me was the encouragement and support which I received from my fellow graduate students, particularly and .

Finally, I wish to express my deep gratitude to my husband, , and to my children, and . Had they done any less or any differently this work could not have been completed.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

# CHAPTER I

# THE PROBLEM

## Statement of the Problem

The introduction into multiple-choice testing of response modes and scoring rules which require more from examinees than simply marking a single best choice of an answer to a multiple-choice item was intended to extract from the test responses greater amounts of information about examinees' knowledge and, by doing so, enhance test score characteristics. It was believed that the additional information in the alternative scores would differentiate more sensitively among examinees, decrease the error variance in the test scores, and result in improved test reliability and validity. Empirical studies comparing alternative scores to conventional number-right scores have generally shown a modest increase in reliability and negligible changes in validity.

As soon as one moves away from conventional number-right scoring, however, the additional information acquired about examinees' knowledge places upon them the necessity of making more decisions about the multiple-choice items. One of the earliest departures from conventional number-right scoring was the imposition of a "correction-for-guessing" on multiple-choice tests. Ethical considerations require that examinees be instructed to answer all questions *unless* an answer would represent a sheer guess (Davis, 1967). However, what constitutes a sheer

guess for one examinee may represent "playing a hunch" to another examinee. It cannot be assumed that the basis on which all examinees decide to omit an item or to guess is the same, nor that their decisions are accurate reflections of the degree of information and partial information involved in the responses. This notion has been demonstrated by previous research which has shown that, when forced to guess at questions previously left blank under corrected-for-guessing directions, examinees are, on the average, able to guess the correct choice more frequently than would be expected by chance alone (Cross & Frary, 1977; Sherriffs & Boomer, 1954). Moreover, it has been shown that the tendency to refrain from "playing a hunch" may be related to personality differences (Sherriffs & Boomer, 1954; Swineford, 1938; Ziller, 1957).

Thus, the question of concern is whether the increases in reliability accompanying alternative response modes derive, not from the ability of the test to discriminate more precisely, but rather from reliable but extraneous personality differences which are introduced by the alternative procedures. The evidence for this mediating effect of personality is, at present, most clear for confidence-weighting procedures, in which examinees attach numerical weights to their answers, indicating their confidence in the correctness of their responses. Because the choice of these weights requires a subjective appraisal of confidence in one's level of knowledge, it may well be that the expressed confidence reflects personality differences as well as individual differences in the domain measured (Jacobs, 1975).

The alternative response mode of particular interest to this study was suggested by Coombs (1953), and appears to require subjective appraisals similar to those involved in confidence-weighting. The Coombs mode asks examinees to mark all item options which they believe to be *incorrect*, taking care *not* to mark the right answer. A credit of +1 points is awarded for each incorrect options marked; a penalty of $-(n - 1)$ is imposed if the right answer is marked, where $n$ is the number of options per item. Item scores are a sum of the credits and penalty from the correct and incorrect marks. Thus, when misinformation causes an examinee to identify the correct choice as a distractor, a score penalty is imposed, but the penalty is partially offset if, on the same item, partial information has enabled the examinee to identify one or more distractors as incorrect answers. The total test score is a sum of the item scores.

Since the information which is extracted from the responses through alternative procedures ostensibly results from the display of partial information and levels of misinformation, it seemed reasonable to believe that the anticipated increases in reliability would reflect this additional information. It was also possible, however, that personality characteristics would come into play when examinees complied with a set of test directions that would permit evaluation of their partial information and misinformation. The Coombs response mode offers the advantage of allowing total test scores to be partitioned into separate groups of items reflecting complete information, levels of partial information, ignorance, levels of partial misinformation, and

complete misinformation.

The purpose of this study was to use the sensitivity to levels of information and misinformation provided by the Coombs response mode in an examination of the mechanism by which test reliability and validity and personality differences among examinees are affected by an alternative response mode and scoring procedure. This was done by mathematical manipulation of the test scores, based on the score components identified by the Coombs response mode, with the expectation that such manipulation would delineate the particular contribution of each score component to test characteristics, and to demonstrate the differential contribution of personality variables to the expression of the score components.

## Examination of the Effects of Partial Information and Misinformation on Reliability and Validity

Under conventional number-right scoring, partial information cannot be directly expressed. Examinees use their partial information to increase their chances of guessing correctly by eliminating from consideration options which they know to be wrong within an item. Thus, when examinees guess correctly more frequently than is expected by chance from guessing randomly among the entire set of options for an item, they are assumed to be acting from partial information. Score credit is given, however, only when the guess is correct.

Under the Coombs response mode and scoring rule, partial information is directly expressed, and credit is awarded for each bit of

knowledge which an examinee expresses. Theoretically, then, guessing is eliminated, or at least severely curtailed, because the opportunity to express all pieces of information and to receive credit for them in test scores reduces the need to guess.

Using five-option multiple-choice items, it is possible, with the Coombs response mode, to identify nine information and misinformation components within test scores. Whether these score components had similar or differential effects on the reliability and validity of the test used in the study was examined by mathematically removing the score components from the total test scores, isolating, first, the relationship of partial information to conventional number-right scores. Since partial information causes an increment to number-right scores only for guessing correctly among a subset of options that includes the right answer, the score increment resulting from successful guessing among the possible sets and subsets of options was sequentially subtracted by a rescoring procedure used by Cross (1973). It was expected that, since the literature indicates that guessing has a generally depressing effect on reliability, the guessing-free scores, which were produced by the rescoring procedure should be more reliable than test scores containing the guessing components.

A similar procedure was followed to remove the effects of misinformation from the Coombs test scores. Misinformation was defined as being present in any item in which the right answer was incorrectly identified as a distractor. These item scores can only be negative, since the Coombs penalty for the marking of the right answer exceeds

the possible credits for marking distractors correctly on a single item. The relationship of misinformation to reliability was difficult to predict, since it was not known how large the misinformation components in the test scores would be. While it is probable that some guessing will be exhibited in all but the very simplest tests, misinformation score components may be negligible and, thus, have little or no effect on the characteristics of the total test. If, however, the misinformation components proved to be of reasonable magnitude, it might be found that misinformation was a consistent behavior, and its removal from test scores would tend to depress reliability.

Validity was examined by the same method as reliability, with new validity coefficients computed as each level of partial information and misinformation was removed from the test scores. Four concurrent validity criteria were used. The criterion of major interest was an open-ended response form of the multiple-choice items which were answered in the Coombs mode. Because of the additional time required for examinees to respond in the Coombs mode, only the first 25 items of the first section of the test were subject to the experimental response mode; thus, the remaining 35 items in the first part of the test and the two succeeding test sections were also used as validity criteria. Since the literature suggests that the partial information and misinformation score components have little effect on validity, this part of the analysis was essentially exploratory. The primary concern was to determine whether any of the information or misinformation components had differential effects on the validity coefficients.

Examination of the Relationships Between Personality Characteristics and Partial Information and Misinformation

One of the major questions addressed by this study was whether there was significant involvement of personality variables in test performance when examinees were required by the Coombs response mode to appraise the accuracy of their own information. On the assumption that the limited number of personality variables which have been used in previous studies might be insufficient to describe the factors affecting test performance under the Coombs mode, a complete personality instrument and two attitude scales were administered to the examinees in this study. The 24 scales from the personality instrument were reduced by principal components analysis and Varimax rotation to a five factor solution. These five factors and the two attitude scales were the independent variables used in a series of multiple linear regressions. The dependent variables in the regressions were number-right test scores, a test score based on the Coombs response mode with credit given for partial information and penalty for misinformation, an empirical choice-weighted score which accounts for partial information without requiring examinee involvement in the judgment of the weights, and the nine information and misinformation score components derived from the Coombs response mode. In addition, a measure of unwarranted confidence and a standard of assurance, suggested by previous researchers, were also used as dependent variables.

If the enhanced reliability of the Coombs test scores could be at attributed to reliable measurement of personality characteristics

rather than to improved measurement of ability, a significant multiple correlation between the personality and attitude variables and the scores based on the Coombs response mode would be expected. This result was expected to be in contrast to the relationship of personality and attitude variables to conventional number-right test scores, since no judgments about the accuracy of the information reported are required from examinees under the number-right procedure. Delineation of the mechanism by which Coombs scores are affected by personality and attitude variables was expected in the relationships between these variables and the guessing score components and the measure of unwarranted confidence and the standard of assurance. Since the testing literature has not dealt as extensively with the effects of misinformation as it has with the effects of partial information, the misinformation components were considered in an exploratory light.

# CHAPTER II

## REVIEW OF THE LITERATURE

When response modes and scoring rules that account for partial information are used, the expected result is enhanced test reliability. In addition, assuming that the dimension being measured by the test remains invariant across methods of measurement, it is also anticipated that greater precision in measurement, indicated by the reliability coefficient, would produce an increase in the relationships between test scores and validity criteria.

Empirical tests of these expectations have not been uniformly supportive. Increases in reliability coefficients have not been consistent, and there have frequently been negligible changes in validity coefficients accompanying improvements in reliability. Most importantly, it is not clear whether reported enhancements in reliability derive from the ability ostensibly being measured or from some reliably measured by extraneous variables introduced by the conditions attached to the alternative response modes and scoring rules. When the requirements of the scoring rule or response mode impose upon examinees the necessity for judgments on the correctness of their responses, there are, apparently, individual differences in personality characteristics which emerge and which may be measured in conjunction with the ability or knowledge measured by the test.

A diagram of the assumed relationships between the assessment of

partial information and self-assessment in terms of the response modes and scoring rules considered in this study is presented in Figure 1. Each of the entries in the diagram will be defined and discussed in subsequent sections of this chapter.

Briefly, it is assumed that when examinees are asked only to mark what they believe to be the right answer (number-right response mode and scoring rule), partial information is not explicitly measured and there is no involvement of personality characteristics in the responses beyond that usually accompanying a testing situation. The application of a correction-for-guessing to the number-right responses requires some additional judgments from examinees on whether their information is sure enough to risk guessing on items for which they do not know the correct answer. Differential choice-weighting is a mathematical procedure applied to number-right responses. It gives some measure of partial information, but asks for nothing more from examinees than number-right responses. The answer-until-correct response mode provides greater information about partial information levels and is assumed not to invoke personality characteristics associated with self-assessment, but may be related to other types of affective responses. With confidence-weighting and the Coombs method, there is the most precise delineation of partial information and misinformation in responses; however, it is believed that the decisions which must be made in using these response modes are more complex than in other modes and, even though reliability may be considerably enhanced by the use of these modes, it is with these modes that there is the

| SCORING RULE MEASURES PARTIAL INFORMATION | RESPONSE MODE REQUIRES SELF-ASSESSMENT: NO | RESPONSE MODE REQUIRES SELF-ASSESSMENT: YES |
|---|---|---|
| NO | Number-Right (NR) | Number-Right with Correction for Guessing (CG) |
| YES | Number-Right with Differential Choice-Weighting (EMP)<br><br>Answer-Until-Correct (AUC) | Confidence-Weighting (CW)<br><br>Coombs Method (CMBS) |

Figure 1. Comparison of Response Modes and Scoring Rules in Terms of the Measurement of Partial Information and Self-Assessment of Information.

greatest chance that the improvement in reliability may be spurious. Each of these terms will be defined and the assumptions relating to them will be explained in the next sections of thie chapter. A distinction between response modes and scoring rules will be maintained through this discussion. Hakstian and Kansup (1975) noted that two variables are operative in generating different sets of test score scores:

1. the method by which examinees respond (response mode)
2. the method by which the obtained responses are scored (scoring rule).

It is evident from Figure 1 that several scoring rules--number-right, correction-for-guessing, and differential choice-weighting--may be applied to the number-right response mode. It is also possible to use the number-right scoring rule to obtain an actual number-right score from confidence-weighted responses and an inferred number-right score from the answer-until-correct response mode. In this study, an open-ended response and two multiple-choice response modes were used. From these response modes, a variety of scores were calculated using several scoring rules. Consequently, for clarity, the separation of the two components of a testing situation is important.

## Response Modes and Scoring Rules

### Number-Right Response Mode

The response mode under which examinees simply mark a single, best answer for each item will be referred to as the number-right (NR)

response mode. The scoring rule generally used with NR responses is to award one point for a correct choice and no points for an incorrect choice. The NR response mode with NR scoring requires that examinees make only one judgment--that of selecting the best answer for an item.

When an announced correction for guessing (CG) is applied to NR responses, however, the examinee must make additional judgments about his responses since a penalty greater than zero is imposed for each incorrect mark. The corrected-for-guessing (CG) score formula is:

$$S = R - \frac{W}{k - 1}$$

where S = corrected score; R = number of items answered correctly; W = number of items answered incorrectly; and k = number of answer options for each item.

This formula is based on the rather questionable assumption that if an examinee knows the answer he will mark it correctly, or that, in the absence of certainty, he will guess at random among all choices (Cross & Frary, 1977). De Finetti argued, "To assume simply a clear-cut discrimination of perfect knowledge and perfectly blind guessing hinders any real understanding of both and any reasonable approach to the whole question" (de Finetti, 1965, p. 110). The purpose of the CG scoring rule is to remove from the NR score the score credit that would be gained by chance through lucky guessing; but, when the guessing is less than random, *i.e.*, the examinee uses his partial information to eliminate one or more options from the subset within which he

guesses, the CG formula will under-correct, since examinees will be more successful in choosing the correct response from reduced subsets than they would be in guessing at random from the complete set of options. When misinformation is present, the CG formula over-corrects "since the candidate loses not only the point for the question about which (s)he is misinformed, but a fraction of another from questions for which (s)he is entitled to credit" (Rowley & Traub, 1977).

There is some evidence that use of the CG formula suppresses the appropriate use of partial information by some examinees, causing them to omit items for which they know the correct answer (Sherriffs & Boomer, 1954; Slakter, 1968; Votaw, 1936). Lord (1975), a proponent of correction for guessing, criticized these three studies for failing to provide guidance for the examinee to help him find the proper strategy for responding. Cross and Frary (1977) attempted to overcome this criticism in their study of omissive behavior under "do not guess" directions by giving particular attention to control of understanding and compliance with test instructions. Their results supported the earlier studies and suggested that "examinees who understood and conscientiously attempted to follow the directions were underestimating their partial information, and lost score points by not guessing" (Cross & Frary, 1977, p. 320). One explanation which has been offered for the underestimation by some examinees of their partial information, expressed by the reluctance to risk guessing under CG scoring, is that it is a function of personality differences among examinees.

Thus, NR response with NR scoring is placed in the top left box in Figure 1, indicating that the response mode does not require self-assessment and the scoring rule does not measure partial information. The CG scoring rule applied to NR responses is placed in the top right box in Figure 1, since there is no measurement of partial information, but the work of Slakter (1967) and Cross and Frary (1977) suggests that the self-assessment required under the CG scoring rule may elicit differential personality characteristics.

Differential choice-weighting, found in the lower left box of Figure 1, is another scoring procedure which is applied to the NR response mode. Unlike CG, however, it is a passive system from the examinee's point of view, since the score adjustments are determined by outside agents. There is no implicit or overt judgment by examinees of the relative accuracy of a response. Unlike CG, differential choice-weighting allows for the direct assessment of partial information. With differential choice-weighting, adjustments for partial information are made by assigning weights, indicating a relative degree of "correctness," to the answer options. These weights may be assigned beforehand, based on the judgment of the test maker (Logical Choice-Weighting), or by weighting the choices as a function of the ability levels of those examinees selecting each choice (Empirical Choice-Weighting). Ability may be estimated by use of total scores on the test or on some external criterion. The differential choice-weighted score on the test is a summation of the weights attached to the options selected by an examinee. The response strategy which the

examinee is directed to adopt while taking the test is the same as that offered under NR scoring--simply to choose a single best answer for each item. He is not required to assess the confidence he has in the correctness of his response.

Since logical choice-weights could not be determined because of the nature of the test used in this study, empirical choice-weighting was used, with total test scores as the ability measure. Empirical choice-weighting will be referred to as EMP, but references to logical choice-weighting will not be abbreviated.

The expectation for the differential choice-weighting procedures was that the increased differentiation in examinees' scores would result in higher reliability and validity coefficients for tests. Some studies using differential choice-weighting have been encouraging. Davis and Fifer (1959) reported an increase of 50 percent in effective test length from differential choice-weighting. They commented that with differential choice-weighting, "Variance arising from selection by examinees among distracters of unequal merit is obtained; this variance is excluded from measurement when all incorrect choices are weighted equally" (Davis & Fifer, 1959, p. 165). They also expected that differential choice-weighting would increase the reliability of difficult tests more than easy tests, "since the nonchance variance associated with selection among incorrect choices by examinees tends to increase as the number of items marked incorrectly increases" (Davis & Fifer, 1959, p. 165).

There was no change in validity, however, in the Davis and Fifer

study. They offered this explanation:

> The increment of nonchance variance responsible for the increase in test reliability is attributed to the selection among incorrect choices exercised by examinees in the case of items to which they do not know the answer. This variance displayed the same concurrent validity as variance attributable to the selection of correct choices. Thus, scoring with choice weights increased test reliability without reducing test validity (Davis & Fifer, 1959, p. 169).

Wang and Stanley (1970) take slight exception to this conclusion, noting that "the increased reliability of the test should have resulted in a concomitant increase in the validity" (p. 693).

Hendrickson (1971) was able to demonstrate an average effective increase in test length of 49 percent using EMP, a result similar to that of Davis and Fifer (1959). In a cross-correlational validity analysis, however, she found, in all cases, a decrease in validity coefficients with EMP scores. The effect of EMP is to maximize internal consistency reliability, *i.e.*, to make the items in the test more homogeneous. Hendrickson's results suggest that "the more homogeneous a test is made, the more poorly it correlates with something quite different" (Hendrickson, 1971, p. 25). Thus, in this case, the increment in nonchance variance which resulted in the higher reliability coefficients did not bear the same concurrent validity as the original variance of the test, as was apparently the situation in the Davis and Fifer (1959) results.

The difference in the validity results of these two studies raises an important issue in the use of alternative response modes and scoring rules on multiple-choice tests, that of whether or not a

different dimension is being measured when procedures other than NR responses and scoring rules are being used. Mechanically, for validity to remain the same across response modes and scoring rules, the relationship of the test scores to the criterion must remain the same. For this to occur, the relative positions of examinees to each other must be unchanged under the alternative procedures. For validity to increase, the relative positions of examinees on the test must more closely approach their relative positions on the criterion, indicating that the dimension being measured by the test has become more similar to that measured by the criterion. A decrease in validity indicates that the relative positions of examinees have altered in some way which is different from the criterion. Thus, increasing the homogeneity of a test may emphasize its dissimilarity to the validity criterion, as suggested by Hendrickson's (1971) results. Serlin and Kaiser (1978) came to a similar conclusion, based on their results. They found that applying the EMP procedure to a 5-item test gave an increase in reliability from .50 to .77, but that the EMP scores and the NR scores produced a correlation of only .75, indicating that the relative positions of examinees were somewhat different as a function of the scoring rule applied. They concluded from this correlation that, "clearly we are measuring something somewhat different by this procedure of scoring [EMP]" (Serlin & Kaiser, 1978, p. 339). They tended to attribute this "something somewhat different" to the effects of partial information and poor test items.

Kansup and Hakstian (1975) summarized the literature on

differential choice-weighting as showing "insubstantial increases in reliability and no increase in validity" (p. 219) for studies using EMP, and mixed reliability results and no statistically significant increases in validity for logical choice-weighting. Their study using logical choice-weighting showed an increase in internal consistency reliability over NR scores, but a decrease in test-retest reliability. There was no change in validity for the verbal reasoning measures, but a significant decrease in validity for the measures of mathematical reasoning. As an explanation for this difference they proposed that:

> The logical weights assigned incorrect vocabulary item options by raters may approach more readily the latent "degree of correctness" each option has than do those assigned by raters to Mathematical Reasoning item alternatives (Kansup & Hakstian, 1975, p. 225).

The validity criteria used were several course grades, overall grade point average, and additional measures of verbal and mathematical reasoning ability.

Cross and Frary (1978) found increases in both reliability and validity for EMP scores over both NR and CG scores. In a subsequent study, however, Cross, Ross, and Geller (1978) found that logical choice-weighting gave inconsistent validity and reliability results and, although EMP greatly increased the internal consistency reliability coefficients, there was no accompanying increase in validity. They commented:

> With respect to reliability and validity as defined in this study, the data reported here offer no basis for recommending either empirical or logical choice-weighted scoring. Even if the empirical weighting procedure had been successful in improving predictive validity, it would be difficult for an instructor to

> justify scoring achievement tests in this manner unless the weights determined empirically were consistent with his/her judged degree of correctness. Initially it was the intention to use empirical weights as a guide for establishing logical weights. In many cases, however, the empirical weights were in direct conflict with what logic would suggest. It may be that misinformation played a larger role than anticipated, and this offset the order of attractiveness among choices that would have been arrived at through the use of partial information alone (Cross, Ross, & Geller, 1978, p. 7).

In summary, the NR response mode can be scored by the NR scoring rule, CG scoring rule, or EMP scoring, although the directions given to examinees for CG scoring would be different from NR directions. The standard to which other response modes and scoring rules are generally compared to determine changes in reliability and validity is NR responses with NR scoring. The two alternative scoring rules considered thus far may produce increased internal consistency reliability coefficients, but the effects on validity appear to be inconsistent. In addition, the CG scoring rule, because it requires judgment by the examinee of his responses, may introduce unwanted personality variables into the scores. Studies relating to personality contamination of test scores will be discussed later in this chapter.

EMP scoring does not call for a self-assessment by examinees, but there is evidence suggesting that the option weights assigned by EMP may be contrary to logic, and that misinformation on a few items by generally high scoring examinees may distort the weights. In addition, the lower test-retest reliability for logical choice-weighting reported by Kansup and Hakstian (1975) suggests that the use of partial information may not be consistent over tests. Although EMP does

measure partial information among examinees, it is not clear whether the ability measured by the EMP rule is the same as that measured by NR. EMP will maximize internal consistency reliability, but, in doing so, may also negatively affect the relationship of the test to more heterogeneous validity criteria.

Answer-Until-Correct Response Mode

The answer-until-correct (AUC) response mode is found in the same lower left box in Figure 1 as EMP, indicating that AUC also measures partial information without requiring self-assessment from examinees. However, the means by which partial information is obtained through AUC differs sharply from the EMP procedure.

AUC uses special answer sheets on which all answer options are hidden by removable covers. The examinee reads the item, determines which of the options he believes is correct, and uncovers that option on his answer sheet. If the option which he has uncovered contains a special mark indicating that he has indeed chosen correctly, he will receive the maximum number of points for the item and can proceed to the next item. In a 5-option item, he would receive four points. If his initial selection is not correct, however, he would then make another selection among the remaining options. If he is correct on his second choice, he would receive three points. In this manner, the test scores are derived from the number of attempts necessary to uncover the right answer. Some credit is awarded as long as fewer than $n$ options are uncovered. Consequently, the examinee who has been able

to narrow the answer options to a subset of two, but who does not guess correctly the first time, will still receive some improvement in his test score for his partial information over what he would receive under NR scoring.

Although AUC is not one of the response modes used in this study, there are several points raised by studies using AUC which are relevant to this inquiry. With the AUC procedure, it is possible to infer a NR score by counting only those items for which the single correct option is uncovered. A number of studies using inferred number-right scores (INR) have reported increases in reliability for AUC over INR (Gilman & Ferry, 1972; Hanna, 1974; Hanna, 1975). Hanna (1977) reported a study with three conditions of feedback, including AUC and NR, with random assignment of treatment conditions to intact classes. His results suggested that the reliability inferiority of NR in comparison to AUC was similar to the INR scores reported previously.

Gilman and Ferry (1972) reported a split-half reliability coefficient for INR of .79; the split-half coefficient for AUC was .93, an increase in effective test length of 253 percent. They speculated that the increase might be due to the inclusion of partial information in the test scores, but they also suggested that the act of marking additional responses may be somewhat similar to the effect of increasing the test length. They commented:

> If the increase [in] reliability is simply a result of Ss' responding to items more frequently in SSM [self-scoring method - AUC] than in RWSM [right-wrong scoring method - NR],

> it might be argued that this amounts to a spurious increase in reliability (Gilman & Ferry, 1972, p. 207).

If, however, it is assumed that, even though the NR response mode requires the selection of only a single answer option, examinees consider each available option, however briefly, before committing themselves to one answer, the multiple responses to each item are already present and contribute to the answer selection. The emphasis of the Serlin and Kaiser (1978) study discussed in the previous section was the use of all information available from a test. If decisions about each option are already being made as part of the examinee's test-taking strategy, the AUC response mode (and the Coombs response mode which will reviewed subsequently) simply provides an opportunity for examinees to report existing information about their responses. In this sense, a resultant increase in reliability cannot be considered to be spurious. However, the AUC response mode may increase the length of time required to administer the test. Thus, if NR was used for the same length of time, more items could be given with an expected increase in reliability.

Problems with validity, similar to those with EMP, have been encountered with AUC. Hanna (1974, 1975, 1977) has studied the effects of AUC on validity in several different settings. In his 1974 study, he found a nonsignificant increase in validity, using two term papers and an interpretive exercise as criteria. In his 1975 study, with an interpretive exercise, a term paper, and an essay part of the final examination as criteria, he found validity to show slight to

significant decreases with AUC. He hypothesized that "the relative criterion-related validity of AUC and conventional procedures may hinge on built-in biases of the criterion measures used" (Hanna, 1975, p. 178). He viewed criterion-related validity as having two components--reliability and relevance. He defined relevance as the estimated correlation between obtained scores on the criterion and true scores on the experimental test. Since the reliability of the AUC test was expected to be higher than that of the NR test, the influence of relevance on validity was assumed to be the critical factor. Hanna (1977) argued:

> Where a criterion involves only one response to each item (*e.g.* in a conventional objective test), scores from AUC testing should be less relevant (*i.e.* less valid after correction for attenuation in the predictor) than scores from conventional testing. This is because the tests are simply measuring somewhat different attributes. Conversely, where a criterion involves the skillful use of feedback in seeking answers (*e.g.* in the game Twenty Questions), scores from AUC testing should be more relevant than scores from conventional testing because the tests would be tapping the same attributes. . . . In situations where the criterion measure is essentially an AUC measure, then both reliability and relevance should contribute to the validity superiority of AUC procedures. But in more typical situations where the criterion measure requires only one response per item, reliability and relevance should contribute to opposite outcomes; the difference between criterion-related validities arising from AUC and conventional procedures should be relatively small and difficult to predict (Hanna, 1977, p. 3).

The results from his test of this argument were of essentially no differences in the validity coefficients for three graduated feedback conditions and the conventional right/wrong completion test used as the criterion. His measures of relevance were nonsignificant but increasing in magnitude as the testing situation moved from complete

feedback (AUC) to no feedback (NR), approximating the no feedback condition of the criterion. The split-half reliability coefficients changed in the opposite direction, being greater for AUC than for NR. Hanna (1977) suggested that the increased information provided by AUC was counteracted by the dissimilarity between the complete feedback AUC and the no feedback validity criterion.

Since the primary concern of this study is on ways in which partial information and misinformation function when examinees are required to judge the accuracy of their own information, the AUC response mode has several important drawbacks as a research tool. First, AUC does not, apparently, impose the same sort of self-assessment burden as do the two response modes in the lower right box of Figure 1. There is some question about the affective dimension in examinees' responses under the immediate feedback mechanism in AUC (Hanna, 1974, 1975, 1977), but it is believed to be of a different type of emotional involvement from that elicited when judgments must be made with no feedback.

The most important difficulty with using AUC to examine partial information and misinformation is the amount of inference necessary to detect what type of information or misinformation was involved in the answer selection. Partial information may have been operative in items where fewer than $n$ alternatives were uncovered, but when the examinee uncovered all of the options, the explanation may be either misinformation or bad luck in random guessing. If the right answer is uncovered first, there is no way to determine whether the selection

was from complete information or from fortunate guessing among two or more alternatives, nor is there any way to know whether misinformation was involved in the early choices for an item that is finally answered correctly.

