IMPROVING THE MULTIPLE-CHOICE EXAMINATION QUESTION:
THE ASSESSMENT AND USE OF PARTIAL KNOWLEDGE

A thesis submitted in partial satisfaction of the requirements for the degree of Master of Arts in Psychology
by
Glenn Martin Barge

August, 1975
The thesis of Glenn Martin Barge is approved:

California State University, Northridge
August, 1975
Acknowledgments

I would like to thank Drs. Wittig and Tabachnick for their assistance and helpful criticisms throughout the development and progress of this project. Their expressions of interest and encouragement were deeply appreciated.

I am especially grateful to my chairman, Dr. Richard O'Connell, for his unwavering support and invaluable guidance. His efforts in both time and energy far exceeded all expectations. The model of excellence which he provided served as an inspiration pushing me forward when my strength was low. All he did will be remembered.

Finally, I wish to thank all those members of the staff of the psychology department who assisted me in this study and Dr. Robert Dear for providing the original idea.
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>Abstract</td>
<td>vii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Method</td>
<td>30</td>
</tr>
<tr>
<td>Results</td>
<td>33</td>
</tr>
<tr>
<td>Discussion</td>
<td>49</td>
</tr>
<tr>
<td>References</td>
<td>59</td>
</tr>
<tr>
<td>Appendix</td>
<td>65</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List of Responses in Terms of Levels of Certainty and Types of Outcome</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>List of Testing Procedures in Terms of Penalty and Weighting as Experimental Variables</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>List of Means for Each Dependent Variable and Each Testing Procedure</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>List of Frequencies of the Various Response Classes as Percentages of the Total Number of Responses</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>List of Reliability Coefficients and Improvement Factors for the Four Testing Procedures</td>
<td>41</td>
</tr>
<tr>
<td>6</td>
<td>Analysis of Variance on Mean Number of Omitted Responses</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>Analysis of Variance on Mean CONF Scores</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>Analysis of Variance on Mean G Scores</td>
<td>46</td>
</tr>
<tr>
<td>9</td>
<td>Analysis of Variance on Mean A Scores</td>
<td>48</td>
</tr>
<tr>
<td>10</td>
<td>Analysis of Variance on Mean TPQ Scores</td>
<td>50</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean TPQ Scores for low and high risk-takers in the four testing procedures</td>
<td>51</td>
</tr>
</tbody>
</table>
Abstract

Improving the Multiple-Choice Examination Question:
The Assessment and Use of Partial Knowledge

by
Glenn Martin Barge

Master of Arts in Psychology

August, 1975

This study attempted to evaluate modifications of the multiple-choice examination question using the criteria of validity, practicality, and student acceptance. Validity was assessed with regard to the separate issues of relevance and reliability. Test relevance was examined in view of possible influences of risk-taking tendency and capacity to represent levels of partial knowledge. Test reliability was measured directly using Hoyt's (1941) procedure and indirectly by assessing the amount of guessing behavior. The time required to complete the quizzes served as a practicality measure. Student acceptance was measured by a questionnaire at the termination of the experiment.

40 Ss were matched for academic ability and then assigned to four testing groups: conventional testing, correc-
The groups were defined in terms of two independent variables which were penalty for incorrect responses and weighting of credit for correct responses. Prior to the start of the five week series of quizzes, each S completed the Choice Dilemmas Questionnaire and the Risk-Taking on Objective Examinations test as measures of risk-taking tendency. In addition, all Ss were administered one practice quiz in order to allow some familiarity with the testing procedures.

The Risk-Taking on Objective Examinations test was found to be superior to the Choice Dilemmas Questionnaire as a measure of risk-taking tendency. The level of risk-taking tendency was not influential on testing behavior in any of the groups. Confidence weighting provided a better representation of partial knowledge than did probability testing. Direct comparison of reliability coefficients was not possible due to the small number of subjects used in this experiment. Confidence weighting proved to be a better suppressor of guessing than was probability testing. There were no significant differences when comparing time in testing for the different procedures. Student acceptance was high for conventional testing and probability testing and was low for correction testing and confidence weighting. A number of suggestions were made for further study.
Improving the Multiple-Choice Examination Question:
The Assessment and Use of Partial Knowledge