Confidence-Weighting Response Mode

Confidence-weighting (CW) offers credit for partial information, as does AUC, and also permits the identification of intermediate levels of information and misinformation. With CW, the examinee indicates his own belief in the correctness of the answer options by attaching his own weights to each item. An item which an examinee answers correctly and gives the highest possible weight can be defined as resulting from complete information. A right answer with an intermediate weight can be considered as reflecting partial information. Wrong answers can similarly be attributed to complete and partial misinformation. CW differs from differential choice-weighting in that examinees themselves determine the weight for an item, rather than the weight being set by an outside agent. The weights represent an individual's confidence in the correctness of his choice of an answer and can vary with the individual and the item. The usual way in which this is done is to allot a certain number of points, *e.g.* 10 points, which may be assigned to each item. If the examinee is completely sure of the answer he has selected and marked, he would assign the whole ten points to this item. If he is reasonably sure, he might assign only eight points; if he thinks the chances that he is right are

only moderate, he might assign five points; for a sheer guess, he would use 0, 1, or possibly 2 points for that item. Generally, the scoring rule would be to give positive credits for the weights attached to the correct answers and a loss of the points assigned to wrong responses. A number of variations of this scoring rule have been investigated, including logarithmic transformations. There is, however, no evidence that superior psychometric properties accrue to tests scored by rules more sophisticated than a simple summation of the confidence weights for the items (Kansup & Hakstian, 1975).

Hambleton *et al.* (1970) summarized the expectations held by investigators of CW as follows:

> Advocates of confidence testing have suggested several possible advantages of the procedure. One is that more information is extracted about the examinee's state of knowledge and, therefore, scores should be more valid. Another is that guessing is discouraged and, as a consequence, reliability may be increased (p. 76).

Their study of CW, however, showed decreased reliability and increased validity, compared to NR scores and two methods of differential choice-weighting. The changes were not large and were not assessed for statistical significance because of the small size of the sample. Hambleton *et al.* (1970) suggested that some of their failure to find large differences might be attributed to the easiness of the test for the examinees. The examinees selected the highest confidence weight for their answer choices about 77 percent of the time.

Hopkins, Hakstian, and Hopkins (1970) found a small increase in reliability and a decrease in validity with CW. Michael (1968)

reported a gain in reliability of approximately 62 percent effective test length. Hakstian and Kansup (1975) compared NR, CW, and the Coombs (1953) response mode and found that neither reliability nor validity was consistently increased by the alternative response modes. They commented:

> The preceding results must be considered disappointing to those who believe that substantial additional information can be learned about subjects' performance potential by going beyond conventional 1-0 scoring (Hakstian & Kansup, 1975, p. 228).

Diamond (1975) used a variant of confidence weighting under which examinees marked as many options for an item as they felt they needed to include the right answer. If they were certain of the answer, they were instructed to fill in the appropriate box with a number "1"; if they could not decide between two choices, they were to indicate both of these choices by placing a "1" in one appropriate box and a "2" in the other appropriate box. Up to four choices could be indicated, but no credit was given if all four boxes for an item were filled in with numbers. The items were scored as follows:

> one box filled in and correct = 3 points
> two boxes filled in and one of the two correct = 2 points
> three boxes filled in and one of the three correct = 1 point.

An inferred NR score was determined by counting all the items for which the box for the keyed answer contained a "1."

Scores using this response mode and scoring rule were determined for a pretest, two tests, and a final examination in a measurement course. Diamond (1975) found a sizeable difference in Flanagan reliability estimates between the experimental scores and INR scores

only on the pretest. As the course proceeded, the reliability estimates tended to converge, suggesting that the use of examinees who are unfamiliar with a novel response mode and scoring rule may artificially inflate reliability coefficients.

Two additional points should be considered in this connection, however. On the pretest, the reliability coefficient for the INR scores was only .34; with the experimental response mode, the reliability improved to .46, an increase in effective test length of 65 percent. When test reliability is as low for NR scoring as it was on this pretest, a 65 percent improvement may have practical importance, but enhancement of reliability is easiest on tests with initially low reliability.

The second consideration relates to the amounts of partial information and misinformation in the test scores. It is assumed, from the directions reported by Diamond (1975) that no additional penalty was exacted for misinformation. As long as the right answer was not within the subset marked by the examinee, the item score was, apparently, zero. Whether the misinformation was given with complete confidence when the examinee "knew" a wrong answer was right and selected only one box, or whether the examinee "knew" the right answer was wrong and indicated some guessing among a subset of wrong alternatives, a uniform penalty was imposed. Thus, the scores did not account for differential effects of partial and complete misinformation.

In addition, there is evidence that partial information, and thus the effects of guessing, had essentially dropped out of the test

scores by the final examination. Diamond (1975) reported that, as the course proceeded, examinees were increasingly expressing higher confidence in most of their responses. The correlation between INR scores and the experimental scores on the final examination was .97, indicating that 94 percent of the variance under one scoring rule was accounted for by the other. The major advantage of CW, however, lies in allowing the expression of partial information. If partial information is not expressed and if there is no penalty for misinformation beyond that imposed by NR scoring, then test scores from NR and CW scoring rules are highly "relevant" (Hanna, 1977) to each other, should be expected to correlate highly, and should show similar characteristics. The reliability and validity coefficients would, thus, converge as they did in Diamond's (1975) study. The increase in reliability coefficients with response modes which account for partial information over NR scoring may be, in part, a function of the difficulty of the tests, more difficult tests being more likely to show the expected improvement than easier tests. However, when there is a considerable discrepancy between the reliability coefficients obtained under two scoring rules, the question of whether the same dimension is being measured by both scoring rules is raised again.

Coombs Response Mode

In 1953, Coombs proposed a response mode which, he believed, would "discriminate levels of partial knowledge" and "yield a greater variance of test scores than the conventional procedure for the same

number of items" (Coombs, 1953, p. 308). The Coombs response mode (CMBS) asks examinees to mark all the answer options they believe to be wrong, taking care not to mark the correct answer. One point is awarded for every correctly identified wrong answer and a penalty of $-(n - 1)$ is imposed for identifying the right answer as incorrect. From the various patterns of eliminating answer options and whether or not the right answer is included in the subset of options designated as wrong, a range of information and misinformation score components can be determined.

Like CW, CMBS also assesses partial information and requires the examinee to make judgments about the accuracy of his own information. An inferred CW may be derived from CMBS responses, if it is assumed that the number of options identified as being wrong reflects the confidence of the examinee that the option or options he has left unmarked include the right answer, and if the NR response is also required from the examinee.

The effects of the CW scoring rule and the CMBS scoring rule are identical for complete information and ignorance. A correct answer with the highest confidence weight would receive full credit as would a CMBS response with all distractors correctly marked. An item with a zero confidence weight is equivalent to a CMBS item with none of the item options marked.

The credits awarded for partial information and the penalty for misinformation are different, however. On a 5-option item under the CMBS mode, an examinee who correctly marks three distractors and,

thus, indicates that he believes the correct answer is one of the two unmarked options, would receive three points credit. Assuming a scoring rule under CW which gives credit or penalty in the amount of the attached confidence weight, a guess between two options with a weight of three given to the item would result in a credit of three points only if the selection were correct. If the guessing between the two options is wrong, as it would be half the time, the examinee loses three points. Thus, CW may fail to account for a portion of the correct information which the examinee may possess.

The penalties for misinformation are reversed from CMBS to CW. Under CMBS, an examinee who correctly identifies three of the item options as incorrect but also includes the right answer in his subset of incorrect answers (a total of four marks) will be penalized four points for his inclusion of the right answer, but will also receive one point credit for each correctly identified distractor, giving him a net penalty of minus one point. Under CW, the wrong answer marked with a confidence weight of four results in a 4-point penalty.

At the other extreme of misinformation, the marking of only the right answer as a distractor causes the maximum CMBS penalty (with five options) of four points to be imposed. Under CW, guessing among four alternatives (when the right answer has been excluded from the subset) with an attached confidence weight of one, will bring only a 1-point penalty. The 1-point penalty would also be incurred one fourth of the time from guessing within a subset which contained the right answer.

The advantage of the CMBS mode for a study of information and misinformation components in test scores is its sensitivity to the differential effects of partial information. Any correct bit of information is rewarded under the CMBS scoring rule. With CW, as under NR scoring rules, the presence of partial information results in a score credit only if it leads to a correct answer.

A disadvantage of the CMBS mode arises from the complexity of the decisions required from an examinee to enable him to maximize his expected score (Wang & Stanley, 1970). De Finetti (1965) described CMBS and CW among other methods of assessing partial information through the subjective probabilities examinees attach to each answer option. These subjective probabilities are assumed to reflect "the degree of belief of an individual about the correctness of each of the alternatives at the moment he is facing the problem of how to answer a specific item in a questionnaire" (de Finetti, 1965, p. 88). The emphasis of de Finetti's (1965) discussion is that for each response mode and scoring rule, there is an optimum strategy which must be discerned and applied by the examinees. He assumed that examinees can and will grasp the probabilistic implications of a particular combination of response mode and scoring rule and will act in their own best interests in selecting a response strategy. The rule given by de Finetti for maximizing expected score under the CMBS mode is: "cross out alternatives until the probability $p_h'$ of the $r - h$ alternatives already crossed out plus that of the next one $p_h^*$ when multiplied by the number $h$ of those still left, does not attain 1/2" (de Finetti, 1965, p. 98).

Obviously, the above rule can be presented to examinees in a simpler, less mathematical manner; nevertheless, the decisions which must be made about each item under the CMBS response mode are more complex than those made under NR response mode. Wang and Stanley (1970) commented on the application of mathematical decision theory to testing procedures:

> The success of testing procedures which attempt to control the decision process will be critically dependent on the ability of subjects to effectively use optimum strategies. It is not certain that all students are equally capable of learning to use such strategies (p. 698).

Stanley and Wang (1970) suggested that reporting directly the subjective probabilities, as in CW, may be less taxing on the examinee than the CMBS response mode.

## Expected Item Scores Under Different Scoring Rules

The raw and standardized expected item scores under various levels of information and misinformation for the response modes and scoring rules discussed in this section, with the exception of differential choice-weighting, are presented in Table 1 (Frary, in press). Neither logical choice-weighting nor empirical choice-weighting (EMP) is included in the table because expected item scores under both types of differential choice-weighting are test specific and not generalizable.

The information and misinformation categories given on the left of the table represent the following response patterns on 5-option items for the three response modes which will be used in this study:

Table 1. Raw and Standardized[1] Expected Item Scores Under Various Levels of Information and Misinformation for 5-Choice Item (from Frary, in press).

| Level | Raw NR[2] | Raw CG | Stn. NR/CG | AUC/CMBS Raw | AUC/CMBS Stn. | CW Raw | CW Stn. |
|---|---|---|---|---|---|---|---|
| Information | | | | | | | |
| INFO | 1.00 | 1.00 | 1.00 | 4.00 | 1.00 | 4.00 | 1.00 |
| G2 | .50 | .38 | .38 | 3.00 | .75 | 3.00 | .38 |
| G3 | .33 | .17 | .17 | 2.00 | .50 | 2.00 | .17 |
| G4 | .25 | .06 | .06 | 1.00 | .25 | 1.00 | .06 |
| Ignorance | .20 | .00 | .00 | 0.00 | .00 | 0.00 | .00 |
| Misinformation | | | | | | | |
| MIS1 | -.20 | -.25 | -.25 | -1.00 | -.25 | -4.00 | -1.00 |
| MIS2 | -.20 | -.25 | -.25 | -2.00 | -.50 | -3.00 | -.50 |
| MIS3 | -.20 | -.25 | -.25 | -3.00 | -.75 | -2.00 | -.75 |
| MIS4 | -.20 | -.25 | -.25 | -4.00 | -1.00 | -1.00 | -.25 |

[1]Linear transformation to assign 1 to INFO and 0 to Ignorance.
[2]With expected guessing gain.

INFO - NR: Marking the correct answer.
CMBS: Marking correctly four distractors.
CW: Marking correct answer and attaching weight of 1.00.

G2 - NR: Guessing between two choices.
CMBS: Correctly indicating three distractors.
CW: Guessing between two choices and attaching weight of .75.

G3 - NR: Guessing among three choices
CMBS: Correctly indicating two distractors.
CW: Guessing between three choices and attaching weight of .50.

G4 - NR: Guessing among four choices.
CMBS: Correctly indicating one distractor.
CW: Guessing among four choices and attaching weight of .25.

Ignorance (IG) -
NR: Guessing randomly among five choices.
CMBS: Indicating no choices as distractors.
CW: Guessing among five choices and attaching weight of 0.

MIS1 - NR: Eliminating right choice from guessing subset.
CMBS: Correctly indicating three distractors and the right answer.
CW: Attaching probability of 1.00 to wrong choice.

MIS2 - NR: Eliminating right choice from guessing subset.
CMBS: Correctly indicating two distractors and the right answer.
CW: Attaching probability of .75 to wrong choice.

MIS3 - NR: eliminating right choice from guessing subset.
CMBS: Correctly indicating one distractor and the right answer.
CW: Attaching probability of .50 to wrong choice.

MIS4 - NR: Eliminating right choice from guessing subset.
CMBS: Indicating only the right answer.
CW: Attaching probability of .25 to wrong choice.

From Table 1, it is evident that the rewards for partial information and the penalties for misinformation are not uniform across response modes and scoring rules. The expected credit for partial

information is identical for NR and CW, and less than that given under CMBS. Under NR, all misinformation is penalized the same. The penalties for misinformation under CMBS and CW are graduated, but in opposite directions. The relationships among NR, CMBS, and an inferred CW score as well as an EMP score, using total test scores as the ability criterion, formed the basis for the examination of partial information and misinformation in this study.

## The Relationship of Personality Variables to Response Modes and Scoring Rules

While confidence weighting (CW) and particularly the Coombs response mode (CMBS) show a distinct superiority over other response modes in allowing the partitioning of examinees' responses into information and misinformation categories, these two modes also make the greatest demands upon examinees for subjective appraisal of their own knowledge. There is some empirical evidence which suggests that the inclusion of this subjective element in the testing situation is reflected in score differences which seem likely to be attributable to personality characteristics, rather than to the ability ostensibly being measured.

As long ago as 1936, the question of the presence of unwanted variables in CW test scores was raised by Wiley and Trimble (1936). They asked 59 students to indicate their confidence in their answers in addition to marking what they believed to be the right response on four tests given in a general psychology course. Three of the tests

were regular periodical tests for the course and the fourth was the final examination. The degree of confidence was to be designated as "certain," "doubtful," or "guess." An unspecified penalty for wrong answers marked "certain" was also announced.

For each test, four scores were calculated. "Certainty," "doubt," and "guess" scores were determined by counting the number of items so marked, and an achievement score (CG for true-false items and NR for multiple-choice and matching items) was also computed. There were, thus, four scores for each of the four tests. A correlation matrix was then calculated, giving 24 unique combinations of tests and scores. Wiley and Trimble (1936) reported an average correlation for "certainty" scores on the four tests of .66; for "doubt" scores the average correlation was .58; for "guess" scores, .57. However, the average correlation between pairs of achievement scores on the tests was only .39. There was apparently greater consistency in the way in which examinees reported their perceptions of the accuracy of their knowledge (indicated by "certainty," "doubt," and "guess" scores) than there was in the amount of the knowledge itself (indicated by the achievement scores). Wiley and Trimble concluded that the presence of these constant factors of "certainty," "doubt," and "guess" represented some unknown but relatively stable personality variables which were operative when examinees were required to commit themselves not only to their knowledge but also to an estimation of their own confidence in that knowledge.

Swineford (1938) examined Wiley and Trimble's (1936) hypothesis

in conjunction with Soderquist's (1936) report that confidence weighting in true-false tests improved the reliability of the tests. She noted that the weighted scores might well appear more reliable because they measured both knowledge and some unknown personality trait. She defined this personality trait as "the tendency to gamble," and derived a measure of this tendency from the responses to a true-false test, using the following formula:

$$\text{Gambling} = G = \frac{\text{Errors marked with highest confidence}}{\text{Total errors} + 1/2 \text{ omissions}} \times 100$$

Her results supported the greater consistency of scores in the confidence categories compared to achievement scores, as had been reported by Wiley and Trimble (1936). She also found a higher reliability coefficient for the total weighted test scores (.68) than for the unweighted total test scores (.59), an increase in effective test length of approximately 48 percent.

Swineford (1938) claimed that if the tendency to gamble is "a personality trait which is not intimately associated with the ability being measured, then it should not be permitted to affect the achievement score" (p. 298). She based this assertion on her finding that G scores were negligible related to total unweighted achievement scores (.08), but highly related (.87) to the total number of items marked with highest confidence (called "4's"). These correlations imply that the tendency to assign highest confidence was a total response pattern that was essentially independent of whether the item was right or wrong. Although the G score was derived from a portion of the items

assigned 4's, it does not necessarily follow that the correlation between the two would be high. Had the examinees indicated greatest confidence only in those items which they answered correctly and were also aware of and reported accurately a lower confidence for wrong items, there would have been few or no wrong items marked with highest confidence. Under these circumstances, the G scores would have been close to zero and would have borne a random relationship to the items assigned 4's, as well as a strong relationship to the total unweighted achievement scores--the reverse of the results which Swineford reported.

The G score cannot, however, be viewed solely as a measure of the tendency to gamble. It seems highly likely that some of the items were ambiguous and misinterpreted by examinees and, also, that some of the errors marked with highest confidence represented entrenched misinformation. Swineford (1938) was aware of this problem and attempted to control for it by removing items which more than one-half of the students answered incorrectly and those for which the percentage of errors marked four exceeded 50 percent, except for those answered incorrectly by fewer than 50 students. Using this reduced test, she recalculated her results and found essentially the same relationships among the scores as before, indicating that ambiguities and misinformation did not materially affect the correlation coefficients. Nevertheless, the possibility that the G scores include misinformation should be noted. Swineford seemed to believe that misinformation would play a small role in an achievement test designed for a course

in which most students attended the lectures and read the assignments. On a standardized achievement test, however, which may be given independently of a course, misinformed responses may be more frequent.

Other investigations (Cross & Frary, 1977; Hansen, 1971; Jacobs, 1971; Sherriffs & Boomer, 1954) have also revealed the presence of extraneous factors in tests using response modes which require examinees to appraise their own knowledge. Sherriffs and Boomer (1954) reported that students who scored on the maladjusted end of the A-scale of the Minnesota Multiphasic Personality Inventory (MMPI) were handicapped by a penalty for guessing imposed on a true-false test. The penalty accruing to the group of more maladjusted students came in part from their omitting more items than the well-adjusted students, and omitting a larger proportion of items that they could answer correctly.

The study by Sherrifs and Boomer (1954) dealt with CG scoring applied to true-false items. It has a direct relationship to Swineford's (1938) study and at least a partial relationship to the Wiley and Trimble (1936) results. A portion of the items in the tests used by Wiley and Trimble were true-false items with CG scoring, and, consequently, the effects of the CW response mode were probably, in some degree, contaminated by the CG directions. In Swineford's study, all test items were scored with the CG rule. If, as Sherriffs and Boomer suggest, more anxious students tend to omit more items under CG scoring, then more anxious students would also tend to have smaller G scores, since the denominator of the G ratio would be increased by one one-half the number of omitted items without any increase in the

numerator. There is also the possibility of reluctance on the part of more anxious students to indicate highest confidence in many of their answers, and thus they would tend to have fewer right *and* wrong answers marked with highest confidence. While the tendency to gamble may be expected to be less pronounced among anxious examinees, the inclusion of the omitted items in the G score formula may somewhat obscure the response characteristics. The number of wrong items is a conglomerate of items on which examinees guessed poorly or were misinformed, while omitted items may represent a somewhat different process. The rationale for including one half of the omitted items in the G score formula is that, on true-false tests, random guessing would result in half of the items being right and half being wrong. There is some evidence, however, which suggests that at least some examinees who omit items are underestimating their partial information and may be able to answer more items than the random-guessing probability would indicate. Cross and Frary (1977) in an examination of omissive behavior on a 4-option multiple-choice test reported that:

> Low risk takers, unlike higher risk takers, may be more successful than would be expected by chance, in guessing the correct answer to items omitted under formula-scoring [CG] directions (p. 317).

Cross and Frary also used the A-scale from the MMPI as a personality measure, but were unable to demonstrate a strong linear relationship between guessing performance and scores on the A-scale. They found a correlation of -.14 between A-scale scores and one of their measures of guessing and .03 for another measure of guessing. However,

Sherriffs and Boomer (1954) were unable to demonstrate a strong linear relationship between A-scale scores and CG scores or between A-scale scores and NR scores. The correlations, respectively, of -.23 and -.22 were significantly different from zero, but the relationships can cannot be considered strong ones, and there is no evidence from these correlations that A-scale scores were related to the CG scores in any different way from their relationship to NR scores. The difference reported by Sherriffs and Boomer was based on a comparison of the CG scores of extreme groups paired on NR scores. The mean CG score for the low A-scale group was significantly higher than the mean for the high A-scale group.

Cross and Frary (1977) did not examine extreme groups in terms of scores on the A-scale. They did, however, separate the examinees into two groups on the basis of their understanding and attempt to comply with the CG instructions. Those who reported that they had complied with the directions scored significantly lower on the A-scale than those who reported that they had frequently ignored the directions, but there was no difference between the groups either on the number of items omitted or on NR and CG scores. These results suggest that the *less* anxious students were more likely to report that they had complied with the test directions, but the difference was not reflected, apparently, in omissive behavior, nor did either group appear to be significantly penalized in terms of total test scores.

It is difficult to compare the Cross and Frary (1977) results to those of Sherriffs and Boomer (1954) since different criteria were

used to divide examinees into groups. It seems likely that at least some of the extreme scorers on the A-scale may not have been included in the Cross and Frary groups, since examinees who could not present evidence of having understood the directions, as well as examinees who understood but answered all questions, with no items omitted, were excluded from the groups formed by Cross and Frary. The different bases for groupings were dictated by the emphases of the respective studies, and, while the Cross and Frary results raise some questions about the role of anxiety in test behavior, they do not necessarily contradict the Sherriffs and Boomer findings.

Jacobs (1971) investigated the relationship between Swineford's (1938) measure of "tendency to gamble," and four scales from the California Psychological Inventory (CPI). The four scales from the CPI constituted a measure of poise, ascendance, and self-assurance. Jacobs reported a significant multiple correlation of .39 between the CPI scales and the "tendency to gamble," which he had modified by excluding the omitted-item term and renamed a measure of "unwarranted confidence." Jacobs observed that confidence weighting was biased against ascendant or self-confident examinees rather than against more anxious examinees, since self-confident examinees tended to give a greater number of incorrect responses with the highest confidence weighting, thus incurring a greater penalty for misinformation than more anxious examinees. Jacobs also used two penalty levels, to which examinees were randomly assigned. He found no significant difference in the measures of unwarranted confidence as a function of penalty

level, but he did report that the correlation between NR and CW scores declined from .88 in the low penalty group to .095 for the high penalty group.

This marked decline in the correlations between NR and an alternative response mode and scoring rule raises the question discussed previously of whether the alternative procedures measure a different dimension than NR scores. Traub and Fisher (1977) attempted to investigate this question directly by using both verbal comprehension and mathematical reasoning tests with a constructed or open-ended response mode, the standard NR response mode, and the CMBS response mode. It was their expectation that a "format factor" would emerge, reflecting the effects of "the longer and more involved instructions associated with the Coombs format" (p. 356), or the different type of memory assumed to be associated with open-ended responses as opposed to multiple-choice tests. Their results suggested a possible, but weak, open-ended response format, but no evidence of either a CMBS response format or a NR response format.

Jacobs (1971) concluded, based on the sharp difference in correlations between NR and CW scores under high and low penalty and his finding that under the high penalty condition the reliability coefficient for CW scores declined from .87 to .39, that:

> Confidence-weighting seems to be a questionable measurement technique, since it is contaminated by individual differences in personality. Still open is the question of whether this effect is greater for this particular procedure [CW] than for others, *e.g.*, Coombs' (1953) type of directions (Jacobs, 1971, p. 18).

Thus, Jacobs appears to attribute differences in the dimensions

measured by NR and CW to differences in personality among examinees which are evoked by the charactersitics of the response mode.

Hansen's (1971) study seems to confirm Jacobs' finding that, in the face of score loss for wrong responses, the more confident examinees would tend to be penalized more severely than less confident examinees. He used an estimate of the tendency of an individual to show certainty in CW tests. Although he reported nonsignificant correlations between certainty and test anxiety, he found that risk-taking, as measured by the Kogan and Wallach (1964) Choice Dilemmas Questionnaire, was significantly and negatively related to response certainty. Those examinees who indicated that they would advise others to enter situations in which the probability of success was low also tended to exhibit a high degree of confidence in their responses on tests.

These studies are difficult to compare because of differences in methodology and instrumentation. It seems, however, that there are two ways in which personality characteristics can be involved in response modes which require judgments by examinees on their knowledge. The first, indicated by Sherriffs and Boomer (1954), is that extremely anxious students tend to omit items for which they could make the correct response if forced to do so. This result suggests that these students may fail to receive credit which they could claim for partial information which they possess but do not reveal. Thus, the primary effect of anxiety appears to be the suppression of partial information. The lack of a significant correlation between anxiety and guessing scores in the Cross and Frary (1977) study, however, tends to

indicate that this effect is not activated except where there is extreme, and possibly maladjusted, expression of anxiety.

The second important effect of personality variables on test scores appears to be in increasing the susceptibility to misinformation penalties among self-confident examinees. Self-confident examinees, apparently, are willing to take greater risks than less confident examinees and express this tendency by claiming greater confidence in answers than is warranted by the certainty of their information. Thus, because they mark a high percentage of their answers in this way, they are more likely to receive a penalty for misinformation than more cautious examinees who answer fewer questions with highest confidence.

Similar effects are found in studies using the CMBS response mode mode. Hritz and Jacobs (1970) reported that the behavior of students stratified by their performance on a measure of risk-taking was highly predictive of their behavior on legitimate CMBS items. The various strata did not perform differently, however, when required to answer in the NR mode. Thus, the implication is that differences in risk-taking appeared only under a condition of increased risk, when penalties greater than zero were imposed for wrong responses.

Continuing this line of research, Jacobs (1975) looked at the effects of increased penalty levels on Slakter's (1967) risk-taking measure and the number of item options responded to on the legitimate test items. The 87 examinees in the study were randomly assigned to three treatments derived from variations of the CMBS response mode.

For the first treatment, the usual CMBS instructions, indicating credits for all partial information and the penalty for marking the correct answer, were given. For the second treatment, the "fair score" developed by Arnold and Arnold (1970) was explained. As in the CMBS scoring rule, the assigned item scores are the same as the expected item scores, but the amounts of credit and penalty for the "fair score" are the same as those given in the standardized NR/CG column in Table 1, rather than the standardized AUC/CMBS scores. In the third treatment, the instructions used by Arnold and Arnold (1970) in their study, in which the penalty for guessing was unspecified but indicated as bringing a zero expected gain, were given. These sets of instructions constituted the low penalty condition. Under the increased penalty condition, the examinees were given a second, parallel test, and the penalties doubled under the first two treatments. For the third treatment (the Arnold and Arnold instructions), the penalty was again unspecified, but examinees were told that guessing would result in an expected loss, rather than a zero gain.

For all three treatment groups, the increased penalty, whether specified or unspecified, resulted in a decrease in both the mean number of "no-content" items attempted and the mean number of response attempts on the legitimate items. Thus, CMBS responses are also affected by different penalty conditions, with examinees tending to more conservative behavior under increased risk. Jacobs commented that "the question of the effect on test validity and the possible interaction with subject (attribute) variables needs investigation" (p. 28).

Coombs, Milholland, and Womer (1956) were concerned with individual differences in CMBS responses, although they did not characterize the variable which they studied as one of personality. They argued that the responses of an individual using the experimental method [CMBS]:

> are probably to some extent a function of his willingness to take a chance by going beyond his sure knowledge. It is conceivable that each individual, independently of his knowledge, sets up a criterion level of "degree of certainty of being right" which serves as a threshold for responding (Coombs, *et al.*, 1956, p. 29).

These investigators called this threshold the individual's "standard of assurance." The higher the standard of assurance, the more certain an individual must be that he is correct before he will respond.

Coombs *et al.* (1956) devised a criterion index of the standard of assurance from tests on which examinees responded both in the CMBS mode and in the NR mode. This criterion index was based on the difference between the NR score and a theoretical NR score. The theoretical NR score was determined from the patterns of marking distractors and represents the NR score expected if the examinee is expressing his true degree of knowledge in the CMBS responses. To the extent that the actual NR score exceeds the expected NR score, the individual has, for some reason, failed to appraise his level of information correctly, or has made inadequate use of his partial information. The greater the disparity, the higher the individual's standard of assurance, and the more information he has which is below his threshold of assurance.

The criterion index of the standard of assurance produced moderate internal consistency reliability coefficients for two of the tests used, but not for the third. Correlation coefficients between the criterion index and NR and CMBS scores ranged from .237 to -.255. The pattern of these relationships was not consistent, and, viewing the total test scores as measures of achievement, the authors concluded that the criterion index of the standard of assurance was unrelated to ability.