The multiple-choice test question is a widely used method of assessing performance on aptitude and achievement examinations. Its popularity as a measuring instrument is, to a large extent, accounted for by the ease with which it can be administered and scored. It is important to note that the most prevalent method of scoring, by counting the number of correct responses only, is based upon an assumed dichotomy of complete knowledge and complete ignorance with regard to each particular question. In using this scoring method, a correct choice receives full credit regardless of whether the choice resulted from complete knowledge, some degree of partial knowledge, or a complete absence of knowledge as with pure guesses. All incorrect choices receive no credit irrespective of any possible distinctions between choices resulting from pure guesses, some degree of partial knowledge, or some degree of misinformation. Thus, an "all-or-none" model for scoring ignores the existence of partial knowledge with its
accompanying complications. This model for scoring is not universally accepted and numerous arguments have been made against it. In challenging the "all-or-none" model, Coombs (1953) argued that "examinees with less than complete information on a given subject have considerable partial information and that this may be used as a valid basis for discriminating among them." This argument has been supported by the findings of a great number of studies (Coombs, Milholland, & Womer, 1956; Davis & Fifer, 1959; de Finetti, 1965; Dressel & Schmid, 1953; Echternacht, Boldt, & Sellman, 1972; Michael, 1968; Rippey, 1968ab, 1970; Shuford, Albert, & Massengill, 1966; Soderquist, 1936; Stanley & Wang, 1970; Stokes, 1966). The subjective reports of examinees lend further support to a rejection of the "all-or-none" model for scoring and the acceptance of a partial knowledge model. Thus, questions arise as to the validity of the multiple-choice question as it is commonly applied.

The purpose of this study was to determine whether the multiple-choice question could be modified in its application so as to improve its validity as a measuring instrument of educational achievement and aptitude.

One View of Validity

Cureton (1951) proposed that validity may be analyzed in terms of two separate components which he termed "relevance" and "reliability". The relevance of a testing instrument may be thought of as the degree to which it
accurately measures the trait or traits it purports to measure. The relevance of a test is decreased to the extent that it measures other lasting and systematic traits. For example, an arithmetic test that is highly dependent upon reading ability may be said to possess less relevance than an arithmetic test that is less dependent upon the ability to read. Relevance may also be seen as the sensitivity of the instrument in assessing meaningful differences. One would not use a bathroom scale to weigh gold bars when a difference of less than an ounce is important. The reliability of a testing instrument may be seen as a measure of the consistency with which it measures a trait. A reliable test is insensitive to temporary and/or chance factors. Such factors are present and are particularly potent whenever choices are made with some degree of uncertainty.

It may be instructive to view the test score of any individual as being derived from the formula: \( S = T + E_1 + E_2 \), where \( S \) is the obtained score, \( T \) is the "true" score, \( E_1 \) is the influence of lasting and systematic errors of measurement, and \( E_2 \) is the influence of temporary and/or chance errors. Improvements in validity would require a reduction or elimination of the error components of the obtained score. Using this conceptualization of validity, the multiple-choice question was considered.

**The Multiple-Choice Question**

**The Question**

In this study we are concerned with the multiple-choice
question which is made up of a stem, one correct alternative, and three incorrect or distractor alternatives. The following question can serve as an illustration: Who founded the psychoanalytic school of psychology? (a) Adler, (b) Freud, (c) Fechner, (d) Wundt. There is one correct choice (b), and three distractors (a), (c), and (d). The manner in which the examinee responds can be viewed not only in terms of his knowledge of the correct answer but also in terms of the certainty with which he holds that knowledge.

Classes of Response

We may view any test answer as being within one of three general classes of response. These classes are delineated in terms of the examinee's level of certainty in his choice. The three levels are complete certainty (Class 1), partial certainty (Class 2), or complete uncertainty (Class 3). These classes of response are presented in Table 1 in terms of levels of certainty and types of outcome. Class 1 responses may be termed complete information (1A) or complete misinformation (1B). There are two subclasses of partial knowledge responses (Class 2) defined in terms of the decision processes involved. The responses may be termed "informed guesses" when the examinee can eliminate, with complete certainty, one or more of the alternatives before choosing among the remaining alternatives in a random manner. The probability of a correct choice (2A) or incorrect choice (2B) depends upon the
<table>
<thead>
<tr>
<th>Levels of Certainty</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correct</strong></td>
<td><strong>Incorrect</strong></td>
</tr>
<tr>
<td>Complete</td>
<td>1A</td>
</tr>
<tr>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>&quot;enlightened guess&quot;</td>
<td>2C</td>
</tr>
<tr>
<td>&quot;informed guess&quot;</td>
<td>2A</td>
</tr>
<tr>
<td><strong>Complete Uncertainty</strong></td>
<td></td>
</tr>
<tr>
<td>&quot;random guess&quot;</td>
<td>3A</td>
</tr>
</tbody>
</table>
number of choices eliminated. Responses may be termed "enlightened guesses" when they are made with some, but less than complete, certainty. The choices will be either correct (2C) or incorrect (2D). Class 2D responses might be thought of as partial misinformation. The third class of response occurs when the examinee has absolutely no knowledge of the answer and makes a purely random choice from among the four alternatives. One would expect that, in terms of long-term probability, one quarter of such responses would be correct (3A) and three quarters would be incorrect (3B).