The criterion index can only be computed when both NR and CMBS responses are made on the same test. An attempt by Coombs *et al.* (1956) to find an estimate of the standard of assurance so that the measure could be used when NR scores were not available proved unsuccessful. The proposed estimate used the number of alternatives or distractors the individual correctly crossed out compared to the total number he crossed out. The difference between these two numbers represented the number of correct alternatives incorrectly marked, and was considered as an estimate of the standard of assurance. The more correct alternatives incorrectly crossed out, the lower the standard of assurance and, thus, the less sure an individual had to be before risking a response. The correlation coefficients between this estimate and the tests which were administered independently of the tests from which the estimate was computed ranged from -.342 to -.469. These correlations were statistically significant and from them, the authors concluded that the estimate, unlike the criterion, was related to ability.

In the definitions of information and misinformation score components presented in Table 1, the incorrect marking of the correct option in the CMBS response mode is considered indicative of a level of misinformation. If, as is customary in multiple-choice tests, each item has only one correct option, the maximum possible number of incorrectly marked correct options is the number of items on the test. Coombs *et al.* (1956) apparently defined the incorrect marking of the correct option as evidence of risk-taking, rather than as some level of "knowing" that a wrong option is correct or that the right option is incorrect. These types of misinformed responses appear to have been viewed by Coombs *et al.* as the opposite of partial information or as an overextension of partial information. From the results of their study, it appears that this supposition was not accurate. The finding that the criterion and the estimate of the criterion produced insignificant correlations with each other suggests that misinformation and partial information may operate somewhat independently in test scores.

## Summary

The studies reviewed in this chapter seem to suggest strongly that the use of response modes which require examinees to assess and report on the accuracy of their own information produces a somewhat different effect that the use of the NR response mode. Apparently, the imposition of a penalty greater than zero for wrong responses is accompanied by different response patterns that may be a function of

personality differences. However, the way in which personality variables affect the responses is not completely clear.

One reasonably consistent effect of the alternative response modes and scoring rules has been enhanced internal consistency reliability, although changes in reliability coefficients have not often been assessed for statistical significance. The most frequent effect on criterion-related validity, however, has been the reverse of the effect on reliability. To what the changes (or, in some cases, the lack of change) can be attributed is also not clear.

The primary difference between NR scores and the alternative response modes and scoring rules in Figure 1 (with the exception of CG) is in the measurement of partial information and the explicit awarding of credit for partial information and penalty for misinformation. This study seeks to explain the relationships among personality variables, test characteristics, and alternative response modes and scoring rules. The literature suggests that these are legitimate concerns which can be fruitfully pursued. It is in the context of these studies, with particular emphasis on the works of Jacobs (1971, 1975), Coombs *et al.* (1956), and Frary (in press) that this study was undertaken. The CMBS response mode was selected as the main vehicle for this investigation because of its sensitivity to the partial information and misinformation score components which are the major concerns of the study.

# CHAPTER III

# METHODS

## Data Collection

Data for this study were collected at Virginia Polytechnic Institute and State University, Blacksburg, Virginia, during the academic year 1978-79. A battery of six tests designed to provide information foraa longitudinal assessment of the undergraduate teacher training program was administered to 360 students by the College of Education. The experimental procedures for this study were incorporated into this existing testing program. At the conclusion of the testing sessions, students were offered an opportunity to have their experimental responses excluded from the data analysis. None chose to do so.

The tests were given to six groups of students: three groups of student teachers; two groups of freshmen; and, one group of sophomores. The student teachers received all six instruments in the battery of tests during one session of approximately four hours length. The freshmen and sophomores were tested in two sessions of approximately two and a half hours each, held on consecutive nights. Of the 360 sets of responses obtained, 278 were usable for this study. Responses of examinees who failed to comply with all the experimental instructions or failed to complete at least 25 items in the experimental response mode were eliminated. Students who were not present for the group testing sessions and took the tests individually or in

small "make-up" groups were also excluded.

## Instruments

Four of the six instruments administered to these groups were used in this study. The instruments used were the Missouri College English Test (Callis & Johnson, 1964), the Adjective Check List (Gough, 1952), the Rokeach Dogmatism scale (Rokeach, 1960), and the Rotter Internal-External Locus of Control scale (Rotter, 1966). The instrument for which the experimental instructions were given was the first 25 items of the Missouri College English Test (MOCOL). These 25 items are part of a section of 60 items on which examinees indicate whether there is an error in punctuation, an error in capitalization, an error in spelling, an error in grammar, or no error in underlined phrases presented in paragraphs. The convention score is a sum of the correct responses. For the 25 items in MOCOL, there were four punctuation errors, five capitalization errors, seven errors in spelling, five errors in grammar, and four items which did not contain an error.

Personality dimensions were measured by the Adjective Check List (ADJ). The ADJ consists of a list of 300 positive and negative words describing human behaviors and attributes. Respondents select and mark those words which they perceive as applying to themselves. Twenty-four scales are derived from linear combinations of the adjectives chosen. Because raw scores on these scales are, in part, a function of the total number of adjectives checked, the raw scores were converted to standard scores, using tables given in the test manual

(Gough & Heilbrun, 1965). Among the ADJ scales are measures of personal adjustment, self-confidence, ascendancy, and need for achievement.

Although the ADJ does not directly parallel either the A-scale used by Sherriffs and Boomer (1954) and Cross and Frary (1977) or the CPI scales used by Jacobs (1971), there are a number of significant correlations between these scales and the ADJ (Gough & Heilbrun, 1965). These relationships are reported in Table 2. Correlations greater than or equal to an absolute value of .20 are significant at $p < .05$. These results suggest similarity between the ADJ scales and the A-scale and the four CPI scales, but they also indicate that the ADJ scales measure dimensions not included in the other instruments. Because of the differences, the ADJ is assumed to provide a broader representation of personality characteristics than is present in the previous studies.

The Rotter Internal-External Locus of Control Scale (ROTTER) is a measure of "the perception of events, whether positive or negative, as being a consequence of one's own actions and thereby potentially under personal control" (Lefcourt, 1976, p. 29). The scale is composed of 29 forced choice items, of which six items are fillers. The possible score range is from 0 to 23. Lower scores suggest belief in personal control; higher scores reflect greater reliance on "luck," "fate," or some other agent of control lying outside the individual.

The Rokeach Dogmatism Scale (ROK) is designed to measure individual differences in openness or closedness of belief systems. The

Table 2. Correlations between Scales of the Adjective Check List and Scales Used in Previous Studies for a Sample of 100 Males (Gough & Heilbrun, 1965, p. 28).

| ADJ Scales | CPI Scales[1] | | | | A-Scale[2] |
|---|---|---|---|---|---|
| | Do | Sy | Sa | Ie | |
| No. Adjectives Checked | .17 | .29 | .25 | -.17 | .04 |
| Defensiveness | .27 | .18 | .04 | .14 | -.42 |
| No. Favorable Adjectives | .34 | .28 | .12 | .26 | -.40 |
| No. Unfavorable Adjectives | -.11 | -.06 | .10 | -.21 | .29 |
| Self-Confidence | .57 | .47 | .38 | .22 | -.37 |
| Self-Control | -.04 | -.15 | -.20 | .17 | -.16 |
| Lability | .12 | .32 | .27 | .09 | .06 |
| Personal Adjustment | .21 | .15 | -.05 | .30 | -.30 |
| Need for Achievement | .49 | .24 | .21 | .26 | -.40 |
| Need for Dominance | .60 | .38 | .33 | .26 | -.41 |
| Need for Endurance | .29 | .05 | .01 | .20 | -.36 |
| Need for Order | .17 | -.04 | -.11 | .17 | -.29 |
| Need for Intraception | .20 | .16 | -.03 | .30 | -.28 |
| Need for Nurturance | .03 | .04 | -.11 | .16 | -.14 |
| Need for Affiliation | .21 | .18 | .01 | .18 | -.28 |
| Need for Heterosexuality | .10 | .20 | .10 | -.14 | -.13 |
| Need for Exhibition | .40 | .37 | .41 | -.05 | -.09 |
| Need for Autonomy | .33 | .16 | .29 | .18 | -.14 |
| Need for Aggression | .22 | .12 | .28 | -.09 | .07 |
| Need for Change | .30 | .35 | .41 | .03 | -.19 |
| Need for Succorance | -.31 | -.12 | -.10 | -.27 | .43 |
| Need for Abasement | -.53 | -.38 | -.34 | -.18 | .34 |
| Need for Deference | -.42 | -.35 | -.34 | -.06 | .18 |
| Counseling Readiness | -.45 | -.41 | -.30 | -.06 | .33 |

[1]California Psychological Inventory Scales used by Jacobs (1971). Do=Dominance; Sy=Sociability; Sa=Self-acceptance; Ie=Intellectual Efficiency.

[2]A-Scale from Minnesota Multiphasic Personality Inventory used by Sherriffs and Boomer (1954) and Cross and Frary (1977).

extent to which a person's belief system is open is "the extent to which the person can receive, evaluate, and act on relevant information received from the outside on its own intrinsic merits, unencumbered by irrelevant factors in the situation arising from within the person or from the outside" (Rokeach, 1960, p. 57). On this scale, examinees respond to 40 statements on a 6-point Likert scale, ranging from 0 (strongly disagree) to 5 (strongly agree). The scores can range from 0 to 200.

## The Response Modes and Protocol

The examinees were instructed to respond to the first section of the Missouri College English Test in three ways. One response mode was to indicate, in the conventional manner, the single option believed correct for each item. The same set of options (spelling, punctuation, *etc.*) was used for all the questions in this section. The second response mode required writing the needed correction in the test booklet. The third way of responding was with the response mode proposed by Coombs (1953). Under this response mode, examinees are instructed to avoid marking the right answer and to indicate in some specified manner all those item options which they believe are wrong. Since the emphases of the conventional number-right response (NR) and the Coombs mode (CMBS) are opposite, *i.e.*, NR asks for the single right answer and CMBS for all wrong options, they are easily used on the same answer sheet.

In this study, examinees were asked to indicate item options they

believed were wrong by placing cross marks over them; the right answer was indicated by darkening the appropriate circle on the answer sheet. The sequence which they were to follow in using these three response modes was:

1. Read the item and then place cross marks on all the options believed to be wrong.
2. Select the single right answer from among the options (or option, if four options have been crossed out) not marked and darken the corresponding circle.
3. Make the appropriate correction in the test booklet.
4. Go on to the next item.

The initial instructions to the examinees implied that the three response modes would be used for all 60 items in the first section of the test. After approximately 30 minutes, however, the examinees were told that the multiple responses need be used only for the first 25 items and that they could complete the remaining items in the conventional manner by simply filling in the spaces corresponding to the answers.

Two penalty conditions were imposed. The first group of freshmen and the first group of student teachers tested were told that the CMBS mode was an attempt to determine the thought processes followed in making decisions about the answers marked right. No penalty was announced or implied. The emphasis was on the cooperative giving of information.

The remaining four groups of students were informed that the three response modes were being used to determine the best way to

score their tests, and that they would receive three separate scores, based, respectively, on the three sets of responses. Under the NR mode, they would receive four points for each correct answer and zero points for incorrect responses; the same scoring rule would be applied to corrections in the test booklet. The examinees were advised that their best strategy under these scoring rules was to guess even if they were not sure of the correct answer.

Under the CMBS mode, the examinees were told that they would receive one point for each correctly identified wrong option (distractor), but that they would be penalized four points if they marked the right answer as a distractor. The effects of various patterns of marking responses were demonstrated and the students were instructed not to guess if they were not sure an answer option was a distractor. The emphasis of this explanation was on allowing students to receive at least some credit for whatever information they might have about an item even when that information was partial and not complete.

Questions were answered from the groups during and after the explanations of the response modes, and short individual explanations were given during the first few minutes of work on the test when necessary. Although the Missouri College English Test is designed to be administered in 40 minutes, examinees were assured that they would be allowed ample time to complete the test with the added responses. Most examinees were able to finish in under an hour.

Because for all groups of students, except the sophomores, the primary purpose of the tests was to provide a data base from which a

longitudinal assessment of the teacher training program could be made, it was necessary to assume that college students had been socialized to perform their best on tests administered in a formal setting. The sophomores, however, were being tested as part of their application to the professional studies component of the program which would begin in the junior year. They were the only group whose scores would be reported individually to their course advisors. Consequently, the sophomores were assumed to be operating under a greater condition of risk than the freshmen and student teachers.

## Scores

The use of the CMBS mode allows the categorization of each item according to the amount of information or misinformation reported by the examinee. The categories for 5-option items used in this study are as follows:

Complete Information (INFO) - Marking out all four incorrect choices. INFO assumes that there is no guessing as to the answer and that the examinee has complete confidence in his choice.

Guessing among two alternatives (G2) - Marking out three of the four incorrect choices, but not the right choice. G2 indicates that the examinee could narrow his choice of the correct answer to two of the options and suggests that a forced choice among those two options would be made with reasonable confidence.

Guessing among three alternatives (G3) - Marking out two of the four incorrect choices, but not the right choice. G3 indicates that the examinee could narrow his choice of the correct answer to only three of the options and, thus, would probably not be very confident of a forced selection of a correct answer.

Guessing among four alternatives (G4) - Marking out only one of

the four incorrect choices, but not the right choice. G4 suggests very little information about the item and probably very low confidence in a forced selection of a right answer.

Ignorance (IG) - Marking out none of the incorrect choices or the right choice. IG indicates that the examinee has no information about the item and would probably have no confidence in a forced right choice.

Partial Misinformation Pattern 1 (MIS1) - Marking out three incorrect choices and the right choice. MIS1 indicates that the examinee has both information and misinformation about the item, but is very confident that the choice he has not marked out is the correct answer.

Partial Misinformation Pattern 2 (MIS2) - Marking out two incorrect choices and the right choice. MIS2 indicates that the examinee is not completely confident about the right answer and would have to guess among the two choices he has left unmarked. However, he is confident that the correct choice is not the right answer.

Partial Misinformation Pattern 3 (MIS3) - Marking out only one incorrect choice and the right choice. MIS3 suggests that the examinee would not be very confident about a forced right choice from among the three options he has not marked, but he is reasonably confident that his misinformed choice is not the right answer.

Complete Misinformation (MIS4) - Marking out the right choice only. MIS4 suggests that the examinee would probably indicate very low confidence in a forced right choice, but that he is highly confident that his misinformed choice is not the right answer.

A number of total scores can be directly determined or inferred from these patterns of responses. The scores calculated in this study were:

1. Number Right Score (NR) - Sum of items for which single correct answer was chosen from among the five alternatives.

   NR = 1(Number of Right Answers) + 0(Number of Wrong or Omitted Answers)

2. Coombs Score (CMBS) - Sum of frequency of responses falling into the nine information-misinformation categories, weighted

proportionally to the Coombs scoring rule. The original weights were adjusted to assign 1.00 (rather than 4.00) to INFO and -1.00 (rather than -4.00) to MIS4.

$$\begin{aligned}CMBS = {} & (1.00)INFO + (.75)G2 + (.50)G3 + (.25)G4 + (0)IG \\ & + (-.25)MIS1 + (-.50)MIS2 + (-.75)MIS3 + (-1.00)MIS4\end{aligned}$$

3. Expected Number Right Score (ENR) - Sum of frequency of responses in the nine information-misinformation categories, with probability weights adjusted to give 0 probability for IG.

$$\begin{aligned}ENR = {} & (1.00)INFO + (.38)G2 + (.17)G3 + (.06)G4 + (0)IG + \\ & (-.25)MIS1 + (-.25)MIS2 + (-.25)MIS3 + (-.25)MIS4\end{aligned}$$

4. Inferred Confidence Weighting Score (ICW) - Sum of weighted responses, based on the assumption that the number of marked options reflects the confidence an examinee would have in each choice of right answer.

$$\begin{aligned}ICW = {} & (1.00)INFO + (.75)G2_R + (.50)G3_R + (.25)G4_R + (0)IG \\ & + (-1.00)MIS1 + (-.75)MIS2 + (-.75)G2_W + (-.50)MIS3 \\ & + (-.50)G3_W + (-.25)MIS4 + (-.25)G4_W\end{aligned}$$

where $G2_R$, $G3_R$, and $G4_R$ indicate guessing successfully; and $G2_W$, $G3_W$, and $G4_W$ indicate unsuccessful guessing.

5. Empirical Choice Weighted Score (EMP) - Sum of the empirical weights of the options chosen. The weights were determined using the average NR score as the criterion. Although most of the data analyses were conducted separately on the three groups of examinees, the empirical choice weights were, for increased stability, computed from the total group. Using the group mean and standard deviation, $z$-scores were determined for examinees. The empirical weight for each option in each item was obtained by averaging the $z$-scores of the examinees choosing the option. The EMP score for each examinee was the sum of the positive and negative empirical choice-weights attached to the options which he chose. For comparison with other types of scores, these sums of average $z$-scores were reconverted to standard scores, based on the original NR mean and standard deviation.

# CHAPTER IV

## RESULTS

For the data analysis, the 278 examinees were considered to constitute three separate groups because of differences in the penalties imposed and in the personal consequences of the scores for portions of the total group. Group 1 was composed of freshmen, transfer students, and student teachers for whom the CMBS response mode carried no penalty. Group 2 included freshmen, transfer students, and student teachers who were informed of the penalty under the CMBS mode and were advised not to guess under this mode, but whose scores would not be used for decision-making about them individually. Group 3 was composed of sophomores who were operating under the same instructions as group 2, but who were aware that their scores on the Missouri College English Tests, of which the 25-item MOCOL was a part, could be used in the determination of their eligibility for upper division professional studies.

One of the factors affecting test reliability is the difficulty of the items making up the test. Other things being equal, a test with a greater proportion of items which are passed and failed by equal numbers of examinees (item difficulty = .50) will have a higher internal consistency reliability coefficient than a test with large proportions of very easy items (item difficulty > .50) or very difficult (item difficulty < .50) items. The item difficulties for the

responses of the three groups of examinees in this study on the 25 items of the MOCOL are presented in Tables 3, 4, and 5. Also given are the proportions of information (INFO) reported by examinees (marking out four distractors), proportions of partial information (combined G2, G3 and G4), proportions of ignorance (IG), and proportions of misinformation (combined MIS1, MIS2, MIS3 and MIS4).

An examination of Tables 3, 4, and 5 indicates that group 1 found the test somewhat more difficult than groups 2 and 3, and that all of the items were answered correctly by a larger proportion of the examinees in group 3 than in groups 1 and 2. In group 1, nine of the items were missed by more than 40 percent of the examinees; in group 2, seven of the items proved to be this difficult; in group 3, only five items.

The relative difficulty of the test was reflected in the mean NR scores for the three groups. Only the means for groups 1 and 3 proved to be significantly different at the .05 level ($t$ = -2.91, $df$ = 190, $p$ = .004). Although group 3 appeared to be the most homogeneous of the three groups, there were no significant differences in the variances of the scores at the .05 level. Since the emphasis here was on demonstrating the null hypothesis, care must also be taken to avoid a Type II error. No significant differences in variances could be detected even at $p$ = .15, but the means for groups 2 and 3 were different at the .11 level.

Among the factors enhancing test reliability are items of middle difficulty, as discussed previously, and administering the test

Table 3. Item Difficulties and Proportions of Information, Partial Information, Ignorance and Misinformation in Items of MOCOL for Group 1 ($n$=69).

| Item | Prop.[1] Right | Prop. Info. | Prop. Par. Info. | Prop. Ig. | Prop. Misin. | Prop. Wrong |
|---|---|---|---|---|---|---|
| 1 | .36 | .29 | .20 | .10 | .41 | .64 |
| 2 | .67 | .51 | .14 | .15 | .20 | .33 |
| 3 | .90 | .83 | .06 | .03 | .09 | .10 |
| 4 | .67 | .51 | .12 | .14 | .23 | .33 |
| 5 | .86 | .77 | .09 | .03 | .12 | .14 |
| 6 | .74 | .52 | .14 | .14 | .19 | .26 |
| 7 | .75 | .61 | .16 | .04 | .19 | .25 |
| 8 | .68 | .49 | .25 | .06 | .20 | .32 |
| 9 | .46 | .26 | .26 | .17 | .30 | .54 |
| 10 | .83 | .65 | .14 | .07 | .13 | .17 |
| 11 | .64 | .52 | .10 | .07 | .30 | .36 |
| 12 | .71 | .58 | .12 | .10 | .20 | .29 |
| 13 | .80 | .67 | .10 | .04 | .19 | .20 |
| 14 | .65 | .52 | .17 | .10 | .20 | .35 |
| 15 | .55 | .39 | .23 | .09 | .29 | .45 |
| 16 | .78 | .67 | .10 | .06 | .17 | .22 |
| 17 | .41 | .30 | .16 | .10 | .43 | .59 |
| 18 | .54 | .46 | .07 | .09 | .38 | .46 |
| 19 | .78 | .71 | .06 | .04 | .19 | .22 |
| 20 | .57 | .49 | .12 | .04 | .35 | .43 |
| 21 | .45 | .39 | .09 | .07 | .45 | .55 |
| 22 | .70 | .47 | .25 | .15 | .13 | .30 |
| 23 | .41 | .19 | .30 | .13 | .38 | .59 |
| 24 | .55 | .36 | .19 | .14 | .30 | .45 |
| 25 | .67 | .55 | .13 | .07 | .25 | .33 |
| Averages | .65 | .51 | .15 | .09 | .25 | .35 |

Mean NR Score 16.10, standard deviation 3.85

[1]Item Difficulty

Table 4. Item Difficulties and Proportions of Information, Partial Information, Ignorance and Misinformation in Items of MOCOL for Group 2 ($n$=86).

| Item | Prop.[1] Right | Prop. Info. | Prop. Par. Info. | Prop. Ig. | Prop. Misin. | Prop. Wrong |
|---|---|---|---|---|---|---|
| 1 | .45 | .37 | .21 | .00 | .42 | .55 |
| 2 | .72 | .57 | .24 | .01 | .17 | .28 |
| 3 | .91 | .87 | .05 | .00 | .08 | .09 |
| 4 | .51 | .47 | .09 | .01 | .43 | .49 |
| 5 | .86 | .79 | .08 | .01 | .12 | .14 |
| 6 | .79 | .64 | .17 | .03 | .15 | .21 |
| 7 | .63 | .55 | .16 | .01 | .28 | .37 |
| 8 | .73 | .54 | .21 | .03 | .22 | .27 |
| 9 | .47 | .28 | .34 | .05 | .34 | .53 |
| 10 | .78 | .71 | .09 | .01 | .19 | .22 |
| 11 | .67 | .63 | .10 | .01 | .26 | .33 |
| 12 | .69 | .57 | .10 | .03 | .29 | .31 |
| 13 | .83 | .70 | .16 | .01 | .13 | .17 |
| 14 | .69 | .53 | .27 | .00 | .20 | .31 |
| 15 | .78 | .69 | .19 | .02 | .10 | .22 |
| 16 | .73 | .64 | .13 | .01 | .22 | .27 |
| 17 | .36 | .28 | .21 | .00 | .51 | .64 |
| 18 | .54 | .43 | .12 | .02 | .43 | .46 |
| 19 | .87 | .77 | .12 | .00 | .12 | .13 |
| 20 | .70 | .59 | .09 | .01 | .30 | .30 |
| 21 | .58 | .51 | .14 | .03 | .31 | .42 |
| 22 | .79 | .63 | .20 | .03 | .14 | .21 |
| 23 | .38 | .22 | .31 | .03 | .43 | .62 |
| 24 | .62 | .45 | .27 | .01 | .27 | .38 |
| 25 | .76 | .65 | .12 | .02 | .21 | .24 |
| Averages | .67 | .56 | .17 | .02 | .25 | .33 |

Mean NR Score 16.83, standard deviation 3.72

[1]Item Difficulty

Table 5. Item Difficulties and Proportions of Information, Partial Information, Ignorance and Misinformation in Items of MOCOL for Group 3 ($n$=123)

| Item | Prop.[1] Right | Prop. Info. | Prop. Par. Info. | Prop. Ig. | Prop. Misin. | Prop. Wrong |
|---|---|---|---|---|---|---|
| 1 | .53 | .42 | .25 | .02 | .31 | .47 |
| 2 | .68 | .39 | .41 | .03 | .16 | .33 |
| 3 | .93 | .86 | .07 | .01 | .07 | .07 |
| 4 | .58 | .49 | .18 | .01 | .33 | .42 |
| 5 | .97 | .93 | .05 | .00 | .02 | .03 |
| 6 | .78 | .48 | .43 | .00 | .09 | .22 |
| 7 | .69 | .54 | .24 | .01 | .21 | .31 |
| 8 | .81 | .54 | .33 | .00 | .13 | .19 |
| 9 | .55 | .24 | .46 | .02 | .28 | .45 |
| 10 | .90 | .70 | .22 | .00 | .08 | .10 |
| 11 | .73 | .59 | .18 | .02 | .21 | .27 |
| 12 | .78 | .61 | .19 | .01 | .20 | .22 |
| 13 | .83 | .64 | .25 | .01 | .10 | .17 |
| 14 | .63 | .33 | .45 | .02 | .21 | .37 |
| 15 | .68 | .44 | .41 | .01 | .15 | .32 |
| 16 | .72 | .45 | .33 | .02 | .21 | .28 |
| 17 | .47 | .30 | .31 | .02 | .37 | .53 |
| 18 | .64 | .50 | .21 | .00 | .28 | .36 |
| 19 | .91 | .85 | .08 | .00 | .07 | .09 |
| 20 | .62 | .52 | .16 | .00 | .32 | .38 |
| 21 | .67 | .55 | .20 | .02 | .24 | .33 |
| 22 | .74 | .52 | .34 | .02 | .12 | .26 |
| 23 | .42 | .20 | .41 | .01 | .38 | .58 |
| 24 | .61 | .39 | .41 | .00 | .20 | .39 |
| 25 | .81 | .65 | .21 | .00 | .14 | .19 |
| Averages | .71 | .53 | .27 | .01 | .20 | .29 |

Mean NR Score 17.65, standard deviation 3.35

[1]Item Difficulty

to a group with a wide range of ability rather than to a more homogeneous group (Ebel, 1972, p. 427). Since the mean item difficulties in groups 1 and 2 were closer to .50 than in group 3, and, in addition, the variances of the NR scores in groups 1 and 2 were larger than the variance in group 3, it was expected that internal consistency reliability, estimated by Cronbach's alpha, would be higher in groups 1 and 2 than in group 3. This expectation was supported by the data. For group 1, the alpha coefficient was .67; group 2, .66; group 3, .60. An approximation of the *F*-statistic, designed to test for differences in independent alpha coefficients (Feldt, 1969), showed no significant differences between the pairs of reliability estimates, however. Thus, under the NR scoring mode, group 1 scored significantly lower than group 3 at the .05 level and group 2 was lower than group 3 at the .11 level, but there were no differences in the variances or the alpha reliability coefficients among the groups. For all three groups, the test tended to be of moderate to low difficulty.

It is also evident from Tables 3, 4, and 5 that both partial information and misinformation were operative in the responses of examinees to the test questions. For each of the 25 items, some part of the patterns of marking distractors under the CMBS mode could be categorized as indicative of guessing based on partial information or of responses from misinformation. There also appeared to be a greater occurrence of IG (marking none of the item alternatives) in group 1 than in the other groups. This result was not unexpected and may be

accounted for by the greater personal consequences of the test scores for examinees in group 3, which may have caused them to attempt to gain even small numbers of CMBS points by marking any distractors they felt sure about, rather than accepting a certain zero CMBS score by marking no options. Because of the relatively small sizes of the groups, however, some of the smaller proportions given represent very individual responses to items.

## Information and Misinformation Score Components

A comparison of the mean number of items falling into the nine score components for the three groups is presented in Table 6. For these data, the total responses of examinees were viewed as divisible into the types of responses indicated under the CMBS mode. Thus, each examinee had an INFO score, which was the number of items on which he correctly marked four distractors; a G2 score, the number of items with three correctly marked distractors; a G3 score, *etc.*, until all 25 items were placed into the information or misinformation component described by the examinee's CMBS responses. The mean INFO score is the sum of all the INFO scores made by examinees in a group divided by the number in that group. The remaining information and misinformation category scores were determined in the same way. *T*-tests were used to examine differences between the means because of unequal variances among the groups.

There are relatively few significant differences at the .05 level between the average numbers of items on which examinees in the three

Table 6. *T*-tests for Differences in Mean Number of Items in Information and Misinformation Score Components.

| Score Component | Group | N | Mean | S.D. | *t* | d.f. | prob. |
|---|---|---|---|---|---|---|---|
| *INFO* | 1 | 69 | 12.72 | 5.05 | -1.64 | 153 | .103 |
| | 2 | 86 | 14.07 | 5.09 | | | |
| | 1 | 69 | 12.72 | 5.05 | -0.56 | 190 | .574 |
| | 3 | 123 | 13.14 | 4.79 | | | |
| | 2 | 86 | 14.07 | 5.09 | 1.35 | 207 | .179 |
| | 3 | 123 | 13.14 | 4.79 | | | |
| *G2* | 1 | 69 | 3.03 | 3.02 | -0.31 | 152.96[1] | .759 |
| | 2 | 86 | 3.20 | 3.83 | | | |
| | 1 | 69 | 3.03 | 3.02 | -4.28 | 190 | .001 |
| | 3 | 123 | 5.12 | 3.38 | | | |
| | 2 | 86 | 3.20 | 3.83 | -3.84 | 207 | .001 |
| | 3 | 123 | 5.12 | 3.38 | | | |
| *G3* | 1 | 69 | 0.52 | 1.09 | -0.92 | 149.66[1] | .358 |
| | 2 | 86 | 0.72 | 1.59 | | | |
| | 1 | 69 | 0.52 | 1.09 | -3.87 | 185.96[1] | .001 |
| | 3 | 123 | 1.47 | 2.30 | | | |
| | 2 | 86 | 0.72 | 1.59 | -2.79 | 206.99[1] | .006 |
| | 3 | 123 | 1.47 | 2.30 | | | |
| *G4* | 1 | 69 | 0.20 | 0.93 | -0.34 | 153 | .734 |
| | 2 | 86 | 0.26 | 0.98 | | | |
| | 1 | 69 | 0.20 | 0.93 | 0.32 | 100.83[1] | .748 |
| | 3 | 123 | 0.16 | 0.61 | | | |
| | 2 | 86 | 0.26 | 0.98 | 0.78 | 129.62[1] | .436 |
| | 3 | 123 | 0.16 | 0.61 | | | |

Table 6 (continued). *T*-tests for Differences in Mean Number of Items in Information and Misinformation Score Components.

| Score Component | Group | N | Mean | S.D. | *t* | d.f. | prob. |
|---|---|---|---|---|---|---|---|
| *IG* | 1 | 69 | 2.25 | 4.08 | 3.45 | 85.99[1] | .001 |
| | 2 | 86 | 0.44 | 1.66 | | | |
| | 1 | 69 | 2.25 | 4.08 | 4.09 | 70.68[1] | .001 |
| | 3 | 123 | 0.22 | 0.76 | | | |
| | 2 | 86 | 0.44 | 1.66 | 1.16 | 110.42[1] | .248 |
| | 3 | 123 | 0.22 | 0.76 | | | |
| *MIS1* | 1 | 69 | 5.35 | 3.73 | 0.39 | 153 | .698 |
| | 2 | 86 | 5.12 | 3.65 | | | |
| | 1 | 69 | 5.35 | 3.73 | 3.68 | 110.02[1] | .001 |
| | 3 | 123 | 3.46 | 2.75 | | | |
| | 2 | 86 | 5.12 | 3.65 | 3.55 | 149.32[1] | .001 |
| | 3 | 123 | 3.46 | 2.75 | | | |
| *MIS2* | 1 | 69 | 0.83 | 1.06 | -0.26 | 153 | .796 |
| | 2 | 86 | 0.87 | 1.14 | | | |
| | 1 | 69 | 0.83 | 1.06 | -1.52 | 190 | .131 |
| | 3 | 123 | 1.11 | 1.31 | | | |
| | 2 | 86 | 0.87 | 1.14 | -1.34 | 207 | .182 |
| | 3 | 123 | 1.11 | 1.31 | | | |
| *MIS3* | 1 | 69 | 0.09 | 0.33 | -1.41 | 108.71[1] | .163 |
| | 2 | 86 | 0.24 | 0.97 | | | |
| | 1 | 69 | 0.09 | 0.33 | -3.34 | 189.99[1] | .001 |
| | 3 | 123 | 0.31 | 0.59 | | | |
| | 2 | 86 | 0.24 | 0.97 | -0.55 | 128.51[1] | .582 |
| | 3 | 123 | 0.31 | 0.59 | | | |

Table 6 (continued). *T*-tests for Differences in Mean Number of Items in Information and Misinformation Score Components

| Score Component | Group | N | Mean | S.D. | *t* | d.f. | prob. |
|---|---|---|---|---|---|---|---|
| *MIS4* | 1 | 69 | 0.01 | 0.12 | -1.08 | 94.76[1] | .282 |
| | 2 | 86 | 0.08 | 0.56 | | | |
| | 1 | 69 | 0.01 | 0.12 | 0.38 | 111.38[1] | .703 |
| | 3 | 123 | 0.01 | 0.09 | | | |
| | 2 | 86 | 0.08 | 0.56 | 1.21 | 88.12[1] | .230 |
| | 3 | 123 | 0.01 | 0.09 | | | |

[1]Unequal variances. Approximate *t* with adjusted degrees of freedom.

groups reported their perceptions of their own information and misinformation levels. In general, group 3 tended to differ more frequently from group 1 than from group 2, but group 3 also differed from group 2 in several categories. Group 1 showed a significant difference from both groups 2 and 3 only on IG. This difference may represent a geniune perception by some of the examinees in group 1 as to the state of their knowledge, but, since group 1 was operating under no penalty for failing to cooperate with the directions, the difference may also represent a tendency to ignore or misunderstand the directions for responding. In no case were all three groups different from each other in a single information or misinformation category.

Because of the significant difference in mean NR scores between groups 1 and 3, it was expected that there would also be a significant difference in the INFO scores between these two groups. That this did not occur suggests that examinees in group 3 did not claim the highest possible CMBS score on items which they could, nevertheless, answer correctly. This same tendency is reflected in the larger G2 and G3 score components for group 3 as compared to groups 1 and 2.

The fewer examinees in group 3 who marked four item options are also reflected in MIS1 scores. Group 3 examinees exhibited less of this misinformation type than did the other two groups. There were no differences on MIS2 and MIS4, and group 3 differed only from group 1 on MIS3.

The many significant differences between the variances of the

groups reflect differences in the utilization of the available responses for the groups. There appears to be no consistent pattern in the differences, however.

Of concern also are the reliabilities of the information and misinformation category scores. These are shown in Table 7. Alpha reliability will be used throughout this report, except where specifically noted. The computation formula for alpha reliability is:

$$r_{xx} = \frac{k}{k - 1}\left(1 - \frac{\Sigma s_i^2}{s_x^2}\right)$$

where $r_{xx}$ is the alpha reliability coefficient, $k$ is the number of items on the test, $s_i^2$ is the variance of an item on the test, and $s_x^2$ is the variance of the total test scores.

For all three groups of examinees there were items on the test for which some of the available response categories were not used, or were used only by one individual. This produced zero variances for those items within a given response category, *e.g.*, if no examinees in group 1 marked three options including the right answer (MIS3) on item 3, the variance for that item on the MIS3 score would be zero. From the formula for alpha reliability, it can be seen that, when there are sparse responses, not only can high variability among the scores of a few individuals unduly affect the computation of the coefficient, zero variances also reduce effective test length, making the weight $k(k-1)$ inappropriate. As can be seen from the mean information and

Table 7. Alpha Reliability Coefficients for Information and Misinformation Category Scores by Groups.

| Group | *n* | INFO | G2 | G3 | G4 | IG | MIS1 | MIS2 | MIS3 | MIS4 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 69 | .81 | .74 | .60 | .80[1] | .91 | .74 | .30 | ---[2] | ---[2] |
| 2 | 86 | .82 | .85 | .76 | .77 | .88 | .74 | .37 | .77[1] | .77[1] |
| 3 | 123 | .79 | .68 | .77 | .59[1] | .65 | .64 | .41 | .13[1] | ---[2] |

[1]Between 10 and 20 items with zero variances.
[2]More than 20 items with zero variances.

misinformation scores in Table 6, only for INFO, G2, and MIS1 were there average scores for all groups greater than one, suggesting that the item variances for many items may represent very few individuals, and different individuals from those included in the variances for other items. Consequently, most of these reliability coefficients must be viewed with caution, since the number of items for which there was greater than zero variance differs considerably from one score set to another.

The INFO reliability coefficients, which represent the largest number of examinee responses, indicate the consistency of CMBS responses made under complete information. As expected, the INFO coefficients are higher than the corresponding NR coefficients, since NR scores contain the less reliable guessing and misinformation components. In general, it appears that the expressions of partial information were more consistent than the expressions of misinformation. Again, however, because of the sparsity of responses in some categories, the interpretation must be considered tentative.

## Reliability and Validity of Total Scores

One of the major concerns of this study is the effect of the score components of information and misinformation on the reliability and validity coefficients of the total scores on the 25-item test. The authors of the Missouri College English Test (Callis & Johnson, 1965) reported a split-half internal consistency reliability coefficient, corrected by the Spearman-Brown formula, for the 90-item test

of .94. This coefficient was based on the responses of 3,381 male and female university students, who showed a mean total test score of 58.3, s.d. of 14.7. The mean item difficulty reported by Callis & Johnson (1965) for the total Missouri College English Test was .61, somewhat lower than the mean item difficulty observed for the first 25 items used in this analysis ($\bar{p}$ = .68).

The reliability coefficients for NR scores for the groups in this study were not expected to be as large as .94 for several reasons. One, the groups used for the calculations were far smaller than the norm group and chosen from only one college in the university. Consequently, the attenuating effect of greater group homogeneity was anticipated. Second, the experimental test was far shorter than the 90-item total test. However, when the Spearman-Brown correction formula was applied to the reliability coefficients for the 25-item test in the three groups, and the reliability estimates, thus, projected to those of a test 3.6 times as long (90 items/25 items), the resulting reliability coefficients were .88 (group 1), .87 (group 2), and .84 (group 3). These results suggest that, were the test lengthened, the responses would be similar in internal consistency to those of the norm group. Third, the 25 items making up the experimental test were somewhat easier for the three groups in this study than for the sample of university students in the norm group. The expectation of lower reliability for the experimental test was confirmed, since the NR reliability coefficients for the three groups were all less than .70.

Descriptive statistics and intercorrelations of the test scores

under the five scoring rules used in this study are presented in Table 8. The differences in mean scores reflect the relative emphases placed on partial information and misinformation under the different scoring rules. The very low scores for ICW show the effects of the reverse of the CMBS penalty for misinformation. Under CMBS, some credit is given for partial information, even though the right answer is among the marked options. Under the assumption used to derive ICW scores, these items received the greatest penalty, since they are taken to be wrong answers about which the examinee has the greatest confidence. While it is possible that, under an actual confidence-weighted response mode and scoring rule administration of the test, examinees may have been more sparing in their indications of the high confidence imputed to their choices with the derived scores, the comparison of CMBS with ICW underscores the differences in penalties and credits given by the two rules. Highly confident examinees are more susceptible to severe misinformation penalties under ICW than under CMBS. The very high correlations between NR scores and ICW scores also demonstrate the similarities in expected score gain for partial information and penalty for misinformation between these two systems (see Table 1).

The EMP scores, which have been standardized in Table 8 to facilitate comparison with other scores, have the highest reliability coefficients of the five scores, but, from the magnitudes of their correlations with NR scores, the difference between the two scoring rules does not materially affect the relative standing of examinees within

Table 8. Means, Standard Deviations, Alpha Reliability Coefficients, and Intercorrelations for NR, ONR, CMBS, ICW, and EMP scores by Group.

| | Mean | S.D. | Alpha Reliability | Correlation | | | |
|---|---|---|---|---|---|---|---|
| | | | | ONR | CMBS | ICW | EMP |
| GROUP 1 | $n$=69 | | | | | | |
| NR | 16.10 | 3.85 | .67 | .87 | .87 | .97 | .96 |
| ONR | 14.55 | 3.92 | .69 | | .79 | .85 | .83 |
| CMBS | 13.48 | 5.26 | .76 | | | .90 | .85 |
| ICW | 7.55 | 6.90 | .68 | | | | .94 |
| EMP | 16.21 | 3.71 | .79 | | | | |
| GROUP 2 | $n$=86 | | | | | | |
| NR | 16.83 | 3.72 | .66 | .89 | .95 | .98 | .96 |
| ONR | 16.06 | 3.87 | .69 | | .84 | .90 | .84 |
| CMBS | 14.91 | 5.12 | .75 | | | .94 | .94 |
| ICW | 8.92 | 6.72 | .66 | | | | .93 |
| EMP | 16.74 | 3.91 | .82 | | | | |
| GROUP 3 | $n$=123 | | | | | | |
| NR | 17.65 | 3.35 | .60 | .87 | .90 | .98 | .95 |
| ONR | 16.33 | 3.47 | .63 | | .84 | .87 | .84 |
| CMBS | 16.10 | 3.96 | .63 | | | .94 | .88 |
| ICW | 10.27 | 6.03 | .62 | | | | .94 |
| EMP | 17.66 | 3.63 | .78 | | | | |

their groups. The choice weights attached to the item options by the EMP scoring rule are shown in Table 9. For all items, except item 5, the largest positive choice weights coincided with the keyed right answers. The cause of the high choice weight for a wrong answer in item 5 was found to lie with the responses of only two examinees out of the 278 from which the choice weights were calculated. These two examinees, who chose option 3 on item 5, earned NR scores of 20 and 18, somewhat above the mean of 17.01 (s.d. 3.62) of the whole group. The average of their $z$-scores was greater than the average $z$-score of the 252 examinees who correctly chose option 2. Consequently, the choice weight for that item was unduly affected by the selection of these examinees.

There were significant differences between mean scores for groups 1 and 3 under all scoring rules, group 3 scores being higher. There was also a significant difference between mean scores for groups 2 and 3 with the ICW scoring rule, and between groups 1 and 2 on ONR scores, using $t$-tests. $F_{max}$ tests also showed group 3 more homogeneous than either group 1 or group 2 when the CMBS and ICW scoring rules were used. There were no differences among group variances with the other scoring rules and no differences between groups 1 and 2.

## Tests for Differences in Reliability Coefficients Under Different Response Modes

From Table 8, it is obvious that there are fluctuations in reliability coefficients when different scoring rules are used. In terms of absolute magnitude, the expectation that the CMBS mode would

Table 9. Empirical Choice Weights and Keyed Right Answers Based on the Responses of 278 Examinees.

| Item | Options | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 Capital | | 2 Grammar | | 3 Punctuation | | 4 Spelling | | 5 No Error | |
| 1 | -0.72 | 5[1] | -0.08 | 11 | -0.92 | 3 | 0.54* | 130 | -0.49 | 129 |
| 2 | -0.92 | 6 | -0.54 | 46 | -0.45 | 30 | -0.94 | 5 | 0.25* | 191 |
| 3 | -1.94 | 1 | -1.52 | 4 | -0.28 | 1 | 0.10* | 253 | -0.92 | 19 |
| 4 | -0.60 | 6 | -0.69 | 20 | -0.71 | 27 | 0.34* | 161 | -0.28 | 64 |
| 5 | -1.20 | 3 | 0.11* | 252 | 0.55 | 2 | -0.77 | 13 | -1.76 | 8 |
| 6 | -0.58 | 11 | -0.63 | 33 | -0.76 | 8 | -0.78 | 11 | 0.19* | 215 |
| 7 | -0.97 | 2 | 0.22* | 188 | -0.46 | 56 | -1.66 | 4 | -0.25 | 28 |
| 8 | -1.44 | 5 | -0.57 | 16 | 0.22* | 210 | -1.05 | 10 | -0.50 | 37 |
| 9 | -0.65 | 12 | -0.27 | 26 | -0.54 | 15 | -0.32 | 86 | 0.36* | 139 |
| 10 | 0.13* | 238 | -1.14 | 8 | -0.25 | 8 | -1.11 | 8 | -0.64 | 16 |
| 11 | -0.19 | 2 | -0.20 | 50 | -1.29 | 6 | 0.22* | 189 | -0.79 | 30 |
| 12 | -0.62 | 4 | -0.72 | 22 | 0.20* | 204 | -1.25 | 6 | -0.34 | 42 |
| 13 | 0.16* | 229 | -0.07 | 4 | -0.95 | 7 | -0.94 | 8 | -0.73 | 30 |
| 14 | 0.21* | 184 | -1.11 | 1 | -0.33 | 33 | -0.74 | 9 | -0.41 | 51 |
| 15 | -0.92 | 9 | -0.56 | 7 | 0.13* | 188 | -0.58 | 13 | -0.09 | 61 |
| 16 | 0.16* | 206 | -0.65 | 6 | -0.97 | 2 | -1.59 | 4 | -0.36 | 60 |
| 17 | -0.56 | 5 | 0.41* | 116 | -1.71 | 6 | -0.38 | 11 | -0.22 | 140 |
| 18 | -1.11 | 1 | -0.50 | 30 | 0.30 | 162 | -2.07 | 2 | -0.34 | 83 |
| 19 | -0.83 | 2 | 0.14* | 241 | -1.11 | 3 | -0.83 | 22 | -1.11 | 10 |
| 20 | -0.56 | 3 | -0.59 | 56 | -0.65 | 9 | 0.24* | 175 | -0.05 | 35 |
| 21 | -0.76 | 4 | 0.33* | 162 | -1.11 | 3 | 0.00 | 0 | -0.43 | 109 |
| 22 | -1.02 | 3 | -2.07 | 4 | -0.66 | 8 | 0.15* | 207 | -0.27 | 56 |
| 23 | -0.45 | 10 | -0.45 | 52 | 0.08 | 106 | 0.00 | 0 | 0.18* | 110 |
| 24 | 0.23* | 171 | -0.51 | 27 | -0.32 | 21 | -0.67 | 12 | -0.23 | 47 |
| 25 | -0.60 | 7 | -0.45 | 5 | -0.72 | 30 | 0.19* | 209 | -0.43 | 27 |

*Keyed right answer.

[1]Numbers in italics indicate number of students selecting each option.

produce a higher reliability coefficient than the NR mode appears to have been confirmed. Whether or not these differences in reliability coefficients are statistically significant is very difficult to assess, however. Lord (1975) discussed the problem of locating good tests for comparing reliability coefficients:

1. Only a lower bound to the test reliability (coefficient alpha, for example) can be estimated without splitting the test; good significance tests or confidence intervals for lower bounds are hard to find.

2. Showing that one lower bound is higher than the other does *not* show that one reliability is higher than the other.

3. If the test is split (to obtain a split-half correlation, for example), the choice of split affects the outcome; thus, fluctuations due to choice of split should be taken into account. Good significance tests that do this are not readily available (Lord, 1975, p. 10).

In the studies reported earlier, the most frequently used measure of differences between reliability coefficients has been the coefficient of effective length (CEL), defined by Coombs, *et al.* (1956) as:

$$\text{CEL} = k = \frac{r_{11}(1 - r_{22})}{r_{22}(1 - r_{11})}$$

where $k$ is a multiplier indicating the number of times the original test would be increased in length by the experimental procedure; $r_{11}$ is the reliability of the test administered by the experimental method method; and $r_{22}$ is the reliability of the test as conventionally administered (Coombs, *et al.*, 1956, p. 25). The application of this formula allows the derivation of a percentage increase which can be used to estimate the number of test items which could be removed from

the more reliable test and still maintain the reliability level of the less reliable test.

The CEL, however, is not a test of statistical significance. It can give only what Lord (1975) called evidence of a "worthwhile improvement" (p. 9). There are some methods for setting bounds on reliability coefficients which will be discussed below, but they require the division of the test into equivalent halves. If, to avoid the effects of bias in the choice of split in a test, a Kuder-Richardson formula is used to estimate reliability, the possibilities for tests of statistical significance are very few. A complicating factor in this study is that many of the desired comparisons are between dependent observations. Feldt (1969) proposed a test for differences between the Kuder-Richardson formula 20 (KR20) or Cronbach's (1947) generalization of KR20 (coefficient alpha). In the development of this test, Feldt demonstrated that the distribution of statistic $W$ approximated the $F$-distribution with adjusted degrees of freedom. He defined the statistic as follows:

$$W = \frac{1 - r_1}{1 - r_2} \quad \text{at } \nu_1, \nu_2$$

where $r_1$ = KR20 or alpha coefficient for test 1; $r_2$ = KR20 or alpha coefficient for test 2; $\nu_1 = (2A^2)/(2B-AB-A^2)$; and $\nu_2 = (2A)/(A-1)$. $A$ and $B$ are determined by calculating four $df$ values:

$$df_1 = N_1 - 1$$

$$df_2 = (N_1 - 1)(k_1 - 1)$$

$$df_3 = (N_2-1)(k_2-1)$$

$$df_4 = N_2-1$$

where $N$ = number of examinees taking test and $k$ = number of items on test. These values are then used to find $A$ and $B$:

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

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

This test, however, assumes independent samples. It was, therefore, used only to test for differences between pairs of groups on reliability coefficients under each of the scoring rules. No differences, significant at the .05 level, were found. The reliabilities for all groups, under the same response mode and scoring rule, were approximately the same.

The other methods used in this study to examine reliability coefficients were based on split-half estimates. In an effort to satisfy the assumption for split-half estimates of equivalent half-tests, the items were assigned to the half-tests on the basis of their type and their item variance. The 25 items were divided into punctuation items, spelling items, capitalization items, grammar items, and items with no errors. The variance for each item within the total group of 278 examinees was determined, and the individual variances within the items types were paired according to size. It was generally possible

to make close matches for the item variances, so that the first punctuation item assigned to half-test 1 was very similar in variance to the first punctuation item assigned to half-test 2. The same was true for the other item types. After the pairs of items were assigned, the odd items of each type were paired on their item variances alone and also assigned to the half-tests. Because there were 25 items in the total test, one half-test had an additional item.

To assess how successful the splitting procedure had been in producing equivalent half-tests, a test of the significance of the differences between correlated variances (Ferguson, 1971, p. 167) was performed. The results of this test for scores determined by the NR, CMBS, and EMP procedures are given in Table 10. The mean for Half1 was, not unexpectedly, higher than the mean for Half2, since Half1 contained the 13 items while Half2 contained only 12. This difference in test length raised the possibility that the scores from Half1 would show greater dispersion than scores on Half2, but it was assumed that the effects of adding only one item would be minimal. All other thing things being equal, an increase of 8.33 percent in length for a 12-item test with original reliability of .67 would produce a reliability coefficient of .69. A coefficient of .82 would be improved to .83.

The results presented in Table 10 indicate that the partitioning of the test into two halves was most successful for groups 1 and 3. The variances of the half tests were significantly different for group 2 examinees on NR and EMP scores. Group 2 included the examinee with

Table 10. Comparisons of Half-Test Means and Variances for All Three Groups of Examinees.

| Score | Mean | Variance | Covariance | Correlation | *df* | *t* |
|---|---|---|---|---|---|---|
| Group 1 | | | | | | |
| NR: Half1 | 8.90 | 5.12 | 2.40 | .48 | 67 | 0.235 |
| Half2 | 7.20 | 4.87 | | | | |
| CMBS: Half1 | 7.61 | 9.79 | 4.96 | .56 | 67 | 0.999 |
| Half2 | 5.87 | 8.00 | | | | |
| EMP: Half1[1] | -0.35 | 4.65 | 2.18 | .55 | 67 | 1.579 |
| Half2 | -0.41 | 3.38 | | | | |
| Group 2 | | | | | | |
| NR: Half1 | 9.26 | 5.44 | 2.62 | .63 | 84 | 3.279[2] |
| Half2 | 7.57 | 3.14 | | | | |
| CMBS: Half1 | 8.47 | 8.60 | 5.29 | .68 | 84 | 1.256 |
| Half2 | 6.44 | 7.04 | | | | |
| EMP: Half1 | -0.18 | 5.34 | 2.74 | .69 | 84 | 3.874[2] |
| Half2 | -0.08 | 2.95 | | | | |
| Group 3 | | | | | | |
| NR: Half1 | 9.66 | 3.64 | 1.98 | .55 | 121 | 0.010 |
| Half2 | 7.99 | 3.63 | | | | |
| CMBS: Half1 | 9.00 | 4.99 | 2.62 | .50 | 121 | -0.064 |
| Half2 | 7.10 | 5.47 | | | | |
| EMP: Half1 | 0.32 | 3.36 | 1.86 | .62 | 121 | 1.660 |
| Half2 | 0.29 | 2.65 | | | | |

[1]EMP computations on *z*-scores
[2]Significant at $\alpha < .05$

the most extreme score. This examinee had a NR score of only 3.00 and received $z$-scores on the half tests of -10.10 and -8.55. The removal of this examinee from the computations for group 2 produced slightly lower $t$-values for NR and EMP scores and a slightly higher value for CMBS scores, but the NR and EMP variances continued to show a significant difference. The difference for NR and EMP and not for CMBS was attributed to the fact that NR and EMP scores were determined from the same response mode and that positive choice weights were assigned for EMP scores based on NR scores. Although total NR, CMBS, and EMP scores all showed intercorrelations of about .95 in group 2, the choice of split, apparently, emphasized differences in the responses to items on the two tests which were masked in the total scores. The CMBS, derived from a different response mode, appeared to level out these differences in variance and give the appearance of equivalent forms.

The reliability coefficients used in determining confidence intervals and the confidence intervals calculated by the procedure described below are given in Table 11. Coefficient alpha, previously reported, is shown for comparison purposes. The split-half estimate given in Table 11 is the correlation coefficient between the half-tests, shown in Table 10, increased to full length by the Spearman-Brown correction, which is:

$$r_{xx'} = \frac{2r_{12}}{1 + r_{12}}$$

Table 11. Reliability Estimates for NR, CMBS, and EMP for All Three Groups Determined by Coefficient Alpha, the Split-Half Method Corrected by the Spearman-Brown Formula, and a Split-Half Estimate, $\hat{\alpha}$, Described by Kristoff (1972).

| Test | Alpha | CEL | Split-Half | CEL | $\hat{\alpha}$ | 95% C. I. for $\hat{\alpha}$ |
|---|---|---|---|---|---|---|
| Group 1 $n$=69 | | | | | | |
| NR | .67 | | .65 | | .66 | .45 to .79 |
| CMBS | .76 | 1.60 | .72 | 1.32 | .72 | .55 to .82 |
| EMP | .79 | 1.85 | .71 | 1.32 | .70 | .52 to .81 |
| Group 2 $n$=86 | | | | | | |
| NR | .66 | | .77 | | .76 | .64 to .84 |
| CMBS | .75 | 1.55 | .81 | 1.27 | .81 | .71 to .88 |
| EMP | .82 | 2.35 | .82 | 1.36 | .80 | .72 to .88 |
| Group 3 $n$=123 | | | | | | |
| NR | .60 | | .71 | | .71 | .59 to .80 |
| CMBS | .63 | 1.14 | .67 | 0.83 | .67 | .53 to .77 |
| EMP | .78 | 2.36 | .75 | 1.23 | .76 | .66 to .83 |

where $r_{xx'}$ is the reliability estimate for the full length test, and $r_{12}$ is the correlation of the two half-tests.

The third reliability estimate in Table 11, $\hat{\alpha}$, is also a split-half estimate, defined by Kristoff (1972) as:

$$\hat{\alpha} = \frac{4s_{12}}{s_{11} + s_{22} + 2s_{12}}$$

where $s_{11}$ is the variance of half-test 1, $s_{22}$ is the variance of half-test 2, and $s_{12}$ is the covariance of the two half-tests. This formula is mathematically similar to the corrected half-test correlation and gives almost identical results.

Although coefficient alpha and the split-half procedure are, theoretically, estimates of the same quantity, there appear to be differences in the observations for these groups. Some of the difference can probably be attributed to bias introduced in the choice of test split. The EMP procedure used in this study is designed to maximize coefficient alpha. For some groups and some test splits, the advantages of the EMP procedure may be lost. If a low-scoring examinee should happen to have all of his right answers on one half-test and only wrong answers on the other half-test, he would constitute an extreme case that would distort the correlation coefficient on which the split-half estimate is based, particularly in a relatively small sample. Cronbach (1951) demonstrated that coefficient alpha is the mean of all possible split-half reliability estimates. In that context, the different estimates in Table 11 may be viewed as over- and

under-estimates of the mean of split-half coefficients.

Kristoff (1972) discussed both the condition when the half-tests can be assumed to be parallel and condition when they are not necessarily regarded as parallel. The formula presented for testing point hypotheses and determining confidence intervals when the parts are assumed to be parallel is:

$$t = \frac{\hat{\alpha}-\rho_t}{2\sqrt{(1-\hat{\alpha})(1-\rho_t)}}\sqrt{N-1}\ ,\ df = N-1$$

This formula was used to provide the 95 percent confidence intervals shown in Table 11 for the reliability coefficients in groups 1 and 3. When the parallelism of the half-tests cannot be assumed, the following formula was given:

$$t = \frac{\hat{\alpha}-\alpha}{\hat{\alpha}\sqrt{1-\alpha}} \cdot \frac{r}{\sqrt{1-r^2}}\sqrt{N-2}\ ,\ df = N-2$$

This expression was used for confidence intervals in group 2.