Scoring

The most common method of scoring is to assign one point credit for correct choices and no credit for all other responses. There is no differential treatment of responses from the three classes and so neither levels of certainty nor guessing behavior have any effect on the examinee's score. Thus we are confronted with the two major problems associated with the conventional multiple-choice question.

The Two Problems of Conventional Testing

Guessing

In so far as the obtained score is inflated by random guessing, the reliability of the test is decreased (Michael, 1968). Random guessing contributes both temporary and/or chance effects and thereby decreases the validity of the testing instrument. In terms of the classes of response, purely random but correct choices (3A) and partially random
but correct choices (2A) both create threats to test validity. The number of these responses should be reduced, eliminated, or treated in some way so as to minimize their influence on test scores.

**Partial Knowledge**

The inability of conventional scoring to measure levels of partial information has been documented (Cureton, 1966; Diamond & Evans, 1973; Lyerly, 1951; Michael, 1968; Quereshi, 1974). In postulating the existence of partial knowledge, Cureton (1966) suggested a seven-step scale, not unlike the responses classes presented in this study, and argued the inability of conventional scoring to differentiate between steps. Samejima (1972) asserted the weakness and superficial correctness of items chosen with less than absolute certainty and the inappropriateness of scoring such items identically. In questioning the "all-or-none" model of scoring, a number of researchers (Cureton, 1966; Frary, 1969; Glass & Wiley, 1964; Lyerly, 1951) suggested that those choices made with less than complete certainty are not necessarily random in nature and must not be treated as such in scoring. With regard to the classification of responses, all those with correct outcomes (1A, 2A, 2C, and 3A) receive one point credit and all those with incorrect outcomes (1B, 2B, 2D, and 3B) receive no credit. With this failure to distinguish responses of varying certainty, the conventional application of multiple-choice questions would seem to be a crude measuring instrument.
The errors in measurement created by this coarseness would decrease the relevance of the test.

**Two Possible Solutions**

Serious attempts at improving the multiple-choice examination have spanned a period of about fifty years. One strategy, that has been seen as a solution for the guessing problem, has been the use of penalty formulae.

**Penalty for Guessing**

The penalty for guessing solution generally involves the use of special instructions and correction formulae. The instructions inform the examinee that points will be taken from the score in relation to the number of incorrect answers. The examinee is usually admonished against making random guesses (Thorndike, 1971). The effects of this strategy are twofold. First, the instructions may have the effect of inhibiting purely random guesses and in doing so would increase the number of omitted responses (Cronbach, 1970; Ebel, 1965; Lindquist, 1953; Thorndike, 1971). It can be stated that to the extent that guesses are reduced in number the solution is of some value (Thorndike, 1971). The second component of the effect is the mathematical removal of a portion of the correct choices thought to be the result of random guessing. The use of this adjustment is based upon an acceptance of the "all-or-none" model of scoring. There is no distinction made between different classes of response and it is assumed that all incorrect choices were the result of random guessing. Nevertheless,
this solution is utilized quite often and is the approach used in some standardized tests such as the Graduate Record Examination. For this reason the penalty for guessing strategy was chosen for consideration in this study.

Weighting of Alternatives

While the penalty for guessing strategy is employed to increase test reliability, the weighting approach is an attempt to assess partial knowledge and thereby increase the relevance of the test. Increased relevance should result from a differentiation of the response classes as they are reflected in the test scores. While the weighting of alternatives may take on several forms, it is basically a method of discovering the test-taker's degree of confidence in his or her choices. Greater credit is given for correct choices with high confidence than for choices made with low confidence.