It is evident, from the confidence intervals in Table 11, that the alpha coefficients and the split-half estimates within each group for the same test cannot be considered as different, although the NR alpha coefficients in groups 2 and 3 lie on the extreme lower edge of the interval. It is possible to test point hypothesis with the above formulas, substituting the comparison value for $\rho_t$ or $\alpha$, but the spread of the confidence intervals indicates that such tests would not be significant. Obtaining more precise confidence intervals would require far larger sample sizes than are available in this study. To

demonstrate a significant difference between the extreme reliability coefficients in group 1 (.66 for NR and .72 for CMBS) at the .05 level, a sample size in excess of 400 cases would be needed.

The width of the obtained confidence intervals is a function of both the magnitude of the reliability estimates and the size of the samples. The invervals for groups 2 and 3 are somewhat smaller than those for group 1. Based on the Kristoff (1972) procedure, however, the question of whether the CMBS and EMP procedures produced statistically significant increases in reliability estimates over the NR procedure must be answered in the negative. The confidence intervals for NR scores in each group contain all of the values, alpha, split-half, and $\hat{\alpha}$, that were found for the alternative scoring and response methods.

## Tests for Differences in Validity Coefficients Under Different Response Modes

Since validity coefficients are zero-order Pearson product-moment correlations, a test for significant differences between correlated correlation coefficients, developed by Hotelling (1940), was used for this part of the analysis. The Hotelling formula is:

$$t_{d_r} = (r_{12}-r_{13})\left[\frac{(N-3)(1+r_{23})}{2(1-r_{23}^2-r_{12}^2-r_{13}^2+2r_{23}r_{12}r_{13})}\right]^{1/2}$$

(from Guilford, 1965, p. 190)

where $X_1$ is the criterion test, $X_2$ is one test being related to the criterion, and $X_3$ is the other test, given to the same group of

examinees and also being related to the criterion. Critical values for the test were determined at $p = 2$ and $df = N - 3$ from tables by Jensen and Howe (1968).

Four validity criteria were used. The first was the open-ended number right scores (ONR) based on the corrections made by examinees in their test booklets. The second was the NR score on the remaining 35 items in Part I of the test (RTEST). These items were identical in type and response options to the 25 items in the experimental test. The third criterion was composed of NR scores on the 10 items in Part II of the test (EXPS). In these items, the examinees were asked to select the best of four sentences expressing an idea. The final criterion was the score obtained from the 20 items in Part III of the test (COH). In this part of the test, examinees were directed to arrange four sets of five items into coherent paragraphs.

The obtained validity coefficients and $t$-tests for significant differences between the NR coefficients and CMBS and EMP coefficients are given in Table 12. Several significant declines in validity were found when the alternative response modes and scoring rules were used. In group 1, the use of the CMBS procedure compared to NR scores produced significant decreases in validity with ONR scores and EXPS scores. In group 2, the validity coefficients for both CMBS and EMP scores with ONR scores were significantly lower than the validity coefficient for ONR with NR scores. EMP scores also showed significantly lower coefficients with RTEST and EXPS compared to the coefficient for NR with RTEST and EXPS. For group 3, only EMP with COH

Table 12. Comparisons of Validity Coefficients for NR, CMBS, and EMP Scores by Group.

| Score | Validity Coefficients | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ONR | $t$-tests[1] | RTEST | $t$-tests | COH | $t$-tests | EXPS | $t$-tests |
| Group 1 | | | | | | | | |
| NR | .87 | | .71 | | .35 | | .40 | |
| CMBS | .79 | 2.61[2] | .64 | 1.59 | .27 | 1.36 | .25 | 2.67[2] |
| EMP | .83 | 2.33 | .73 | -0.84 | .32 | 0.92 | .43 | -0.96 |
| Group 2 | | | | | | | | |
| NR | .89 | | .73 | | .26 | | .50 | |
| CMBS | .84 | 3.16[3] | .72 | 0.42 | .31 | -1.53 | .50 | 0.00 |
| EMP | .84 | 3.55[3] | .67 | 2.87[3] | .20 | 2.04 | .41 | 3.50[3] |
| Group 3 | | | | | | | | |
| NR | .87 | | .60 | | .27 | | .42 | |
| CMBS | .84 | 1.55 | .62 | -0.63 | .25 | 0.51 | .45 | -0.82 |
| EMP | .84 | 2.12 | .61 | -0.44 | .20 | 2.56[4] | .41 | 0.38 |

[1] $t$-tests between NR and CMBS, and NR and EMP
[2] significant at .05 [C.V. (.05,2,66) = 2.52]
[3] significant at .05 [C.V. (.05,2,83) = 2.51]
[4] significant at .05 [C.V. (.05,2,120) = 2.49]

showed a significant decline in validity as compared to NR.

## Effects of Partial Information and Misinformation on Reliability and Validity

The second major concern of this study was on the differential effects of the various partial information and misinformation score components on reliability and validity coefficients. These effects were investigated by sequentially removing from test scores the gain or loss in score from partial information or misinformation.

Since the standard to which alternative response modes and scoring rules are compared to determine whether their use had been beneficial or detrimental to test reliability and validity is NR scores, NR scores were used as the reference point for the removal of the effects of partial information on scores. Items which were counted as right under NR scoring, but which had been identified under the CMBS mode as items on which an examinee had guessed, were sequentially rescored to add no credit to the total score. The procedure followed was identical to that used by Cross (1973) to examine the effects of guessing on multiple-choice test scores. Four sets of scores, each representing a different degree of partial information, were computed. For the first sets of scores, each item for each examinee was scored 1 if at least one distractor was correctly marked and the right answer was indicated. The item was scored 0 otherwise. Thus, the score credit from guessing successfully among five alternatives was removed from the NR scores. This score is referred to as RG5.

The second set of scores was determined by giving one point if at

least two distractors were marked and the right answer indicated, and 0 otherwise. This rescoring removes the effects of the guessing with no partial information (G5) which was removed in the first set of scores, plus the effects of guessing with a small amount of partial information (G4). These scores are referred to as RG5/4.

For the third set of scores, the tests were again rescored to give one point if at least three distractors were marked and the right answer indicated, and 0 otherwise. These scores are referred to as RG5/4/3, and show the effects of the removal of moderate levels of partial information.

The final set of scores was determined by awarding one point if all four distractors were marked and the right answer indicated, and 0 otherwise. These scores reflect only those items for which it appeared that the examinees knew the answer with a substantial degree of assurance. Thus, these scores are an indication of complete information with no apparent guessing involved. These scores are referred to as GF scores (RG5/4/3/2) and are identical to the INFO scores shown previously in Tables 6 and 7.

These four sets of scores in comparison to NR scores are shown in Table 13. As noted before (see Table 6), group 1 was significantly different from the other two groups in the frequency of reported guessing among five alternatives. This effect is evidenced again by the greater decrease in mean scores for group 1 when the gains to NR scores from apparently random guessing are removed. The progressive changes in the mean scores in Table 13 reflect the frequency of use

Table 13. Reliability Changes for NR Scores from Sequentially Removing Guessing Score Components.

| Score | Mean | Standard Deviation | Alpha Reliability | CEL | 95 percent C.I. for NR scores (from α) |
|---|---|---|---|---|---|
| Group 1 $n$=69 | | | | | |
| NR | 16.10 | 3.85 | .67 | | |
| RG5 | 15.20 | 4.43 | .75 | 1.48 | |
| RG5/4 | 15.06 | 4.68 | .78 | 1.75 | .45 to .79 |
| RG5/4/3 | 14.75 | 4.91 | .80 | 1.97 | |
| GF (RG5/4/3/2) | 12.73 | 5.05 | .81 | 2.10 | |
| Group 2 $n$=86 | | | | | |
| NR | 16.83 | 3.72 | .66 | | |
| RG5 | 16.66 | 3.87 | .69 | 1.15 | |
| RG5/4 | 16.55 | 3.99 | .71 | 1.26 | .64 to .84 |
| RG5/4/3 | 16.16 | 4.19 | .73 | 1.39 | |
| GF (RG5/4/3/2) | 14.06 | 5.09 | .82 | 2.35 | |
| Group 3 $n$=123 | | | | | |
| NR | 17.65 | 3.35 | .60 | | |
| RG5 | 17.59 | 3.37 | .61 | 1.03 | |
| RG5/4 | 17.50 | 3.40 | .61 | 1.03 | .59 to .80 |
| RG5/4/3 | 16.71 | 3.94 | .71 | 1.63 | |
| GF (RG5/4/3/2) | 13.13 | 4.80 | .79 | 2.51 | |

and the success associated with the guessing categories. The increasing dispersion of the scores when guessing is removed is shown by the progressively larger standard deviations.

It is also evident, from the data in Table 13, that reliability increased when guessing was removed. Frary (1968, 1969) has presented a theoretical framework in which this analysis can be viewed. He indicated that enhancement of reliability by removing guessing scores could be expected when the correlation between true scores and guessing scores was negative. Such a situation would occur when the better prepared examinees, who inevitably have higher true scores, also have lower guessing scores, while less well-prepared examinees tend to have lower true scores and higher guessing scores. This follows, intuitively, since better prepared examinees will have less necessity for guessing since they know more, and less well prepared examinees will be forced to guess more frequently since they have less information. These results appear to be consistent with this explanation. With the removal of each guessing components, there was an increase in the reliability estimates.

The question of whether these changes are statistically significant remains difficult to assess. In terms of the conventionally used coefficient of effective length (CEL), the removal of guessing had a marked effect on test reliability, increasing the effective length to that of a test more than twice as long for all groups. In the context of the NR 95 percent confidence intervals for $\hat{\alpha}$, however, only the coefficients for group 1 slightly exceed the range established for

NR scores. The same coefficients for groups 2 and 3 are contained within the 95 percent confidence intervals, but were a 90 percent interval chosen, the extreme values, *i.e.*, NR and GF, would be considered significantly different.

It must be noted, however, that although partial information and guessing are highly related concepts, they are not synonymous when they are discussed in terms of both NR responses and CMBS responses. The credits which constitute NR scores are a combination of complete information, partial information, and guessing. Partial information in NR scores operates to allow an examinee to reduce the set of options from which he selects his answer so that the probability of choosing correctly is greater than it would be if the answer were chosen from the complete set of options. Whatever credit accrues to NR scores from partial information is a function of the more or less fortunate choosing of the single right answer.

With the CMBS response mode, the scoring emphasis is essentially the reverse. All of the score points do not lie in single right answers for items, but are spread throughout the item options. Each five-option item contains four equal portions of credit and a penalty. Credits accrue to CMBS scores from any expression of information, either partial or complete, and, theoretically, there is no guessing in CMBS scores. Thus, the removal of the guessing components from NR scores reflects the effects of partial information on NR scores, but not in the same way in which partial information is involved in CMBS scores.

In the context of Frary's (1968, 1969) theoretical model of the effects of guessing scores on test reliability, the implication is that the removal of a score component which is negatively related to true score will improve reliability. He suggested, further, that the removal of a score component having zero correlation with true score might also improve test reliability, but that a score component having a positive correlation with true score would have little or no effect on reliability. It does not seem likely, with the CMBS response mode and scoring rule, that the expression of partial information would be negatively related to expressions of complete information. Examinees with higher levels of complete information would also be expected to have higher levels of partial information. Thus, the removal of the score components reflecting partial information should have minimal effect on the reliability of scores based on Coombs responses.

To assess the accuracy of these expectations, the CMBS test score scores were also subjected to a rescoring procedure, similar to that followed in Table 13 for NR scores. The results of this procedure are given in Table 14. The top line of the table shows the average CMBS test scores, for purposes of comparison. Four new sets of scores were obtained and are shown below the CMBS test score. These scores differ from those in Table 13 in that guessing is not a factor. While, under NR scoring, examinees receive credit for partial information only if it enables them to guess correctly, indicating that the right answer lies within a subset of the total number of options under the CMBS mode is sufficient to gain some amount of positive credit for an item.

Table 14. Reliability Changes from Sequentially Removing Partial Information Components from CMBS Scores.

| Score | Mean | Standard Deviation | Alpha Reliability |
|---|---|---|---|
| Group 1 $n$=69 | | | |
| CMBS scores | 13.48 | 5.26 | .76 |
| RMIS | 15.31 | 4.65 | .81 |
| RMIS/P4 | 15.26 | 4.75 | .81 |
| RMIS/P4/P3 | 15.00 | 4.92 | .82 |
| GF (RMIS/P4/P3/P2) | 12.73 | 5.05 | .81 |
| Group 2 $n$=86 | | | |
| CMBS scores | 14.91 | 5.12 | .75 |
| RMIS | 16.89 | 3.87 | .74 |
| RMIS/P4 | 16.83 | 3.97 | .75 |
| RMIS/P4/P3 | 16.47 | 4.22 | .77 |
| GF (RMIS/P4/P3/P2) | 14.06 | 5.09 | .82 |
| Group 3 $n$=123 | | | |
| CMBS scores | 16.10 | 3.96 | .63 |
| RMIS | 17.76 | 3.06 | .66 |
| RMIS/P4 | 17.72 | 3.09 | .66 |
| RMIS/P4/P3 | 16.98 | 3.71 | .73 |
| GF (RMIS/P4/P3/P2) | 13.14 | 4.79 | .79 |

To obtain the first set of new scores shown in Table 14, all items in which the right answer was marked as a distractor, regardless of how many distractors were correctly identified, were rescored as 0. All of these items would carry negative item scores. The remaining items were scored with appropriate CMBS credits. This first manipulation demonstrates the effect of the removal of all misinformation from CMBS scores, leaving only partial information and complete information. This set of scores is called RMIS.

The second set of scores shows the results of rescoring all misinformation items as 0 and also all items on which only one distractor was correctly identified. This set of scores is referred to as RMIS/P4. It can be seen in the table that, while the removal of misinformation produced an increase in mean test scores, the removal of the .25 credit for correctly identifying one distractor caused a slight decrease in mean scores.

For the third set of scores, all misinformation items and all items with no more than two correctly marked distractors were scored as zero. The remaining items received the appropriate CMBS credits. This set of scores is designated as RMIS/P4/P3.

The final set of scores was determined by rescoring to zero all misinformation items and all items for which no more than three distractors were correctly marked. These scores, then, reflect only those items for which all four distractors were correctly marked and are identical to the GF scores in Table 13.

Inspection of Table 14 indicates that the expected changes

occurred only in group 1. In group 1, there was a modest increase in reliability when the misinformation components were removed, but, essentially, no differences as the partial information components were removed. This suggests that, for the examinees in group 1, partial information may be positively related to complete information.

The same results were not obtained in groups 2 and 3, however, In group 2, the removal of misinformation produced a negligible change change, but there was a conspicuous improvement in reliability with the removal of all partial information. A similar effect was found in group 3, and the improvement in reliability with the removal of all partial information was even more striking than that found in group 2. These results suggest that the relationship between partial information and misinformation in groups 2 and 3 is not the same as in group 1, and that partial information may bear a negative correlation to complete information in groups 2 and 3. As in the previous reliability analyses, however, no statistically significant changes can be demonstrated.

Some significant declines in validity were revealed when the guessing components were removed from NR scores (see Table 15). The test for significance used here was the Hotelling $t$ with the same critical values as in Table 12. ONR scores proved to be particularly sensitive to the removal of guessing. This result was somewhat surprising, since it was assumed that when examinees made changes in their test booklets a minimum of guessing would be involved, and, thus, that ONR scores might be more similar to GF scores than to NR

Table 15. Validity Changes for NR Scores from Sequentially Removing Guessing Score Components

| Score | Validity Coefficients | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ONR | *t*-tests[1] | RTEST | *t*-tests | COH | *t*-tests | EXPS | *t*-tests |
| Group 1 *n*=69 | | | | | | | | |
| NR | .87 | | .71 | | .34 | | .40 | |
| RG5 | .82 | 1.73 | .58 | 3.09[2] | .30 | 0.88 | .27 | 2.40 |
| RG5/4 | .78 | 2.83[2] | .59 | 2.62[2] | .29 | 0.82 | .27 | 2.21 |
| RG5/4/3 | .76 | 3.14[2] | .57 | 2.77[2] | .26 | 1.19 | .24 | 2.47 |
| GF (RG5/4/3/2) | .69 | 3.85[2] | .47 | 3.46[2] | .15 | 2.07 | .13 | 3.06[2] |
| Group 2 *n*=86 | | | | | | | | |
| NR | .89 | | .73 | | .26 | | .50 | |
| RG5 | .88 | 1.00 | .72 | 0.67 | .30 | -1.94 | .52 | -1.07 |
| RG5/4 | .87 | 2.00 | .72 | 0.67 | .30 | -1.94 | .51 | -0.53 |
| RG5/4/3 | .88 | 0.56 | .70 | 1.07 | .27 | -0.25 | .48 | 0.56 |
| GF (RG5/4/3/2) | .72 | 4.52[2] | .47 | 4.34[2] | .18 | 0.94 | .30 | 2.64[2] |
| Group 3 *n*=123 | | | | | | | | |
| NR | .87 | | .60 | | .27 | | .42 | |
| RG5 | .87 | 0.00 | .61 | -0.98 | .26 | 0.81 | .42 | 0.00 |
| RG5/4 | .87 | 0.00 | .61 | -0.98 | .27 | 0.00 | .42 | 0.00 |
| RG5/4/3 | .82 | 2.52 | .62 | -0.63 | .26 | 0.25 | .41 | 0.27 |
| GF (RG5/4/3/2) | .61 | 6.74[2] | .49 | 1.76 | .21 | 0.78 | .26 | 2.22 |

[1] *t*-tests between NR and each of the removal-of-guessing scores.
[2] significant at .05 level

scores. Since this did not occur, it seems likely that some examinees lacked confidence in their knowledge and, thus, tended to report some degree of guessing for items on which they knew the correct answer. Removal of the partial information score components from the CMBS scores gave almost identical results to those in Table 12 for NR scores.

In assessing the effects of misinformation on test scores, the CMBS scores were used as the reference point, since it is on these scores that the effects of misinformation are most directly measurable. The differential effects of various levels of misinformation were determined by a sequential rescoring similar to that used for partial information. The results are shown in Table 16.

The first score given in Table 16 is the CMBS score, reported previously. Four additional sets of scores were obtained. The first score, referred to as RM4, was determined by rescoring as zero all items on which the right answer was the only option marked. Other items were scored as customary under the CMBS mode. The second set of scores was gained by rescoring as zero all items for which the right answer was marked as a distractor and no more than one distractor was correctly marked. These scores are referred to as RM4/3. The third set of scores was the result of rescoring to zero all items for which the right answer was marked as a distractor and no more than two distractors were correctly marked. These scores are called RM4/3/2. The final set of scores represented a misinformation free score and was determined by rescoring to zero all items on which the right answer

Table 16. Sequential Rescoring of CMBS Scores to Obtain MF Score.

| Score | Mean | S.D. | Alpha Reliability | CEL | Validity Coefficients | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | ONR | RTEST | COH | EXPS |
| Group 1 | | | | | | | | |
| CMBS | 13.48 | 5.26 | .76 | | .79 | .64 | .27 | .25 |
| RM4 | 14.82 | 4.94 | .80 | 1.26 | .75 | .57 | .23 | .20 |
| RM4/3 | 15.23 | 4.75 | .81 | 1.35 | .75 | .58 | .25 | .20 |
| RM4/3/2 | 15.29 | 4.68 | .81 | 1.35 | .75 | .59 | .26 | .21 |
| MF | 15.31 | 4.65 | .81 | 1.35 | .75 | .58 | .26 | .21 |
| Group 2 | | | | | | | | |
| CMBS | 14.91 | 5.12 | .75 | | .84 | .72 | .31 | .50 |
| RM4 | 16.19 | 4.96 | .78 | 1.12 | .83 | .68 | .29 | .47 |
| RM4/3 | 16.63 | 4.56 | .80 | 1.33 | .82 | .69 | .30 | .47 |
| RM4/3/2 | 16.81 | 4.14 | .77 | 1.12 | .84 | .72 | .32 | .50 |
| MF | 16.89 | 3.87 | .74 | 0.95 | .86 | .73 | .33 | .52 |
| Group 3 | | | | | | | | |
| CMBS | 16.10 | 3.96 | .63 | | .84 | .62 | .25 | .38 |
| RM4 | 16.96 | 3.64 | .66 | 1.14 | .82 | .61 | .25 | .43 |
| RM4/3 | 17.52 | 3.30 | .66 | 1.14 | .83 | .63 | .25 | .45 |
| RM4/3/2 | 17.75 | 3.08 | .66 | 1.14 | .83 | .63 | .25 | .44 |
| MF | 17.76 | 3.06 | .66 | 1.14 | .84 | .63 | .26 | .44 |

was marked as a distractor.

The MF score differs from the NR score to the extent that the CMBS credit for partial information differed from the incidence of successful guessing in the NR mode. For example, an examinee who indicated guessing on items for which he knew the correct answer would receive a lower MF score than NR score, since he would be claiming less than the maximum CMBS credit for these items. On the other hand, if he appropriately indicated guessing on items about which he was not sure and on which he guessed incorrectly for his NR score, his MF score could exceed his NR score.

It is evident, from examination of the data in Table 16, that the effect of the misinformation categories on test reliability and validity was minimal. A similar analysis was performed on ENR scores to examine whether differences could be found when NR scores were used as the basis for comparison. The results were almost identical to those reported in Table 16 for the CMBS scores. No trends across groups could be observed and, since the changes were so small when the misinformation components were removed, no tests for significant differences were made.

## Summary of Reliability and Validity Effects

Although both CMBS and EMP scoring produced reliability coefficients of greater magnitudes than NR scoring, the differences could not be shown to be statistically significant. There were, however, several significant declines in the correlations with validity

criteria when CMBS and EMP were used. The validity differences were most pronounced in group 2.

The removal of the effects of guessing from NR scores suggested that the guessing component depresses reliability. The increases in reliability as guessing components were rescored exceeded the 95 percent confidence intervals only for group 1. The changes in groups 2 and 3 produced a considerable increase in effective test length, but could not be shown to be statistically significant. The removal of partial information from CMBS scores produced no change in reliability in group 1, but increases were shown in groups 2 and 3. The removal of guessing from NR scores and the removal of partial information from CMBS scores also tended to decrease criterion-related validity. For NR scores in all three groups, the decrease was significant in relation to ONR. For groups 1 and 2, the declines with RTEST and EXPS were also significant.

The removal of misinformation score components from CMBS scores and also from ENR scores had negligible effects on reliability and validity in all three groups.

Personality Characteristics and Test Scores

Because of the large number of scales produced by the Adjective Check List and the considerable intercorrelation among some of the scales (Gough & Heilbrun, 1965, p. 29), the number of variables was reduced by a principal components factor analysis, using subprogram Factor from SPSS (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). Using a minimum eigen-value of 1.00 as the criterion for selecting the number of factors, a 5-factor structure, which accounted for almost 77 percent of the variance among the scales, was determined and rotated to final solution by the Varimax procedure. The factor loadings from this solution are presented in Table 17. Factor scores for each examinee on the five factors were output for further analyses.

To facilitate further discussion of the factors, a subjective interpretation of the patterns of loadings was made, based on the descriptions given in the Adjective Check List manual (Gough & Heilbrun, 1965), and the factors named. The scales loading most heavily on Factor I seemed to reflect a self-perception of dependable, self-controlled, serious, and conventionally responsible behavior. This factor was named Conformist.

Factor 2 appeared to describe a more forceful, dominant, confident individual who was determined to do well and sure of his ability. Factor 2 was named Assertive. Factor 3 suggested a mercurial individual who was impulsive and intolerant of prolonged effort and tedium. Factor 3 was named Spontaneous.

Only two scales loaded clearly on Factor 4; however, the factor

Table 17. Factor Loadings on Rotated 5-Factor Solution for 24 Scales from Adjective Check List.

| Scale | Factor 1[1] | Factor 2[2] | Factor 3[3] | Factor 4[4] | Factor 5[5] |
|---|---|---|---|---|---|
| No. Adjectives Checked | .07 | .03 | .07 | -.02 | .88 |
| No. Favorable Adjectives | .86 | .23 | .09 | .16 | .19 |
| No. Unfavorable Adjectives | -.76 | .02 | -.11 | -.16 | .32 |
| Defensiveness | .82 | .11 | -.19 | .29 | .09 |
| Self-Confidence | .12 | .85 | .11 | .09 | .07 |
| Self-Control | .79 | -.26 | -.31 | -.00 | -.10 |
| Lability | .06 | .16 | .74 | -.03 | -.06 |
| Personal Adjustment | .84 | -.01 | -.02 | .15 | .00 |
| Need for Achievement | .54 | .70 | -.13 | -.06 | .10 |
| Need for Dominance | .26 | .91 | -.01 | .07 | .04 |
| Need for Endurance | .69 | .31 | -.51 | -.06 | -.10 |
| Need for Order | .66 | .17 | -.58 | -.21 | .06 |
| Need for Intraception | .82 | -.09 | .11 | -.09 | .10 |
| Need for Nurturance | .78 | -.18 | .07 | .44 | -.04 |
| Need for Affiliation | .67 | .11 | .12 | .49 | .18 |
| Need for Heterosexuality | .19 | .25 | .13 | .69 | .19 |
| Need for Exhibition | -.31 | .72 | .09 | .41 | .12 |
| Need for Autonomy | -.42 | .72 | .29 | -.15 | .06 |
| Need for Aggression | -.74 | .54 | .02 | -.10 | .08 |
| Need for Change | .02 | .32 | .73 | .07 | .17 |
| Need for Succorance | -.42 | -.58 | -.29 | .01 | .32 |
| Need for Abasement | .13 | -.85 | -.07 | -.11 | .10 |
| Need for Deference | .46 | -.74 | -.31 | .09 | -.05 |
| Counseling Readiness | -.19 | .04 | .07 | -.74 | .25 |
| Variance Explained by Factor | 7.64 | 5.48 | 2.17 | 1.92 | 1.28 |
| Percent of Total Variance | 31.84 | 22.83 | 9.03 | 7.98 | 5.31 |

[1]Named the Conformist Factor
[2]Named the Assertive Factor
[3]Named the Spontaneous Factor
[4]Named the Sociable Factor
[5]Named the Enthusiastic Factor

accounted for almost eight percent of the total variance. These scales indicated an uncomplicated, outgoing, pleasure-seeking outlook. Factor 4 was named Sociable.

The single scale loading unambiguously on Factor 5 reflects surgency, drive, and a relative absence of repressive tendencies. A slightly negative correlation with intelligence was reported by Gough and Heilbrun (1965), and the conjecture was made that checking many adjectives as applying to oneself may indicate an "exhuberance in behavior" that springs "more from shallowness and inattention to ambiguities than from a deep level of involvement" (p. 7). Factor 5 was summarized as Enthusiastic.

Although the factor scores were derived from the responses of all the examinees for whom Adjective Check List scores were available (273 examinees), it could not be assumed that all three groups would exhibit equivalent overall levels of each factor, since there was no random assignment of examinees to groups. Groups 1 and 2 were composed of a mixture of freshmen, transfer students who could be at any academic level, and senior student teachers. These groups were expected to show highly similar scores on the personality factors because of the similarity of the academic levels of the examinees within them. Group 3 was made up almost entirely of sophomores, and, to the extent that the relative homogeneity of the group related to the measured personality characteristics, group 3 could be found to differ from groups 1 and 2. Group 3 was also the largest of the three groups and, since the factor scores were a function of the combined groups, group 3

would have the greatest effect on the factor scores, and thus emphasize whatever differences might be present.

On the other hand, it was hoped that sources of similarity, such as the selection of the teacher training program as a major and the choice of this specific university, would act to reduce the differences. All the examinees, except for one small group of student teachers included in group 2, were given the Adjective Check List before the Missouri College English Test. For the freshmen and sophomores, the administration of these two instruments occurred in two successive testing sessions, 24 hours apart. This separation of the two instruments was assumed to eliminate the possibility of influence on the Adjective Check List responses from the additional stress that was expected to be introduced with the Missouri College English Test.

To assess whether the expectation of similarity among the three groups was attained, a multivariate analysis of variance (MANOVA) was performed on the seven personality and attitude measures with group membership as the independent variable (SAS, 1979). The resulting Wilks' criterion was .87950, at $p = 7$, $\nu_H = 2$, $\nu_E = 252$. The critical value corresponding to these parameters at the .05 level was .91263. Since the calculated value did not exceed the critical value, the MANOVA was significant at $p < .05$. Thus, the assumption of initial equivalence of groups was found to be incorrect.

Simultaneous confidence intervals on the extreme pairs of means for each of the characteristics did not reveal which of the variables was contributing to the significant difference. Simultaneous

confidence intervals provide considerable protection against reporting as significant, differences which are, in fact, spurious, but, since such a conservative approach can be viewed as biased in favor of the desired outcome, another procedure (Hummel & Sligo, 1971) which is less likely to overlook real differences between groups, was followed. Hummel and Sligo (1971) argued that a completely multivariate approach, using MANOVA and simultaneous confidence intervals, was extremely conservative; they suggested that a combination approach, using MANOVA with univariate *F*-tests, provided reasonable and consistent control over experimentwise error rates, but was not so extremely conservative. The results of univariate *F*-tests for the variables in the MANOVA indicated that group 1 was significantly different from groups 2 and 3 on the Enthusiastic factor and on the dogmatism scale. No significant differences between groups 2 and 3 were found.

The differences between group 1 and the other groups may possibly be attributable to the time in the school year when the testing occurred. The instruments were administered to group 1 during the fall quarter, while groups 2 and 3 were tested in the winter and spring. Group 1 was primarily composed of freshmen and transfer students in their first quarter at the university. Group 2 was also primarily freshmen and transfer students, but at the time of the testing, these freshmen and transfer students were in the winter quarter and, thus, had passed through one quarter of adjustment to the university environment. The univariate *F*-tests revealed that group 1 was significantly less enthusiastic ($\overline{X}$ = -0.25, S.D. = 0.85) than either group 2

($\overline{X}$ = 0.16, S.D. = 1.08) or group 3 ($\overline{X}$ = 0.06, S.D. = 0.98), and more dogmatic ($\overline{X}$ = 93.95, S.D. = 17.60) than group 2 ($\overline{X}$ = 86.45, S.D. = 17.49) or group 3 ($\overline{X}$ = 89.08, S.D. = 16.40). These indications of greater feelings of repression and of less openness to change in group 1 seem consonant with what might be the expected response of new students in a complex, unfamiliar university environment. The lack of difference between groups 2 and 3 suggests that, after the first quarter in the new situation, the groups may become more similar. It should be noted, however, that the freshmen and transfer students in groups 1 and 2 were, essentially, self-selected to the groups. Those in group 1 were the students who responded to the first request to participate in the testing. The students in group 2 were those who could not or did not respond to the first announcement. It is, thus, possible that the differences between group 1 and the other groups are due to factors other than the length of time at the university and may persist. No means for examining this possibility was included in the study, and, therefore, the remaining data analyses must be viewed in the context of the initial differences among the groups.

The first question addressed by the regression analysis was whether differing amounts of variance in test scores would be explained by the personality and attitude variables under different penalty conditions. The $R^2$ and signficance for the regressions by group on NR, CMBS, ICW, and EMP scores are shown in Table 18 under the heading *Experimental Scores*. These results indicate that there was no significant involvement of personality variables in the no-penalty

Table 18. Multiple Linear Regressions of Personality Factor Scores and Attitude Scales on Test Scores.

| Dependent Variable | Multiple R | $R^2$ | *F*-statistic for Total Regression |
|---|---|---|---|
| Group 1 *n*=62 | | | |
| *Experimental Scores* | | | |
| NR | .15 | .02 | 0.18 |
| CMBS | .27 | .07 | 0.62 |
| ICW | .18 | .03 | 0.25 |
| EMP | .08 | .01 | 0.05 |
| *Criterion Scores* | | | |
| RTEST | .22 | .05 | 0.40 |
| EXPS | .26 | .07 | 0.54 |
| COH | .32 | .10 | 0.88 |
| Group 2 *n*=77 | | | |
| *Experimental Scores* | | | |
| NR | .56 | .32 | 4.60[1] |
| CMBS | .56 | .32 | 4.59[1] |
| ICW | .55 | .30 | 4.21[1] |
| EMP | .55 | .30 | 4.21[1] |
| *Criterion Scores* | | | |
| RTEST | .54 | .29 | 4.01[1] |
| EXPS | .54 | .29 | 4.11[1] |
| COH | .29 | .08 | 0.89 |
| Group 3 *n*=116 | | | |
| *Experimental Scores* | | | |
| NR | .36 | .13 | 2.33[1] |
| CMBS | .37 | .14 | 2.43[1] |
| ICW | .36 | .13 | 2.24[1] |
| EMP | .39 | .15 | 2.77[1] |

Table 18 (continued). Multiple Linear Regressions of Personality Factor Scores and Attitude Scales on Test Scores.

| Dependent Variable | Multiple R | $R^2$ | *F*-statistic for Total Regression |
|---|---|---|---|
| **Group** 3 (continued) | | | |
| *Criterion Scores* | | | |
| RTEST | .41 | .16 | 3.05[1] |
| EXPS | .26 | .07 | 1.07 |
| COH | .31 | .10 | 1.70 |

[1]significant at $p \leq .05$

condition (group 1), and significant involvement in all the experimental scores for groups 2 and 3. Thus, as far as can be determined from these data, there was a difference in the action of the personality and attitude variables under the no-penalty condition, as opposed to the two penalty conditions. The fact that initial differences in the personality and attitude variables between group 1 and the other groups could be shown tends to cloud this result, however, and make attribution of cause speculative. The expectation that there would be a greater involvement of the affective characteristics in group 3, which was assumed to be operating under a condition of greater risk than group 2, was not supported by the data. Between 30 and 32 percent of the variance in group 2 scores could be explained by the personality and attitude variables, while only between 13 and 15 percent of the variance in group 3 scores was accounted for by these variables.

It was also expected that, in groups 2 and 3, the NR and EMP scores would show little or no relationship to affective characteristics while the CMBS and ICW scores would be significantly related to the affective variables. This difference was expected to manifest itself in large $R^2$ values for CMBS and ICW scores, and small $R^2$ values for NR and EMP scores. That this did not occur is obvious from the results in Table 18. Approximately the same amounts of variance were explained by each of the experimental scores within the groups. Rather than a differential effect related to CMBS and ICW responses, there appears to be a diffuse effect over all test scores.

This pervasive effect is a function of the high intercorrelations among the sets of scores within the groups. There was, apparently, little change in relative positions of examinees, regardless of the response mode or scoring rule used. In addition, the test items proved to be of low to moderate difficulty for the examinees, and, thus, there was less need for guessing and less use of the various response patterns offered by the CMBS response mode than would be found in a more difficult test. When the effects of NR scores were removed from CMBS scores by partial correlation, the amount of additional variance in CMBS scores explained by the personality and attitude variables was negligible.

Under the assumption that, if the imposition of the CMBS penalty could be considered a cause of the differences in affective relationships between group 1 and the remaining groups, the removal of the penalty would produce a lessening or disappearance of the effect, a series of regressions was performed on the remaining sections of the test. These results are given under the heading *Criterion Scores* in Table 18. Inspection of these results indicates that this assumption did not prove to be correct, either. There was no dramatic drop in the $R^2$ values between the experimental scores and RTEST, which was composed of the last 35 items in Part I of the Missouri College English Test. Rather, in group 3, RTEST was found to have a larger, although not significantly greater, relationship to the personality and attitude variables than the experimental scores. The expected drop did occur for the 10 items in Part II of the test (EXPS) in group 3,

but not in group 2. There was no involvement of affective characteristics in the last section of the test (COH), but the items in this section were very different from those in the experimental test and in RTEST. The correlations for NR and CMBS with EXPS and COH (see Table 12) ranged from low to moderate, suggesting that these subsequent parts of the Missouri College English Test may measure a different ability from what is measured in the experimental test.

A final attempt to uncover differences between the groups and between the NR and CMBS test scores was made by examining the relative positions of the personality and attitude variables, based on the magnitudes of their standardized beta coefficients (see Table 19). In group 1, the largest beta-weights for both NR and CMBS scores were found for the Assertive factor, while in groups 2 and 3, the largest weights were on the Sociable factor. In addition, in group 3, the dogmatism scale produced a beta-weight significant at less than the ,05 level on the CMBS scores and at the .07 level for NR scores. In group 2, the second largest beta-weights were for the Spontaneous factor on both sets of scores, but these weights were significant only at a probability level of about .20. Thus, there is some evidence that the personality and attitude variables operated differently from one group to another, but there is little basis for attributing causality to penalty levels, since there were initial differences in the groups. There is no evidence at all supporting a different involvement of the personality and attitude variables in the CMBS scores from their involvement in NR scores.

Table 19. Ordering of Standardized Beta-Weights from Regressions of Personality and Attitude Variables on NR and CMBS Test Scores.

| NR | | CMBS | |
|---|---|---|---|
| Group 1 $n$=62 | | | |
| Assertive | -.130 | Assertive | -.217 |
| Spontaneous | -.065 | Enthusiastic | .136 |
| Dogmatism | -.045 | Spontaneous | .068 |
| Enthusiastic | .036 | Conformist | .065 |
| Conformist | .028 | Sociable | .035 |
| Rotter | .018 | Dogmatism | .027 |
| Sociable | -.012 | Rotter | .023 |
| Group 2 $n$=77 | | | |
| Sociable | -.520[1] | Sociable | -.504[1] |
| Spontaneous | .131 | Spontaneous | .129 |
| Assertive | -.040 | Assertive | -.070 |
| Dogmatism | -.038 | Dogmatism | -.066 |
| Conformist | .035 | Conformist | .042 |
| Rotter | -.029 | Enthusiastic | -.026 |
| Enthusiastic | .016 | Rotter | .008 |
| Group 3 $n$=116 | | | |
| Sociable | -.304[1] | Sociable | -.293[1] |
| Dogmatism | -.167 | Dogmatism | -.211[1] |
| Assertive | -.040 | Enthusiastic | .045 |
| Enthusiastic | .037 | Assertive | -.040 |
| Spontaneous | -.014 | Rotter | -.022 |
| Rotter | .013 | Spontaneous | .012 |
| Conformist | .004 | Conformist | .010 |

[1]Standardized beta-weights significant at $p \leq .05$.

The next step in the analysis of the data was originally intended to delineate the ways in which the information and misinformation score components contributed to the affective dimension expected to be unique to or most pronounced in the CMBS responses. Since no differences could be found in the relationships between the affective variables and CMBS and NR scores, the emphasis must be on a further examination of possible differences in response patterns and personality and attitude variables among the three groups of examinees. The results of the regressions of the personality and attitude variables on the information and misinformation score components by groups are shown in Table 20.

It is clear from Table 20 that for each of the groups the variables having the largest beta-weights for the total score regressions (Table 19) proved to be significant in explaining the INFO score component. Since the INFO pattern of marking options was the most frequently used by examinees and has the greatest effect on both NR and CMBS scores, this result was expected. The negative relationship of the Assertive and Sociable factors indicate that the less assertive in group 1 and the less sociable in groups 2 and 3 tended to have higher INFO scores than the more assertive and more sociable. This result is consonant with the literature on self-confidence and alternative response modes, since it suggests that more self-confident examinees may tend to overestimate their knowledge and mark items with greater assurance than is warranted. These examinees would tend to be less successful in the INFO category and more susceptible to the MIS1

Table 20. Multiple Linear Regressions of Personality Factor Scores and Attitude Scales on Information and Misinformation Scores.

| Dependent Variable | Multiple R | $R^2$ | *F*-statistic for Total Regression | Variables with Significant[1] Beta-weights |
|---|---|---|---|---|
| Group 1 *n*=62 | | | | |
| INFO | .43 | .18 | 1.72 | Assertive (-) |
| G2 | .21 | .04 | 0.35 | None |
| G3 | .52 | .26 | 2.79[1] | Conformist (-)<br>Enthusiastic (-) |
| G4 | .46 | .22 | 2.12 | Conformist (-)<br>Enthusiastic (-) |
| IG | .44 | .19 | 1.83 | Assertive (+) |
| MIS1 | .29 | .09 | 0.73 | None |
| MIS2 | .25 | .06 | 0.52 | None |
| MIS3 | .36 | .13 | 1.18 | Enthusiastic (-) |
| MIS4 | .17 | .03 | 0.22 | None |
| Group 2 *n*=77 | | | | |
| INFO | .44 | .20 | 2.40[1] | Sociable (-) |
| G2 | .35 | .12 | 1.35 | Enthusiastic (-) |
| G3 | .19 | .04 | 0.36 | None |
| G4 | .15 | .02 | 0.24 | None |
| IG | .30 | .09 | 1.00 | None |
| MIS1 | .46 | .21 | 2.62[1] | Sociable (+) |
| MIS2 | .32 | .10 | 1.15 | Enthusiastic (-) |
| MIS3 | .31 | .10 | 1.07 | None |
| MIS4 | .17 | .03 | 0.29 | None |
| Group 3 *n*=116 | | | | |
| INFO | .36 | .13 | 2.32[1] | Sociable (-) |
| G2 | .28 | .08 | 1.27 | None |
| G3 | .24 | .06 | 0.96 | None |
| G4 | .27 | .07 | 1.23 | None |
| IG | .32 | .10 | 1.80 | Rotter (-) |
| MIS1 | .25 | .06 | 1.00 | None |
| MIS2 | .32 | .10 | 1.71 | Sociable (+) |
| MIS3 | .31 | .10 | 1.67 | Sociable (+) |
| MIS4 | .19 | .04 | 0.61 | None |

[1]significant at $\alpha \leq .05$

penalty. In this context, MIS1 should be the converse of INFO, and the affective characteristics which produce lower INFO scores should also produce higher MIS1 scores. Confirmation of this position would be found if the factor relating negatively and significantly to INFO scores also related positively and significantly to MIS1 scores. This relationship occurred only in group 2. The Sociable factor gave a negative and significant beta-weight in the regression on INFO and a positive and significant beta-weight in the regression on MIS1. In group 1, the converse of the assertive factor was found in the IG category. Since the definition of IG was marking none of the options as being wrong, it may be that the significance of the Assertive factor in group 1 (the no-penalty group) reflects compliance with the directions, rather than a relationsiip to the alternative response mode itself. More assertive students may have decided not to cooperate with the request to give information about their response decisions, while less assertive students may have acquiesced, at least to the point of marking all of the options in most of the items, regardless of the amount of knowledge which they had about an item. Since the test was relatively easy for all examinees and the penalty for MIS1 was the least of the penalties for misinformation, marking four options for most of the items, whether or not there was complete assurance about the correct answer, would not have the strong effect on CMBS scores that it would on a more difficult test. That this might be the situation in group 1 is suggested by the lower correlation between NR and CMBS scores (.87) than in group 2 (.95) or group 3 (.90) (see Table 8).

Also, although there were no significant differences in the mean number of INFO items among the groups, group 1 had the highest mean number of MIS1 items and was significantly different from group 3 in this category (see Table 6).

In group 3, the converse of the Sociable factor was found in MIS2 and MIS3. The MIS2 and MIS3 patterns are suggestive of both uncertainty and misinformation. The examinee is not sure of the right answer and indicates guessing among two or three alternatives, but he *is* highly confident that the right answer is incorrect. Why this type of response should be given by more sociable, out-going examinees is not clear. It is in the marking of four options (INFO and MIS1) that the aspect of unwarranted confidence or risk-taking is assumed to occur, and penalties fall to the more sociable, out-going individuals. This mechanism was the expected result and is demonstrated in the regressions for group 2. The failure to find a similar result in group 3 suggests that the evocation of the affective dimension in group 3 may be different from that in group 2.

It is also possible, however, that the relative orderings of the personality and attitude variables are approximately the same in group groups 2 and 3, although the same variables did not attain significance at the .05 level in both groups. To examine this possibility and to confirm further the differences between group 1 and groups 2 and 3, the orderings of the standardized beta-weights for the information and misinformation categories are presented in Table 21. G4 and MIS4 represented only negligible portions of the scores of all

Table 21. Orderings of Standardized Beta-Weights from Regressions of Personality and Attitude Variables on Information and Misinformation Score Components.

| INFO | | MIS1 | |
|---|---|---|---|
| Group 1 $n$=62 | | | |
| Assertive | -.30[1] | Conformist | .22 |
| Enthusiastic | .20 | Enthusiastic | .15 |
| Conformist | .20 | Spontaneous | .10 |
| Dogmatism | .12 | Dogmatism | .08 |
| Spontaneous | .08 | Sociable | -.07 |
| Rotter | -.03 | Assertive | -.04 |
| Sociable | .03 | Rotter | -.01 |
| Group 2 $n$=77 | | | |
| Sociable | -.36[1] | Sociable | .38[1] |
| Enthusiastic | .18 | Enthusiastic | .22 |
| Spontaneous | .11 | Spontaneous | -.12 |
| Conformist | .06 | Conformist | .09 |
| Rotter | .05 | Rotter | .04 |
| Dogmatism | .03 | Assertive | .03 |
| Assertive | .01 | Dogmatism | .01 |
| Group 3 $n$=116 | | | |
| Sociable | -.27[1] | Spontaneous | -.14 |
| Dogmatism | -.17 | Dogmatism | .10 |
| Enthusiastic | .16 | Conformist | -.10 |
| Conformist | -.07 | Assertive | .10 |
| Spontaneous | -.04 | Rotter | .05 |
| Assertive | .04 | Sociable | .04 |
| Rotter | -.01 | Enthusiastic | .04 |

Table 21 (continued). Orderings of Standardized Beta-Weights from Regressions of Personality and Attitude Variables on Information and Misinformation Score Components.

| G2 | | MIS2 | |
|---|---|---|---|
| Group 1 $n$=62 | | | |
| Assertive | .12 | Sociable | .18 |
| Rotter | .10 | Dogmatism | .15 |
| Conformist | -.09 | Enthusiastic | .06 |
| Dogmatism | -.08 | Assertive | -.05 |
| Sociable | .07 | Spontaneous | .04 |
| Spontaneous | .02 | Conformist | -.04 |
| Enthusiastic | -.01 | Rotter | -.01 |
| Group 2 $n$=77 | | | |
| Enthusiastic | -.27[1] | Enthusiastic | -.29[1] |
| Dogmatism | -.16 | Rotter | .14 |
| Assertive | -.07 | Dogmatism | .08 |
| Rotter | -.05 | Spontaneous | .07 |
| Spontaneous | -.04 | Assertive | .07 |
| Conformist | -.03 | Conformist | -.05 |
| Sociable | .01 | Sociable | .03 |
| Group 3 $n$=116 | | | |
| Conformist | .17 | Sociable | .21[1] |
| Enthusiastic | -.16 | Spontaneous | .16 |
| Spontaneous | .10 | Conformist | .12 |
| Sociable | .08 | Dogmatism | .11 |
| Assertive | -.07 | Rotter | .10 |
| Rotter | .03 | Enthusiastic | .05 |
| Dogmatism | .03 | Assertive | .04 |

Table 21 (continued). Orderings of Standardized Beta-Weights from Regressions of Personality and Attitude Variables on Information and Misinformation Score Components.

| G3 | | MIS3 | |
|---|---|---|---|
| Group 1 $n$=62 | | | |
| Conformist | -.38[1] | Enthusiastic | -.28[1] |
| Enthusiastic | -.29[1] | Conformist | -.16 |
| Dogmatism | -.21 | Spontaneous | -.13 |
| Sociable | -.16 | Rotter | -.08 |
| Assertive | .07 | Sociable | -.07 |
| Rotter | .05 | Dogmatism | -.06 |
| Spontaneous | .01 | Assertive | -.02 |
| Group 2 $n$=77 | | | |
| Enthusiastic | -.18 | Conformist | -.21 |
| Dogmatism | .10 | Rotter | -.17 |
| Assertive | -.05 | Assertive | .11 |
| Sociable | -.04 | Sociable | .08 |
| Spontaneous | .03 | Spontaneous | -.02 |
| Rotter | -.03 | Dogmatism | .00 |
| Conformist | -.03 | Enthusiastic | .00 |
| Group 3 $n$=116 | | | |
| Sociable | .15 | Sociable | .21[1] |
| Enthusiastic | -.15 | Dogmatism | .13 |
| Dogmatism | .10 | Spontaneous | .10 |
| Assertive | -.07 | Conformist | .09 |
| Rotter | -.03 | Enthusiastic | -.08 |
| Conformist | -.02 | Rotter | -.07 |
| Spontaneous | .01 | Assertive | -.07 |

[1]beta-weights significant at $\alpha \leq .05$.

examinees and are, thus, not included in Table 21.

It is evident from the comparisons of INFO and MIS1 in Table 21 that only in group 2 can the relationships of the personality and attitude variables to MIS1 be considered the converse of the relationships to INFO. The positive, but insignificant, relationship of the Enthusiastic factor to both INFO and MIS1 in group 2 probably reflects the tendency of more confident examinees to mark more options and, thus, to increase, to some degree both INFO and MIS1 scores. It is also evident that, with the exception of the insertion of the dogmatism scale, there is some similarity between INFO in group 2 and INFO in group 3. The ordering of the MIS1 beta-weights in group 3, however, does not appear to be the same as the ordering for INFO in group 3. It should be noted that, when the beta-weights are not significantly different from zero at the selected probability level, the ordering must be viewed as no different from chance. However, the clarity of the patterns for group 2 on the INFO and MIS1 score components is in marked contrast to the more diffuse results for groups 1 and 3, suggesting that there may be a different mechanism operating in groups 1 and 3 from that in group 2.

While a proclivity towards risk-taking would be expressed in INFO and MIS1 scores, a reluctance to take risks was assumed to be most directly related to the guessing components and, probably, to the misinformation components which reflect some tentativeness of response. Thus, significant associations between the personality and attitude variables and the G2, G3, G4, MIS2, MIS3, and MIS4 score components

were expected. Contrary to expectations, a significant total regression was found only in the G3 scores for group 1 (Table 20), although there were significant beta-weights in the G4 and MIS3 components in group 1, for the G2 and MIS2 components in group 2, and for the MIS2 and MIS3 scores in group 3. It seems reasonably obvious from inspection of the comparisons of G2 and MIS2 scores and G3 and MIS3 scores that in no group can the guessing behavior be considered as the converse of the parallel marking pattern which reveals misinformation (Table 21), nor can any distinct relationships be discerned between the guessing and partial misinformation scores and the INFO and MIS1 scores. The extreme variability in the orderings of the personality and attitude variables make generalizations very difficult.

There is some evidence in the significant regression in group 1 on the G3 scores supporting the possibility that the personality and attitude involvement in group 1 tended to be more a function of compliance with the request to give information about the process of responding than a result of using the CMBS response mode. The personality variables giving significant beta-weights in this regression indicate higher G3 scores for examinees who were less conforming and also less enthusiastic. While, from one point of view, it would seem more reasonable for *more* conforming examinees, rather than *less* conforming examinees, to indicate greater amounts of guessing in an effort to comply with the directions, it also seems possible that more conforming individuals might be less amenable to changing their usual method of responding to a test, while the less conforming might be more

flexible. Some support for the latter position is found in the relationship of dogmatism scores to the G3 score component. While dogmatism was not significantly involved at the .05 level, it was significant at the .09 level, and its negative relationship indicates that the less dogmatic, *i.e.*, those examinees with more open belief systems, tended to have higher G3 scores than the more dogmatic.

The positive relationship of the Enthusiastic factor to the INFO and MIS1 scores and the negative relationship to the guessing and partial misinformation scores can also be viewed as generally supporting the expectations suggested by the literature. More confident examinees tended to mark four options for the test items, thus risking penalty for misinformation, while less confident, and by implication more anxious, examinees tended to mark fewer than four options, thus risking the loss of credit for tentatively-held, correct information.

It should also be noted that the Rotter scale, which purports to measure the degree to which an individual perceives himself to control outcomes, was found to produce a significant beta-weight only in the IG category in group 3. It was expected that perception of "luck" would play a much more important part in the regression analysis. The negative relationship of the Rotter scale to IG in group 3 suggests that those who perceived themselves as being in control of outcomes were more likely to refuse to guess at any of the options on items for which they were completely unsure of the correct answers. This result is consistent with locus of control theory, but it is surprising that perception of control did not appear more frequently in the analyses.

In terms of the amounts of variance in the score components which were explained by the personality and attitude variables, the largest $R^2$'s were found in group 1, even though few of them could be shown to be significant at the .05 level. There were also more and different factors involved in group 1 than in the other two groups. The results which were significant appear, in general, to support the previous findings in the types of variables involved and the direction of the involvement. More outgoing, confident examinees appear to be penalized by the CMBS mode, as evidenced by the negative relationship of the Sociable factor to INFO and the positive relationship to MIS1, MIS2, and MIS3. Examinees with greater feelings of repression and anxiety tended to mark fewer options than more enthusiastic examinees. However, these results do not provide for a clear delineation of the relationship between the score components and the personality and attitude variables.

The final step in the data analysis was to examine the relationship of the personality and attitude variables to measures derived from literature on personality and multiple-choice tests. The first of these measures was the "tendency to gamble" (Swineford, 1938) used by Jacobs (1971) and renamed an "expression of unwarranted confidence." The procedure for calculating this measure is based on the definition by Jacobs, as follows:

$$JCON = \frac{MIS1}{W} \times 100$$

where JCON represents Jacobs' confidence measure, MIS1 is the number of items with three options and the right answer identified as distractors, and W is the number of wrong responses. The MIS1 category is considered to be analogous to the confidence-weighted component of "number of errors for which maximum confidence was expressed" (Jacobs, 1971, p. 17).

The second of the measures from the literature was based on the concept of a "standard of assurance" described by Coombs, *et al.*, (1956). Based on responses made under the CMBS mode, a theoretical NR score was determined from the probabilities of selecting the right option under various conditions of guessing and misinformation. This theoretical score was defined earlier as the expected NR score (ENR). The difference between the NR and ENR scores represents information which the examinee has, but which is below his personal threshold of assurance. This index can also be viewed as a measure of the accuracy of the examinee's perception of his own information. It was defined as:

$$DIFF = (NR - ENR)$$

where DIFF is the measure of accuracy, NR is the total number of correct answers selected under the NR response mode, and ENR is an expected score based on CMBS responses. A positive DIFF score represents suppression of partial information or lack of confidence in partial information; a negative DIFF score indicates overconfidence or bad luck in guessing in this situation.

Swineford (1938) based part of her argument for the presence of a personality factor in confidence-weighted scores on relationships between the number of right and wrong items marked with highest confidence weights and other variables. To determine whether the results she predicted were present in these results, the "highest confidence" was imputed to INFO and MIS1 items, and these items were summed for what Swineford called an "All 4's" score.

The final composite measure used in this part of the analysis was an overall measure of the frequency of marking options under the CMBS response mode. This measure was called NOPT and was determined by multiplying the numbers of items in each score component by the number of marked options indicated by the component, *e.g.*, for the INFO and MIS1 score components, four options were marked as distractors. The formula used was:

$$\begin{aligned} \text{NOPT} = {} & 4(\text{Number of INFO items}) + 4\ (\text{Number of MIS1 items}\ ) + \\ & 3(\text{Number of G2 items}) + 3(\text{Number of MIS2 items}) + \\ & 2(\text{Number of G3 items}) + 2(\text{Number of MIS3 items}) + \\ & (\text{Number of G4 items}) + (\text{Number of MIS4 items}). \end{aligned}$$

The means, standard deviations, and correlations of JCON, DIFF, All 4's, and NOPT with each other and with total test scores are given in Table 22.

The mean JCON ratio of number of errors with four marked options to the total number of errors was signficantly lower in group 3 than in the other two groups. However, even in group 3, almost 50 percent of the errors were of the MIS1 type. This result is in contrast to

Table 22. Means, Standard Deviations, and Correlations of JCON, DIFF, All 4's, and NOPT with Each Other and with Total Scores.

| Variables | Means | S.D. | Correlations | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | DIFF | 4's | NOPT | NR | CMBS | ICW | EMP |
| Group 1 | | | | | | | | | |
| JCON | 58.79[1] | 29.55 | -.46 | .79 | .69 | -.10 | .09 | -.13 | -.13 |
| DIFF | 3.69 | 2.86 | --- | -.81 | -.80 | -.16 | -.59 | -.29 | -.17 |
| All 4's | 18.07[2] | 6.01 | | --- | .88 | .17 | .45 | .19 | .13 |
| NOPT | 85.29[3] | 18.65 | | | --- | .22 | .57 | .23 | .20 |
| Group 2 | | | | | | | | | |
| JCON | 63.37[1] | 32.71 | -.44 | .85 | .70 | .05 | .05 | .03 | .07 |
| DIFF | 2.98 | 2.28 | --- | -.79 | -.72 | -.28 | -.47 | -.37 | -.31 |
| All 4's | 19.19[2] | 6.09 | | --- | .86 | .20 | .29 | .21 | .24 |
| NOPT | 91.22[3] | 11.28 | | | --- | .33 | .46 | .30 | .38 |
| Group 3 | | | | | | | | | |
| JCON | 47.92[1] | 31.50 | -.38 | .79 | .72 | .06 | .02 | .06 | .07 |
| DIFF | 3.53 | 2.38 | --- | -.76 | -.71 | -.05 | -.37 | -.20 | -.12 |
| All 4's | 16.60[2] | 5.61 | | --- | .93 | .27 | .35 | .34 | .29 |
| NOPT | 88.82[3] | 12.78 | | | --- | .32 | .42 | .37 | .35 |

[1]Group 3 significantly different from groups 1 and 2 in analysis of variance; no significant difference between groups 1 and 2.
[2]Significant difference between groups 2 and 3; no significant difference between groups 1 and 2 or between groups 1 and 3.
[3]Significant difference between groups 1 and 2. No difference between groups 1 and 3 or between groups 2 and 3.

Jacobs' (1971) finding that no more than 18 percent of the errors made by examinees were made with highest confidence. The difference between these two studies suggests that students may be more careful in the use of the maximum confidence weight under actual confidence-weighting instructions than can be inferred from their marking of four options under CMBS instructions. It is evident from Table 22, however, that group 3, which was operating under the greatest penalty, was either more cautious in the marking of options or better able to appraise the accuracy of their information. There was no difference in the mean JCON scores between groups 1 and 2, although group 2 had the highest JCON score of all the groups. This suggests, again, that the imposition of no penalty and two levels of penalty did not affect the three groups linearly.

The lack of significant differences in the mean DIFF scores also argues against group 3 being better able to appraise the accuracy of their information. For all three groups, an average of approximately three to four points was present in NR scores in comparison to the score which would be expected if the CMBS markings had correctly reflected the accuracy of the information.

There is further evidence for the cautiousness of group 3 in the number of right and wrong items with four options marked (All 4's). Group 3 had the fewest 4's and was significantly different from group 2. In the total frequency of marking options (NOPT), however, group 3 had slightly more than group 1, although the difference was not significant, and it was group 2 which marked the largest number of

options on all items. Thus, in general, group 3 appeared to be more cautious in marking options than group 2, but not necessarily more cautious than group 1. Group 2 examinees appeared to be the most liberal in marking options.

There was considerable overlap among these four measures, particularly for JCON, All 4's, and NOPT, as shown in Table 22. This would be expected, since these three measures are derived from or strongly influenced by the number of items with four options marked. The moderate negative relationship between JCON and DIFF is also logical. A low JCON score indicates a low amount of unwarranted confidence, but it may also show some timidity in marking options, while a high positive DIFF score reflects a discrepancy between the number of items answered correctly and projected score based on the examinee's expressed confidence in his answers, indicative of suppression of partial information or lack of confidence in responses. There is an element of risk-taking in the DIFF score, since the difference between the score expected from the CMBS markings and the NR score is, in part, a measure of the willingness of examinees to risk claiming credit for responses about which they are not completely sure. Less than 25 percent of the variance in JCON is explained by DIFF, however, and the unexplained variance between the measures may be attributable to the amount of unused partial information included in the DIFF measure, as compared to JCON, which is primarily a measure of unwarranted confidence and misinformation.

Coombs *et al.* (1956) and Swineford (1938) concluded that DIFF

and the tendency to gamble (JCON), respectively, were independent of achievement because of the low correlations they obtained for these measures and total test scores. DIFF was independent of NR and EMP achievement scores in groups 1 and 3, and moderate correlations were found for DIFF and NR and EMP in group 2. The correlation between DIFF and CMBS was, in each group, considerably higher than correlations between other scores. This result would seem to be logical, since it is in the CMBS scores that the greatest effect of partial information on scores can be seen. The smaller the discrepancy between NR scores and ENR scores, the more frequent the matching of number of marked options to actual information level and, thus, the higher the CMBS scores would be. JCON was found to be unrelated to all total test scores for all three groups.

To assess whether the personality variables were significantly involved in these measures of general response tendencies, a series of multiple linear regressions was performed. The results are reported in Table 23.

There were no significant total regressions on JCON for any of the groups. In group 2, however, the Enthusiastic factor produced a significant beta-weight, and in group 3, the Spontaneous factor produced a significant beta-weight. When only groups 1 and 2 are compared, the results are consonant with what would be expected based on the results of Jacobs (1971). When no risk is involved (as in group 1), the tendency to gamble or display unwarranted confidence is not active, but under a penalty (as in group 2), more enthusiastic

Table 23. Multiple Linear Regressions of Personality Factor Scores and Attitude Scales on JCON, DIFF, All 4's, and NOPT.

| Dependent Variable | Multiple R | $R^2$ | *F*-statistic for Total Regression | Significant[1] Beta-weights |
|---|---|---|---|---|
| Group 1 $n$=62 | | | | |
| JCON | .34 | .12 | 1.03 | None |
| DIFF | .48 | .23 | 2.26[1] | Assertive (+) |
| All 4's | .49 | .24 | 2.43[1] | Assertive (-)<br>Conformist (+)<br>Enthusiastic (+) |
| NOPT | .50 | .25 | 2.59[1] | Assertive (-)<br>Conformist (+)<br>Enthusiastic (+) |
| Group 2 $n$=77 | | | | |
| JCON | .40 | .16 | 1.92 | Enthusiastic (+) |
| DIFF | .25 | .06 | 0.67 | None |
| All 4's | .32 | .10 | 1.15 | Enthusiastic (+) |
| NOPT | .31 | .09 | 1.01 | None |
| Group 3 $n$=116 | | | | |
| JCON | .32 | .11 | 1.81 | Spontaneous (-) |
| DIFF | .23 | .05 | 0.84 | None |
| All 4's | .33 | .11 | 1.84 | Sociable (-) |
| NOPT | .33 | .11 | 1.94 | Sociable (-) |

[1]significant at $\alpha \leq .05$.

examinees tend to exhibit high confidence in wrong answers. This comparison is complicated, however, by the previous finding that group 1 was significantly less enthusiastic than group 2. While lower levels of a variable do not, themselves, imply a reduced correlation with another variable, if high enthusiasm is a significant determinant of JCON, there would be less opportunity for the relationship to be expressed in a group in which few individuals showed high enthusiasm. In addition, Jacobs (1971) was able to show a multiple correlation between four personality scales and JCON for 72 examinees of .39, significant in his study at $p < .05$. Although it was not possible to show significance of the multiple correlations in this study, in part because of small group sizes and more independent variables, the multiple correlation between JCON and the personality and attitude variables was .34 in group 1 and .40 in group 2. These results, even for group 1, are not widely disparate from Jacobs' correlation, and suggest considerable involvement of the personality and attitude variables in both groups, although group 1 was not under a penalty condition.

The negative relationship of the Spontaneous factor to JCON in group 3 is puzzling. It should be noted, however, that the Spontaneous factor was positively related to MIS1, which forms the basis for the JCON measure, only in group 1 (see Table 21). In both groups 2 and 3, the relationship was negative for MIS1. The same relationships are found in JCON (see the first column in Table 24). This suggests that the less impulsive and more tolerant of tedium were more likely

Table 24. Orderings of Standardized Beta-Weights from Regressions of Personality and Attitude Variables on JCON and DIFF.

| JCON | | DIFF | |
|---|---|---|---|
| Group 1 $n$=62 | | | |
| Conformist | .23 | Assertive | .27[1] |
| Enthusiastic | .17 | Enthusiastic | -.23 |
| Dogmatism | .12 | Spontaneous | -.20 |
| Assertive | -.12 | Dogmatism | -.18 |
| Spontaneous | .10 | Conformist | -.17 |
| Rotter | -.07 | Sociable | -.09 |
| Sociable | -.07 | Rotter | .03 |
| Group 2 $n$=77 | | | |
| Enthusiastic | .31[1] | Sociable | .17 |
| Spontaneous | -.13 | Enthusiastic | -.11 |
| Conformist | .13 | Rotter | -.09 |
| Assertive | -.10 | Spontaneous | -.07 |
| Dogmatism | .06 | Conformist | -.03 |
| Rotter | -.03 | Dogmatism | -.01 |
| Sociable | .00 | Assertive | .00 |
| Group 3 $n$=116 | | | |
| Spontaneous | -.21[1] | Enthusiastic | -.16 |
| Conformist | -.15 | Dogmatism | .13 |
| Sociable | -.11 | Sociable | .10 |
| Enthusiastic | .11 | Rotter | .06 |
| Assertive | .07 | Assertive | -.05 |
| Rotter | .06 | Conformist | .04 |
| Dogmatism | -.03 | Spontaneous | -.01 |

[1]beta weights significant at $\alpha \leq .05$.

to have a higher percentage of errors with four marked options. The expectation for this factor would seem to be the reverse. Those examinees who were more tolerant of tedium would seem more likely to be those who would spend a greater amount of time on an item, evaluating their level of information, while the more impulsive would be more likely simply to mark four options with only cursory attention to the implications of doing so in terms of their scores. This is, obviously, not the appropriate explanation for the behavior of the individuals in groups 2 and 3 in this study. The more impulsive examinees in groups 2 and 3, apparently, tended to mark fewer than four options on items about which they were misinformed. This difference in marking options between the more impulsive and less impulsive examinees did not appear to be a function of ability, however. When the extreme scorers in each group were identified (those whose scores fell more than one standard deviation above the mean and those who fell one standard deviation below the mean), there was no significant difference in mean NR or mean CMBS scores for the extreme scoring examinees on the Spontaneous factor in any of the three groups. Thus, it appears that the more impulsive examinees in groups 2 and 3 were more likely to incur the greater penalties for misinformation, since the pattern of marking four options, one of which is the right answer, results in the lowest misinformation penalty. Although the mean scores for the extremely high impulsives were lower in group 3 than the means for the low impulsives, the differences were not significant, and, in group 2, the mean scores for high impulsives were generally higher

than those of low impulsives, but, again the differences were not significant.

In general, in these three groups of examinees, there was an involvement of personality variables in JCON, the measure of unwarranted confidence, that was similar in magnitude to the multiple correlation obtained by Jacobs (1971), although the total regressions could not be shown to be significant. No significant beta-weights were obtained for any of the variables in group 1, but the Enthusiastic factor produced a significant beta-weight for group 2 and the Spontaneous factor for group 3. An inspection of the directions of relationships and the order of the size of standardized beta-weights within the three groups suggested that different mechanisms were operating in these relationships, but no general inferences could be drawn.

For the DIFF measure, only in group 1 was the total regression significant and only in group 1 was there a significant beta-weight for an affective variable. In group 1, the Assertive factor was positively related to the DIFF score, which was a measure of the discrepancy between the "rightness" or "luck" in guessing an examinee displayed in his NR score and the accuracy with which he perceived his knowledge level as reflected in an ENR score. This result is in line with previously reported findings. In group 1, the more assertive examinees tended to answer correctly more items whan would be expected by their marks under the CMBS response mode. The Assertive factor was negatively related to INFO in group 1 and positively related to IG. If this is interpreted as evidence that more assertive examinees chose

not to comply with the request to give information about the decision process on each item, then it would also be expected that more assertive examinees would appear to be poor estimators of their knowledge levels, hence the positive discrepancy between NR and ENR. In groups 2 and 3, in which the CMBS directions were given, the DIFF measure is independent of personality and attitude variables.

Examination of the orderings of the standardized beta-weights for DIFF in the three groups suggests, again, that different mechanisms are operating in the involvement of the personality and attitude variables in the three groups. It also seems evident that those variables which most strongly affect the JCON scores are not those which were influential on the DIFF scores.

In group 1, both All 4's and NOPT produced significant regressions and several significant beta-weights, reflecting the same characteristics which were significant in the information and misinformation scores in Table 20. Neither All 4's nor NOPT produced significant regressions in groups 2 and 3, although in group 2 the Enthusiastic factor scores gave a significant beta-weight for All 4's, and in group 2, the Sociable factor was significant for both dependent variables. For both of these groups, the factors appearing as significant are those expected by the previous results.

Summary of Personality and Attitude Variables and Test Scores

It was assumed that affective characteristics would be evoked when examinees were asked to judge the accuracy of their own

information under the CMBS response mode, and that, as a result, the regressions of personality and attitude variables on CMBS scores would, particularly in groups 2 and 3, provide a greater explanation of the variance in the responses than regressions of the same variables on NR scores. No evidence was found to support this assumption. In all three groups, essentially the same percentages of variance were accounted for by personality and attitude variables for both CMBS and NR scores. In addition, the standardized beta-weights of the personality and attitude variables in the regressions suggested that the same variables were involved, in approximately the same ways, in both types of scores within groups.

Taken by itself, this result would seem to indicate that the use of the CMBS response mode did not interact with any measured personality characteristic in examinees in a way that was significantly different from NR responses. Such an inference cannot be drawn without qualification, however, since the three groups used in the study were operating under different levels of penalty and there were differences among the groups in the strength and type of relationships between total scores, score components, measures from the literature and the personality and attitude variables.

The differences between the groups seemed to suggest that compliance with the request to give information was the primary determinant of the personality and attitude variables elicited in groups 1, while considerations more directly related to the CMBS mode were involved in groups 2 and 3. There also appeared to be some differences in the

relative importance of the personality and attitude variables between groups 2 and 3, but the differences could not be clearly defined.

In general, it appeared that more confident examinees tended to mark more options than the less confident and, thus, rendered themselves more susceptible to penalties for misinformation. The personality and attitude mechanisms in the guessing patterns, whether indicating partial information or some level of misinformation, were more diffuse, and, though the implication seems strong that more anxious examinees tended to respond more tentatively than the less anxious, the way in which this occurred could not be determined from the regression results.

# CHAPTER V

# DISCUSSION AND CONCLUSIONS

## The Effects of Partial Information and Misinformation on Reliability and Validity

### Reliability

In all three groups of examinees, the removal of the guessing components from total scores resulted in enhanced reliability estimates. In the context of Frary's (1968, 1969) theoretical model of the effects of guessing on reliability, the implication of this finding was that the guessing scores were negatively related to true scores. Using estimates of guessing scores and true scores, determined from the data in the same manner as Frary, it is possible to examine further the application of the theoretical expectations to the observations in this study. The guessing scores are defined as the items which were answered correctly through guessing, either randomly among all five alternatives or with some partial information. The best estimates of true scores for each group are the sums of those items answered correctly on which examinees claimed to have complete information. These true score estimates have been referred to previously as information scores (INFO) and also as guessing-free scores (GF). On repeated administrations of the test, the guessing scores would, of necessity, fluctuate and represent and unreliable source of variance, while the true score estimates would remain

relatively constant, except for mechanical errors in the test-taking.

To facilitate this discussion of the theoretical implications of these results, three correlation matrices are given in Tables 25, 26, and 27. These matrices show the interrelationships between total test scores and the various score components calculated in the analysis section. The abbreviations used in the correlation matrices refer to the following scores:

1. NR - Number-right test scores
2. CMBS - Coombs response mode test scores
3. EMP - Empirical choice-weighted test scores
4. INFO - Number of items answered with four distractors correctly marked.
5. G2 - A partial information measure consisting of those items for which three distractors were correctly identified.
6. G3 - A partial information measure consisting of those items for which two distractors were correctly identified.
7. G4 - A partial information measure consisting of those items for which one distractor was correctly identified.
8. GG2 - A guessing score component consisting of those items for which three distractors were correctly identified and the right answer was indicated between the two remaining options.
9. GG3 - A guessing score component consisting of those items for which two distractors were correctly identifed and the right answer selected from among the three remaining options.
10. GG4 - A guessing score component consisting of those items for which one distractor was correctly identified and the right answer selected from among the four remaining options.
11. GG5 - A guessing score component consisting of those items for which no distractors were marked, but the right answer was selected from among five options.
12. AllP - All partial information, a summation of G2, G3 and G4.

Table 25. Intercorrelations Between Total Test Scores, Information, Misinformation, and Guessing Scores, and Composite Measures for Group 1 ($n$=69).

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 NR * | .86 | .95 | .68 | .17 | -.10 | -.18 | .24 | -.06 | -.13 | -.03 | -.07 | .11 | -.21 | .64 | .30 | .18 | .29 | .76 |
| 2 CMBS | | .97 | .85 | .20 | -.20 | -.37 | .14 | -.21 | -.34 | -.44 | .02 | -.26 | -.56 | .42 | .35 | .27 | .30 | .56 |
| 3 EMP | | | .60 | .22 | -.06 | -.16 | .27 | -.02 | -.12 | .02 | .13 | .18 | -.13 | .70 | .23 | .14 | .25 | .81 |
| 4 INFO | | | | -.28 | -.46 | -.41 | -.32 | -.42 | -.37 | -.47 | -.48 | -.65 | -.58 | .09 | .43 | .35 | .28 | .25 |
| 5 G2 | | | | | .27 | -.15 | .91 | .15 | -.18 | -.16 | .89 | .56 | -.21 | .33 | -.31 | -.06 | .00 | .24 |
| 6 G3 | | | | | | .33 | .27 | .87 | .29 | .16 | .62 | .52 | .25 | .36 | -.26 | -.48 | -.05 | .24 |
| 7 G4 | | | | | | | -.10 | .39 | .98 | .27 | .23 | .37 | .43 | .17 | .04 | -.42 | -.10 | .14 |
| 8 GG2 | | | | | | | | .13 | -.13 | -.08 | .83 | .67 | -.11 | .33 | -.29 | -.08 | -.04 | .24 |
| 9 GG3 | | | | | | | | | .41 | .20 | .50 | .50 | .30 | .35 | -.22 | -.40 | -.11 | .24 |
| 10 GG4 | | | | | | | | | | .28 | .19 | .36 | .42 | .16 | .08 | -.41 | .02 | .15 |
| 11 GG5 | | | | | | | | | | | -.02 | .60 | .89 | .26 | -.07 | -.06 | -.06 | .33 |
| 12 A11P | | | | | | | | | | | | .73 | .02 | .44 | -.33 | -.30 | -.04 | .32 |
| 13 A11G | | | | | | | | | | | | | .57 | .54 | -.28 | -.26 | -.09 | .44 |
| 14 IG | | | | | | | | | | | | | | .37 | -.11 | -.10 | -.29 | .32 |
| 15 MIS1 | | | | | | | | | | | | | | | -.17 | -.13 | -.11 | .94 |
| 16 MIS2 | | | | | | | | | | | | | | | | .30 | .37 | .16 |
| 17 MIS3 | | | | | | | | | | | | | | | | | -.03 | .04 |
| 18 MIS4 | | | | | | | | | | | | | | | | | | .02 |
| 19 A11M | | | | | | | | | | | | | | | | | | -- |

*Definitions of abbreviations in text.

Table 26. Intercorrelations Between Total Test Scores, Information, Misinformation, and Guessing Scores, and Composite Measures for Group 2 ($n$=86)

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 NR* | .95 | .94 | .68 | .11 | -.11 | -.30 | .15 | -.01 | -.24 | -.10 | -.01 | .07 | -.17 | .62 | .20 | .44 | .40 | .86 |
| 2 CMBS | | .94 | .74 | .10 | -.15 | -.41 | .06 | -.11 | -.33 | -.22 | -.05 | -.06 | -.26 | .55 | .22 | .54 | .54 | .84 |
| 3 EMP | | | .67 | .11 | -.14 | -.43 | .12 | -.08 | -.33 | -.08 | -.05 | .03 | -.14 | .50 | .17 | .60 | .61 | .80 |
| 4 INFO | | | | -.56 | -.54 | -.46 | -.57 | -.50 | -.44 | -.16 | -.69 | -.69 | -.17 | .06 | .47 | .43 | .35 | .37 |
| 5 G2 | | | | | .37 | .03 | .96 | .38 | .10 | -.15 | .89 | .87 | -.18 | .53 | -.54 | -.07 | .07 | .36 |
| 6 G3 | | | | | | .67 | .37 | .91 | .68 | .02 | .73 | .63 | -.01 | .40 | -.18 | -.26 | -.18 | .25 |
| 7 G4 | | | | | | | .07 | .59 | .92 | .04 | .44 | .33 | .02 | .24 | -.01 | -.67 | -.67 | -.04 |
| 8 GG2 | | | | | | | | .40 | .14 | -.11 | .87 | .92 | -.15 | .49 | -.50 | -.07 | .05 | .33 |
| 9 GG3 | | | | | | | | | .54 | .06 | .70 | .67 | .02 | .38 | -.16 | -.12 | -.20 | .26 |
| 10 GG4 | | | | | | | | | | -.06 | .48 | .36 | -.05 | .25 | -.05 | -.59 | -.44 | .01 |
| 11 GG5 | | | | | | | | | | | -.10 | .12 | .95 | .02 | .12 | .05 | -.05 | .06 |
| 12 A11P | | | | | | | | | | | | .93 | -.14 | .58 | -.47 | -.27 | -.14 | .35 |
| 13 A11G | | | | | | | | | | | | | .07 | .54 | -.43 | -.15 | -.08 | .35 |
| 14 IG | | | | | | | | | | | | | | -.04 | .17 | .05 | -.02 | .03 |
| 15 MIS1 | | | | | | | | | | | | | | | -.26 | -.16 | -.09 | .86 |
| 16 MIS2 | | | | | | | | | | | | | | | | .10 | .02 | .08 |
| 17 MIS3 | | | | | | | | | | | | | | | | | .83 | .26 |
| 18 MIS4 | | | | | | | | | | | | | | | | | | .29 |
| 19 A11M | | | | | | | | | | | | | | | | | | -- |

*Definitions of abbreviations in text.

Table 27. Intercorrelations Between Total Test Scores, Information, Misinformation, and Guessing Scores, and Composite Measures for Group 3 ($n$=123)

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 NR * | .90 | .92 | .63 | .01 | -.22 | -.11 | .19 | -.11 | -.02 | -.02 | -.11 | .09 | -.18 | .54 | .42 | .40 | .21 | .79 |
| 2 CMBS | | .85 | .72 | -.01 | -.33 | -.14 | .03 | -.28 | -.10 | -.06 | -.18 | -.11 | -.18 | .53 | .51 | .48 | .25 | .84 |
| 3 EMP | | | .60 | .02 | -.28 | -.12 | .15 | -.13 | -.00 | -.11 | -.13 | .05 | -.21 | .49 | .40 | .33 | .24 | .71 |
| 4 INFO | | | | -.64 | -.70 | -.21 | -.57 | -.62 | -.24 | -.01 | -.81 | -.72 | -.07 | -.04 | .48 | .49 | .21 | .28 |
| 5 G2 | | | | | .37 | -.02 | .91 | .32 | .11 | -.13 | .88 | .83 | -.23 | .54 | -.25 | -.25 | -.05 | .34 |
| 6 G3 | | | | | | .19 | .37 | .92 | .24 | .09 | .76 | .70 | .13 | .32 | -.20 | -.33 | -.14 | .14 |
| 7 G4 | | | | | | | -.02 | .21 | .84 | .08 | .23 | .17 | .24 | .13 | .00 | -.11 | -.13 | .10 |
| 8 GG2 | | | | | | | | .33 | .08 | -.10 | -.13 | .05 | -.21 | .52 | -.20 | -.22 | -.05 | .35 |
| 9 GG3 | | | | | | | | | .08 | -.10 | .81 | .90 | .08 | .28 | -.17 | -.23 | -.13 | .14 |
| 10 GG4 | | | | | | | | | | -.04 | .69 | .69 | .10 | .16 | .00 | -.12 | .02 | .13 |
| 11 GG5 | | | | | | | | | | | -.04 | -.01 | .52 | .09 | -.08 | .05 | .02 | .05 |
| 12 A11P | | | | | | | | | | | | .93 | -.07 | .54 | -.27 | -.35 | -.12 | .32 |
| 13 A11G | | | | | | | | | | | | | -.07 | .53 | -.23 | -.27 | -.08 | .34 |
| 14 IG | | | | | | | | | | | | | | .03 | -.02 | .02 | .03 | .03 |
| 15 MIS1 | | | | | | | | | | | | | | | -.18 | -.08 | .05 | .85 |
| 16 MIS2 | | | | | | | | | | | | | | | | .22 | -.01 | .33 |
| 17 MIS3 | | | | | | | | | | | | | | | | | .42 | .24 |
| 18 MIS4 | | | | | | | | | | | | | | | | | | .16 |
| 19 A11M | | | | | | | | | | | | | | | | | | -- |

*Definitions of abbreviations in text.

13. All G - All guessing score components, a summation of GG2, GG3, GG4, and GG5.

14. IG - An ignorance measure consisting of all items for which no distractors were marked.

15. MIS1 - A partial misinformation measure consisting of those items for which three distractors and the right answer were marked.

16. MIS2 - A partial misinformation measure consisting of those items for which two distractors and the right answer were marked.

17. MIS3 - A partial misinformation measure consisting of those items for which one distractor and the right answer were marked.

18. MIS4 - A complete misinformation measure consisting of those items for which only the right answer was marked.

19. All M - All misinformation, a summation of MIS1, MIS2, MIS3, and MIS4.

The distinction between the gains-from-guessing scores that were removed from the NR scores in the analysis section and the partial information scores that were removed from the CMBS scores is maintained by the use of "GG" for the gains-from-guessing scores and "G" for the partial information scores.

Examination of the matrices indicates that the correlations between INFO and the gains from guessing correctly under NR scoring (the GG2, GG3, *etc.* categories) are all negative. In all three groups, the correlations between INFO and the largest guessing components (GG2 and GG3) range from -.32 to -.62. Differences between the responses of group 1 and groups 2 and 3 are again suggested by the relationships found between INFO and the GG5 score which reflects the gain from completely random guessing. In group 1, the correlation between INFO and

GG5 was substantial (-.47); in group 2, the relationship was much less pronounced (-.16); and, in group 3, it was negligible (-.01). However, in groups 2 and 3, examinees marked very few items as guessing among 5 alternatives (respective means, .44 and .32), while this score component was significantly larger in group 1 (mean of 2.25). Thus, there was little expression of completely random guessing in groups 2 and 3, and, as a result, the correlations for these groups may not be meaningful. The average correlations for the gains-from-guessing components were very similar for all three groups ($\bar{r}_1$ = -.39; $\bar{r}_2$ = -.43; $\bar{r}_3$ = -.39), however, and the overall results tend to support the conclusion that when the guessing components are negatively related to true scores, the removal of guessing can be expected to improve internal consistency reliability.

It is also of interest that the correlation between NR scores and GG2 (guessing correctly between two alternatives) is positive, although the remainder of the guessing scores bear a negative relationship to NR as well as to INFO. This positive relationship suggests that the partial information reflected in guessing between two alternatives may be an integral part of the dimension measured by NR scores and, as such, represents a relatively consistent source of score variance. Its removal from NR scores, then, may not affect reliability as importantly as the removal of the negatively related components. This observation seems to be borne out by the effects reported previously in Table 13, which indicate that for group 1, in which the correlation between NR and GG2 is the largest (.24), the major improvements in

reliability come from the removal of the negatively related components. When the effects of guessing between two alternatives are removed, the increase in reliability is only from .80 to .81 in group 1. In groups 2 and 3, in which the correlations are slightly smaller in size (.15 and .19, respectively), the removal of the negatively related components also improves the reliability estimates, but the final removal of the guessing-between-two component produces an increase from .73 to .82 in group 2 and from .71 to .79 in group 3.

The relationship between partial information, as reflected in the CMBS scores, is not so clear, however. If was inferred that the relationship between partial information and true score would be positive, since the removal of the partial information component appeared to have less effect on reliability than the removal of the guessing scores. The correlation matrices in Tables 25, 26 and 27 do not bear out this hypothesis, however. The partial information components, represented as G2, G3, and G4, are related to INFO scores in a negative manner and with similar magnitudes (-.21 to -.70) as those occurring in the relationships of the GG2, GG3, and GG4 scores to INFO INFO. The slight effect on reliability from the removal of partial information was found clearly only in group 1, however; in groups 2 and 3, the effects were more pronounced. This difference between group 1 and groups 2 and 3 is reflected in the average correlations of the partial information components with INFO ($\bar{r}_1$ = -.39; $\bar{r}_2$ = -.52; $\bar{r}_3$ = -.55), and in the relationship of the all partial information measure (A11P) shown in the correlation matrices. For group 1, the

correlation between AllP and INFO was -.48; in group 2, it was -.69; in group 3, -.81. Since the negative relationship between AllP and INFO in groups 2 and 3 was considerably larger than in group 1, the removal of partial information in groups 2 and 3 would be expected to affect reliability more strongly than its removal in group 1. Although it is argued here that the enhancement of reliability found in CMBS scores as compared to NR scores can be attributed to the reduction or elimination of guessing, there appears to be an extraneous component in the partial information scores which may persist under the CMBS mode. It also seems, however, that, to the extent that this extraneous component continues to be present, the effect of the CMBS mode on reliability is reduced. This can best be seen in a comparison of the reliability estimates for NR scores, CMBS scores, and INFO scores in the three groups. These estimates are, respectively: in group 1 .67, .76, .81; in group 2 .66, .75, .82; in group 3 .60, .63, .79. Taking the INFO score as the best estimate of true score, representing the reliability which would be found if there were no guessing in the total scores, the more nearly the reliability of the total scores approaches that of the INFO scores, the less guessing is assumed to be present in the responses. In group 1, the difference between NR reliability and INFO reliability was .14, but the CMBS score reliability, lying between NR and INFO, shows an improvement of .09 over NR reliability and represents 64 percent of the total improvement. In group 2, the increase for CMBS was 56 percent of the total increase, but in group 3, the CMBS showed an increase of only

16 percent of the total improvement. Thus, in groups 1 and 2, where the CMBS mode produced substantial increases in reliability, the CMBS reliability estimates were relatively close to the guessing-free INFO scores. In group 3, where the CMBS mode was accompanied by a very small increase in reliability, the removal of guessing had a far greater effect.

Even though the changes in reliability from the removal of guessing were of considerable size, it was not possible, with the tests used, to demonstrate a statistically significant increase for any group. The finding of enhanced reliability was noted for all groups, however, and is consistent with the work of Frary (1968) and Cross (1973). The weight of the evidence suggests that the removal of guessing components from scores does indeed improve reliability. It seems possible to speculate that the failure to find significance in the improvements may indicate a Type II error.

Even if the reliability improvements from the CMBS mode can be reasonably attributed to the reduction or elimination of guessing, it is still not clear why the amount of the improvement was so different among the groups, particularly in group 3. A possible explanation may lie in the greater amount of stress that was assumed to be present in group 3 because of the personal consequences of the test scores. Group 3 examinees may have misrepresented their partial information in an attempt to maximize their scores and avoid the misinformation penalty. As discussed previously, group 3 average total scores were higher than those in both groups 1 and 2, but there was no difference

in the number of items marked in the INFO category. Group 3 examinees indicated significantly more G2 and G3 items than did examinees in groups 1 and 2, and significantly less MIS1 (see Table 6). Thus, group 3 examinees may have been more cautious and indicated less certainty than was an accurate representation of their levels of information.

Misinformation did not prove to be a major source of unreliability in the CMBS scores. The largest change in CMBS reliability in all three groups occurred when the penalty for MIS4 was removed. MIS4 was the response which brought the most stringent CMBS penalty, but it was also the category into which the fewest responses fell and was absent or negligible for most examinees. While, on the one hand, the MIS4 responses might be expected to be associated with low test scores, they also reflect a considerable degree of assurance in the selection of responses. In these groups of examinees, however, the MIS4 pattern occurred so rarely as to have little or no effect on total scores.

The correlations between MIS4, as well as the other misinformation score components, and total scores (NR, CMBS, and EMP) were all positive, ranging between .14 and .70. Since misinformation scores are negative, a positive correlation would indicate that a smaller negative value (a low misinformation score) would be associated with higher total scores. This is a logically consistent result, since it would be assumed that the more information an examinee possessed, the less misinformation he would have.

The correlations between the true score estimate (INFO) and the

misinformation components were slightly different, however. For MIS2, MIS3, and MIS4, the correlations were positive and of reasonable magnitudes. The relationships between MIS1 and INFO were negligible, however, and in group 3 the correlation was slightly negative (-.04). Since INFO and MIS1 represent essentially the same confidence in the marking of responses--the difference being that INFO items contain four bits of correct information and MIS1 items contain three bits of correct information and one bit of incorrect information--it may be that one of the ways in which guessing occurs in CMBS scores is revealed by this relationship. There seems to be no logical reason why the MIS2, MIS3, and MIS4 scores should be positively related to INFO while the MIS1 scores are independent of INFO unless the MIS1 scores include some guessing beyond the actual knowledge level. The definition of MIS1 assumes that the marking of the right answer as a distractor indicates some degree of assurance on the part of the examinee that he has correctly selected a distractor. However, since there is no way to determine the point in the marking of the options when the right answer is incorrectly identified as a distractor, there is also no way to determine which, if any, of the INFO items are the result of fortunate guessing between two options (one of which is right and one of which is wrong) and which of the MIS1 items come from unfortunate guessing under the same circumstances. MIS1 carries the smallest misinformation penalty, representing a .25 point loss against a potential 1.00 point gain if the guessing had been successful. Thus, some examinees on some items may have attempted to increase

their scores in the short run by guessing on the fourth distractor after correctly identifying three distractors. Because random guessing is not a systematic source of variance, its presence would tend to reduce the association between variables.

Another explanation for the very low correlations between INFO and MIS1 may lie in the possibility that the appearance of one piece of misinformation in the context of three pieces of information is, inherently, a random phenomenon. The three correctly marked distractors in MIS1 items indicate that the misinformed person has almost the same amount of information about the item as the person with complete information. If the incorrect marking of the answer as a distractor is a purposeful act and not the result of guessing between two options, then it would seem unlikely, in the context of considerable correct information about the item, that misinformed responses would be consistent. In MIS2, MIS3, and MIS4 items, where smaller amounts of information are present, a stronger reverse relationship with INFO would be expected.

The composite measure of misinformation (AllM) indicates that the overall relationship between misinformation and INFO is positive and of modest size (.25 in group 1; .37 in group 2; and .28 in group 3). This was the expected result, as were the negative correlations found between the all-gains-from-guessing measure (AllG) and INFO (-.65 in group 1; -.69 in group 2; -.72 in group 3).

An unexpected finding, however, was that the correlations between AllM and AllG for the three groups were positive (.44 in group 1; .35

in group 2; .34 in group 3). These three coefficients from the three groups indicated that examinees with high information scores tended to have lower gains from guessing correctly and lower misinformation scores. The positive correlation between AllG and AllM, however, suggested that high gains from guessing were also associated with low misinformation scores, *i.e.*, misinformation scores that approached zero. On the face of it, this seemed to be a puzzling result, since it appeared reasonable to expect that, if lower guessing and lower misinformation were associated with high information, then lower guessing and lower misinformation should also be directly associated, giving a negative correlation for these particular variables.

The mathematical possibility of the occurrence of the observed positive correlation between AllG and AllM was verified by using a procedure that determines the limits of the correlation between two variables, given their correlations with a third (Stanley & Wang, 1969). The limits of $r_{12}$, given $r_{13}$ and $r_{23}$ , are:

$$r_{13}r_{23} \pm \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}.$$

(Stanley & Wang, 1969, p. 580).

Applying this formula to the correlations for the three groups, the limits for the AllG/AllM correlation in group 1 (.44) were determined as -.90 to .57; for group 2 (.35), the limits were -.93 to .42; for group 3, (.34), -.85 to .49. Thus, while it is highly likely that the AllG/AllM correlations could be negative, since there was larger range of possible negative values within the limits of the coefficients, the

obtained positive values also lay within the limits. The implication of this positive correlation appears to be that low misinformation scores are likely to accompany high levels of partial information as well as high levels of complete information. Although it may seem reasonable to expect that misinformed examinees would also be forced to guess more frequently, the evidence from these data is to the contrary. The result is consistent with the work of Frary, Cross and Lowry (1977) on informed and random guessing.

Validity

The removal of the guessing components from NR scores, as well as the removal of partial information from CMBS scores, tended to decrease criterion-related validity. Since the effects were similar for both sets of scores, only the NR results were reported. Although the changes were not equally graduated for all guessing components, and in some cases the change was negligible, for all the criteria in all three groups there was a considerable loss of validity from NR to GF scores. Many of these decreases were significant at or beyond the .05 level. Frary (1968) argued that for the elimination of guessing score scores to be expected to produce a substantial validity increase, there must be a substantial difference between the predictor and the criterion tests. He continued, "Either random guessing is encouraged on one test and restricted on another or guessing scores are absent or uncorrelated with true scores on one of the tests" (Frary, 1968, p. 43).

It was expected that the substantial difference described by Frary (1968) might be present when the ONR scores were compared to the NR scores. It was assumed that guessing would be minimal on the ONR, which required the examinees to write the corrections they believed to be appropriate in their test booklets, while with the multiple-choice format of the NR responses, the usual amount of guessing was expected to occur. Indeed, guessing was encouraged under NR, and examinees were urged not to leave any item without indicating a choice of a single best answer. Apparently, the guessing was similar on both response modes, and the removal of the guessing components from NR scores did not improve the correlation between NR and ONR. Rather, there were significant decreases in validity in all three groups when the GF scores were determined.

Similar results were found for the other criteria. There was no reason to suppose that guessing would be different on the remaining parts of the multiple-choice test as compared to NR scores on the 25-item test, except to the extent that the special directions accompanying the 25-item test affected behavior. For the groups of examinees in this study, it appears that guessing behavior was highly similar across all the criteria used, and, thus, that the removal of the guessing components from NR scores reduced the similarities between the scores and the criteria, and, thereby, their correlations with each other. CMBS scores, with the partial information components removed, were also compared to the criteria, but the resulting validity coefficients were essentially the same as for the removal of guessing from

the NR scores.

Personality Characteristics and Alternative Procedures

The results of the regressions of personality characteristics on test scores suggested that there may be a generalized affective involvement when CMBS instructions, with a penalty greater than zero for a wrong answer, are given. Further, the generalized effect appeared to continue through other parts of the test even after the CMBS instructions were removed.

It should be noted, however, that more than the CMBS instructions were involved in the testing situation. Examinees were asked to respond to questions in three separate ways. The burden of these complicated instructions, regardless of the actual response mode included, may have been sufficient to produce the emotional responses. However, it seemed quite clear that the effect was different in group 1 from the effect in groups 2 and 3. Since all three groups were asked to do exactly the same things, the difference may be attributable to the penalty that was imposed on groups 2 and 3. On the other hand, the difference may be due to age, experience, or some other unmeasured variable.

Another factor which must be considered, since the affective response was pervasive and not localized in the CMBS scores, is the possibility that the effect observed was the result of novel instructions with a penalty rather than the judgments examinees were required to make about their responses under the CMBS mode. A limitation of this

study and of many of the studies reviewed earlier is that the CMBS response mode, when given for the first time to a group of examinees, constitutes an unfamiliar condition. It is not clear, and it cannot be clear until groups of examinees are used for whom the CMBS mode is a normal way of responding to a test, to what the effects reported are directly due. The implication is that the CMBS mode has effects on examinees beyond the effects of the NR mode, and this implication is in line with other studies which have attempted to provide examinees with sufficient practice with the CMBS mode so that they can use it intelligently and without undue concern about the novelty of the method. However, the problem of separating the effect due to the CMBS mode and the effect from the novelty of the procedure should be investigated further.

Another limitation, inherent in the CMBS mode, is its essentially self-report nature. It is not known, for example, how often examinees simply selected the right answer and marked four distractors without consideration and judgment of each option. To the extent that this occurred, the CMBS responses were a mirror image of NR responses, and the unique features of the CMBS mode did not come into play. This problem could be controlled to some extent by the use of an interactive terminal on which the options were displayed singly, and judgments required at each display. Such an approach, however, may be a serious departure from the usual multiple-choice format, and would likely introduce other extraneous variables into the responses.

In general, the results of this study tend to align with the

expectations suggested by the literature. Internal consistency reliability was enhanced by the CMBS mode and by EMP scoring; validity remained the same or declined; and, there was evidence of involvement of personality and attitude variables in the CMBS scores. However, there was also evidence of personality and attitude involvement in all test scores when the CMBS directions with penalty for wrong answers were given. This suggests that it may not be the CMBS response mode itself that evokes the personality and attitude variables, but the unusual nature of the directions.

In addition, in the two groups which received the complete CMBS instructions, there was some evidence that partial information and complete information were not the same dimension of ability and, possibly, that all bits of information, indicated by marking individual options within items, should not be treated identically under the CMBS mode. The validity results were particularly disturbing in this regard. It seemed clear that CMBS scores did not bear the same relationship to the criteria as NR scores, although not all of the differences were significant. There were also indications that EMP scoring, which also accounts directly for partial information, bore a different relationship to the criteria than did NR scores.

Overall, then, it may be that the major concern with the use of the CMBS mode should not lie in the possible inclusion of personality and attitude variables in the test scores, but in the effects of the direct measurement of partial information. The advantage of enhanced reliability with the CMBS mode may be offset by the indications that

a different dimension of ability is being measured by CMBS responses. There is, however, no reason to assume that the ability dimension measured by conventional test procedures is the best or only dimension that should be included in an evaluation of student achievement. Even if partial information differs in some ways from complete information, the inclusion of partial information may be a desirable goal. Under these circumstances, declines in validity, as compared to conventional scores, would be expected, and, indeed, if they did not occur, the procedure would be suspect.

The results of this study suggest that further investigations into partial information and misinformation would be productive if examinees familiar with the CMBS mode are used and if the test is more difficult so that a more complete representation of the partial information and misinformation score components can be obtained. If it can be established that the personality and attitude involvement in CMBS scores is a function of the novelty of the directions rather than a stress inherent in the response mode itself, the CMBS mode will offer the advantage of assessing partial information without the affective involvement reported for other alternative procedures. To the extent that the direct measurement of partial information is required, the CMBS mode will then provide a reasonable alternative to conventional multiple-choice testing.

The enhancement of reliability that accompanies the CMBS response mode appears to be derived from the reduction in guessing because of the direct inclusion of partial information in the test scores.

However, the relationship of partial information to complete information is far from clear, and the role of misinformation in CMBS scores also requires further definition. Considerable empirical and theoretical work remains to be done.

# CHAPTER VI

## SUMMARY

The purpose of this study was to examine the effects of partial information, misinformation, and personality characteristics on the reliability and validity of alternative response and scoring procedures in multiple-choice tests. The theoretical basis of the study was derived from previous work with various response modes and scoring rules other than conventional number-right scoring. The earlier work suggested that alternative procedures enhanced test reliability, but also might introduce extraneous personality variables into the scores. This affective dimension appeared to be evoked when examinees were asked to report subjective judgments about the accuracy of their own information. Since one of the major differences between conventional number-right scoring and alternative procedures lay in the direct inclusion of the effects of partial information and misinformation in the test scores, these score components were the primary concern of the study.

The Coombs response mode and scoring rule was used as the means for identifying the components of examinees' scores. The Coombs mode asks examinees to mark all options for a multiple-choice item which they believe are wrong. A penalty amounting to four times the credit for each correct mark is exacted if the right answer is marked. Thus, maximizing test scores under the Coombs mode requires that examinees

judge their information levels accurately and mark only those options about which they are reasonably sure. The patterns of marking options allow for the determination of score components of complete information, partial information, ignorance, partial misinformation, and complete misinformation. Thus, the Coombs mode appeared to offer the desired subjective element and, in addition, to provide for delineation of the score components which were assumed to explain differences between the test characteristics of Coombs scores and conventional number-right scores.

The data were collected as part of a general evaluation of undergraduate teacher trainees at Virginia Polytechnic Institute and State University in Blacksburg, Virginia, during the 1978-79 academic year. Six instruments were used in the evaluation, and four of these provided the data for this study. The assessment of personality variables was done with the Adjective Check List (Gough, 1952). Two attitude scales, the Rokeach Dogmatism Scale (Rokeach, 1960), and the Rotter Internal-External Locus of Control Scale (Rotter, 1966) were used. The test on which the Coombs response mode was used was the first 25 items of the Missouri College English Test (Callis & Johnson, 1964).

Responses from 278 examinees, divided into three groups, were used in the study. The first group (group 1) was given the Coombs instructions, but was asked to follow them only to provide information about their decision-making processes. No penalty greater than zero was announced for wrong answers. The second (group 2) and third (group 3) groups were also given the Coombs instructions and were

additionally told that the Coombs penalty of -4 points for 5-option items would be imposed for wrong responses. These groups were assumed to differ in the amount of stress present in the testing situation, since the second group was aware that their responses would be used only in a general assessment of the training program, while the information from group 3 would form part of the basis for decisions on their individual acceptability to upper division professional studies. It was expected that the three groups would constitute a hierarchy of affective involvement, with the no-penalty group (group 1) showing the least effects and group 3, for which the information would be used in a personal way, showing the greatest effects.

This hierarchy was not found, however. Groups 1 and 2 showed the greatest similarities in internal consistency reliability, as estimated by Cronbach's alpha, and for both of these groups, the use of the Coombs response mode and scoring rule was accompanied by considerable increase in reliability over conventional number-right scores. The effect on reliability in group 3 was negligible. No statistically significant differences could be detected for any of the groups, how ever.

Four validity criteria were used: three sections of the Missouri College English Test in multiple-choice format and an open-ended response form of the 25 items subject to the Coombs response mode. With the Coombs mode, some significant declines in validity were found for groups 1 and 2. No significant declines were found for group 3.

In an attempt to isolate the specific roles of the various score

components of complete information, partial information, ignorance, partial misinformation, and complete misinformation on the changes in test characteristics, several rescoring procedures were used. In one set of comparisons, the effects of guessing were sequentially removed from number-right test scores. In each group, guessing had a depressive effect on reliability, and its removal increased the coefficients. From a theoretical viewpoint, it appeared, from this result, that the guessing scores were negatively related to true scores in all groups. This suggested that guessing behavior and the real knowledge of examinees, represented by estimated true scores, were different dimensions. It was expected, however, that partial information, shown in the Coombs responses, would be similar to real knowledge and, thus, have minimal effect on reliability. When the partial information components were removed from Coombs scores, the expected effect was found only in group 1. The removal of partial information enhanced reliability in groups 2 and 3, suggesting that the mechanism by which information and partial information operated in groups 2 and 3 was not the same as that in group 1.

Validity was found to decrease significantly for some of the criteria in all three groups when the guessing components were removed. The effect from the removal of partial information was very similar to the effect from the removal of guessing. The removal of misinformation components, however, had negligible effects on both reliability and validity in all three groups.

In the examination of the involvement of personality and attitude

variables when instructions requiring subjective assessments by examinees of their own knowledge, no unique effect could be attributed to the Coombs response mode. There was no significant personality involvement in any of the test scores in group 1, but in groups 2 and 3, significant personality involvement was found in all scores. An inspection of the ordering of the standardized beta-weights for the number-right and Coombs scores for each group suggested that the personality variables involved in the responses of groups 2 and 3 were different from those in group 1. It appeared that, in group 1, the factors involved in responses were related to compliance with the requested procedures for answering the test, while in groups 2 and 3 there seemed to be evidence that the affective response was more directly related to the Coombs mode. In groups 2 and 3, more outgoing students tended to receive lower test scores in both NR and Coombs response modes.

When the scores were separated into information, partial information, ignorance, partial misinformation, and complete misinformation components, similar results were found. The largest component of the scores for each group was made of those items answered with complete information. In group 1, the largest beta-weight was again interpreted as signifying compliance with the test directions. For groups 2 and 3, the largest beta-weight was interpreted as indicating the same effect as was found for total scores. Close examination of the orderings of the beta-weights for the information and misinformation categories tended to support the conclusion that group 1 was

different from groups 2 and 3, but only minor differences could be found between groups 2 and 3.

A measure of unwarranted confidence in responses, reported in the literature as being significantly correlated with personality variables (Jacobs, 1970) was not found to be so for this group. However, the magnitudes of the multiple correlations resulting from the regression analyses approximated those given by Jacobs. The examination of other measures from the literature suggested again that there may be a difference between the responses of group 1 as compared to groups 2 and 3, but very little difference between groups 2 and 3. No clear patterns of differences between groups 2 and 3 emerged from any of the analyses.

The results of this study suggest that, if the Coombs mode were the customary means of response to a test, the unique involvement of personality characteristics in the Coombs responses may not be as pronounced as that found in other alternative response modes and scoring rules. While the subjective judgments of probabilities under the Coombs mode may appear to be more complex than under other procedures, the fact that any correct response reduces the effect of an incorrect response may produce a situation that is significantly different from the either-penalty-or-credit situation in other procedures. The overall evidence suggests that the Coombs response mode can provide valuable additional information about examinees' knowledge. Whether this information is sufficiently valuable to outweigh the disadvantages of the Coombs mode, particularly in terms of validity, is not clear,

however. Nevertheless, the question of whether the added information can be gathered without the intrusion of extraneous personality variables seems worthy of further study.

## REFERENCES


Arnold, J. C., & Arnold, P. L. On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education*, 1970, *39*, 8-13.

Callis, R. & Johnson, W. *Missouri College English Test*. New York: Harcourt, Brace & World, Inc., 1965.

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

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

Cronbach, L. J. Test "reliability": Its meaning and determination. *Psychometrika*, 1947, *12*, 1-16.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

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

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

Cross, L. H., & Frary, R. B. Empirical choice weighting under "guess" and "do not guess" directions. *Educational and Psychological Measurement*, 1978, *38*, 613-620.

Cross, L. H., Ross, F. K., & Geller, E. S. *Using choice-weighted scoring of multiple-choice tests for the determination of grades in college courses*. Unpublished manuscript, 1978. (Available from College of Education, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 24061).

Davis, F. B. A note on the correction for chance success. *Journal of Experimental Education*, 1967, *35*, 43-47.

Davis, F. B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, *19*, 159-170.

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

Diamond, J. J. A preliminary study of the reliability and validity of a scoring procedure based upon confidence and partial information. *Journal of Educational Measurement*, 1975, *12*, 129-133.

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

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

Ferguson, G. A. *Statistical analysis in psychology and education* (3rd ed.). New York: McGraw-Hill, 1971.

Frary, R. B. *Elimination of the guessing component of multiple-choice test scores: Effect on reliability and validity and an evaluation of related item-weighting*. Doctoral dissertation, Florida State University, 1968.

Frary, R. B. Elimination of the guessing component of multiple-choice test scores: Effect on reliability and validity. *Educational and Psychological Measurement*, 1969, *29*, 665-680.

Frary, R. B. The effect of misinformation, partial information and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement* (in press).

Frary, R. B., Cross, L. H., & Lowry, S. R. Random guessing, correction for guessing, and reliability of multiple-choice test scores. *Journal of Experimental Education*, 1977, *46*, 11-15.

Gilman, D. A., & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, 1972, *9*, 205-207.

Gough, H. G. *The adjective check list*. Palo Alto, California: Consulting Psychologists Press, 1952.

Gough, H. G., & Heilbrun, A. B., Jr. *The adjective check list*. Palo Alto, California: Consulting Psychologists Press, 1965.

Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.

Hakstian, A. R., & Kansup, W. A comparison of several methods of

assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement*, 1975, *12*, 231-239.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, *7*, 75-82.

Hanna, G. S. *Improving reliability and validity of multiple-choice tests with an answer-until-correct procedure.* Paper presented at the meeting of the American Educational Research Association, Chicago, paril 1974.

Hanna, G. S. Incremental reliability and validity of multiple choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, *12*, 175-178.

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

Hansen, R. The influence of variables other than knowledge on probabilistic tests. *Journal of Educational Measurement*, 1971, *8*, 9-14.

Hendrickson, G. F. *The effect of differential option weighting on multiple-choice objective tests* (Report No. 93). Baltimore: The Johns Hopkins University, 1971.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement*, 1973, *33*, 135-141.

Hotelling, H. The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, 1940, *11*, 271-283.

Hritz, R. J., & Jacobs, S. S. *Risk-taking and the assessment of partial knowledge.* Paper presented at the meeting of the American Psychological Association, Miami Beach, Florida, September, 1970.

Hummel, T. J., & Sligo, J. R. Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, 1971, *76*, 49-57.

Jacobs, S. S. Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement*, 1971, *8*, 15-19.

Jacobs, S. S. Behavior on objective tests under theoretically adequate, inadequate, and unspecified scoring rules. *Journal of Educational Measurement*, 1975, *12*, 19-29.

Jensen, D. R., & Howe, R. B. Tables of Hotelling's $T^2$-distribution (Appendix A). In C. Y. Kramer, *A first course in methods of multivariate analysis*. Blacksburg, Va.: Author, 1972.

Kansup, W., & Hakstian, A. R. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, 1975, *12*, 210-229.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston, Inc., 1973.

Kogan, N. & Wallach, M. A. *Risk taking: A study in cognition and personality*. New York: Holt, Rinehart & Winston, 1964.

Kristof, W. On a statistic arising in testing correlations. *Psychometrika*, 1972, *37*, 377-384.

Lefcourt, H. M. *Locus of control: Current trends in theory and research*. Hillsdale, N. J.: Lawrence Erlbaum Associates, Inc., 1976.

Lord, F. M. Formula scoring and number-right scoring. *Journal of Educational Measurement*, 1975, *12*, 7-11.

Michael, J. J. The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, *5*, 307-314.

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

Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, *80*, (Whole No. 609).

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

SAS Institute, Inc. *SAS user's guide: 1979 edition*. Raleigh, N. C.: Author, 1979.

Serlin, R. C., & Kaiser, H. F. A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement*, 1978, *38*, 337-340.

Sherriffs, A. C., & Boomer, D. S. Who is penalized by the penalty for guessing? *The Journal of Educational Psychology*, 1954, *45*, 81-90.

Slakter, M. J. Risk-taking on objective examinations. *American Educational Research Journal*, 1967, *4*, 31-43.

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

Soderquist, H. O. A new method of weighting scores in a true-false test. *Journal of Educational Research*, 1936, *30*, 290-292.

Stanley, J. C., & Wang, M. D. Restrictions on the possible values of $r_{12}$, given $r_{13}$ and $r_{23}$. *Educational and Psychological Measurement*, 1969, *29*, 579-581.

Stanley, J. C., & Wang, M. D. Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, 1970, *30*, 21-35.

Swineford, F. The measurement of a personality trait. *Journal of Educational Psychology*, 1938, *29*, 295-300.

Traub, R. E., & Fisher, C. W. On the equivalence of constructed-response and multiple choice tests. *Applied Psychological Measurement*, 1977, *1*, 355-369.

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

Wang, M. W., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, *40*, 663-705.

Wiley, L. N., & Trimble, O. C. The ordinary objective test as a possible criterion of certain personality traits. *School and Society*, 1936, *43*, 446-448.

Ziller, R. C. A measure of the gambling response-set in objective tests. *Psychometrika*, 1957, *22*, 289-292.

The vita has been removed from the scanned document

# AN INVESTIGATION OF THE RELATIONSHIPS OF TEST CHARACTERISTICS AND PERSONALITY VARIABLES TO PARTIAL INFORMATION AND MISINFORMATION IN MULTIPLE-CHOICE TEST SCORES

by

Mary Burnette Giles

(ABSTRACT)

The purpose of this study was to determine the relationships between selected measures of personality and various score components obtained from a multiple-choice test administered under a response mode proposed by Coombs in 1953. Previous studies using response modes and scoring rules other than the conventional number-right procedures have generally revealed increases in test score reliability and little change in criterion-related validity. The rationale offered for most of these alternative procedures is that the direct inclusion of partial information in test scores provides additional information about levels of examinee knowledge and, therefore, ought to enhance test characteristics. Alternatively, however, the observed increases in test reliability may result from the introduction of a reliable but extraneous source of score variance associated with personality factors evoked when examinees are required to express their assurance about each answer. This study attempted to determine whether the reliability increases accompanying the use of the Coombs mode were due to personality contamination of the scores or to other characteristics of the response mode.

The examinees, 278 teacher trainees in a U.S. university, completed several personality measures and also answered an English achievement test with (1) the Coombs mode, (2) the conventional number-right responses, and (3) an open-ended response format. Their responses were grouped, based on three different penalty conditions for wrong responses. Various total test, information, and misinformation scores were calculated from these responses.

Reliability estimates for test scores under the Coombs mode were higher for all three groups than the estimates for number-right scores. Validity coefficients were unchanged or lower. Multiple linear regressions of the personality variables on the total scores indicated that there was no unique involvement of personality variables in the Coombs scores beyond that also present in number-right scores and in empirical choice-weighted scores. Thus, in these groups of examinees, Coombs directions had a pervasive effect on test responses that did not explain the reliability and validity changes that accompanied the Coombs procedure.

Since the added information about levels of examinee knowledge, assumed to be included in the Coombs scores, was in the form of direct credit for partial information and penalty for misinformation, explanation for the reliability and validity changes was sought in these components. Various rescoring procedures were used to isolate the effects of these components on reliability and validity estimates. From these analyses, it was determined that reliability was enhanced by the Coombs mode to the extent that the guessing components in the scores

were reduced or eliminated. The reduction of guessing was a function of the opportunity provided by the Coombs mode for examinees to express all bits of partial information. Validity was decreased to the extent that removing guessing reduced the similarity of the test to the validity criteria. Misinformation was found to have little effect on either reliability or validity. These results suggest that, under some circumstances, the Coombs mode may provide increased reliability without the personality contamination that is present in other alternative response modes.