Penalty and Weighting as Experimental Variables

There are a number of specific testing procedures which may be classified in terms of the presence or absence of penalty or weighting mechanisms. For the purposes of this study, it was possible to select specific procedures that may be compared in terms of a factorial design. This approach in experimental treatment represents an improvement over previously published studies of testing procedures.

The Four Testing Procedures

Conventional

The conventional method of determining test scores is
to simply count the number of correct responses. There is no penalty imposed for incorrect guesses and therefore few, if any, responses are omitted. This absence of omissions is particularly true with power tests where there are no restrictive time limitations. There is no identification of response classes and therefore no differential treatment of correct choices from the different classes. The same absence of weighting is true with all classes of incorrect response. Responses of classes 1A, 2A, 2C, and 3A all receive the same credit which is usually one point. All responses from classes 1B, 2B, 2D, and 3B receive no credit. Therefore, conventional testing is characterized by its absence of either penalty or weighting.

**Correction**

The most extensively used formula (Cronbach, 1950): \[ S = R - \frac{W}{(n-1)} \], where \( S \) is the corrected score, \( R \) is the number of correct responses, \( W \) is the number of incorrect responses, and \( n \) is the number of alternative choices for each question. It is most often accompanied by instructions, at the time of testing, which explain the penalty imposed for incorrect guessing. The rationale for the use of this particular formula may be explained in the following manner. Since it is not possible to know which of the correct responses were the result of random guessing, we must infer this number from the fact that we view all incorrect responses as random, but alas unlucky, guesses. Therefore, the number of incorrect responses represents
(n-1)/n of the number of total guesses. With our example question about psychoanalytic psychology, n equals four and so if we are concerned with an examination made up of such questions, we would know that three quarters of the guesses would show up as incorrect responses and that the use of the formula would remove the remaining one quarter of the guesses from the number of correct responses. The examinee who chooses an incorrect alternative loses one third of a point. This testing scheme is of course based on the dichotomy of perfect knowledge and perfect ignorance (Cureton, 1966; Glass & Wiley, 1964) and therefore makes no attempt at differentiation of responses of classes 1B, 2B, 2D, and 3B. In other words, all errors are viewed as being the products of complete ignorance and therefore chosen randomly. Thus, we see a testing procedure featuring penalty without any weighting of responses.

**Probability Testing**

The use of the weighting solution is most clearly seen in the probability testing procedure. In general terms, the examinee is asked to assign weights indicating preference or degree of certainty for each of the alternatives on the multiple-choice questions (Rippey, 1968). This system may permit some analysis of the partial knowledge classes of response. The procedure used in this study was the Split Response Technique (SRT) (Stokes, 1966), which allows the examinee a "kitty" of points for each item to be distributed among the various alternatives. Following the
suggestion of Stokes (1966), a kitty of ten points was employed so as to facilitate administration and scoring. The probability testing approach uses weighting but no penalty so it may be compared to conventional and correction testing in terms of an increase in test relevance.

Confidence Weighting

The confidence weighting procedure is an attempt to eliminate guessing and measure levels of knowledge at the same time (Dressel & Schmid, 1953; Soderquist, 1936). The term "confidence weighting" refers to a testing arrangement in which the examinee is required to make two types of decisions. The first decision, as in conventional testing, concerns the choice of one correct alternative. The second decision involves the assignment of a number, representing a degree of confidence, to the chosen alternative. The amount of credit and penalty depends upon the examinee's confidence in his choices. The higher the confidence expressed, the greater the potential gain or loss. In the present study, the test-takers were allowed to assign up to five points to each of their choices with five being an expression of complete certainty. If the choice was correct the student receives five points but if the choice was not correct, a penalty of minus five points was assigned. In this procedure we see both penalty and weighting.

The four procedures may then be considered as the our testing conditions in a two-factor design as is shown in Table 2.
Table 2
List of Testing Procedures in Terms of Penalty and Weighting as Experimental Variables

<table>
<thead>
<tr>
<th>Weighting</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>NO</td>
<td>Conventional Testing</td>
</tr>
<tr>
<td>YES</td>
<td>Probability Testing</td>
</tr>
<tr>
<td>YES</td>
<td>Confidence Weighting</td>
</tr>
</tbody>
</table>
Validity Criteria: Relevance

Those examinations which accurately assess "target" traits and reflect little or no influence from irrelevant traits are said to be highly relevant. Risk-taking tendency has been mentioned, in the literature, as a possible source of systematic errors in test scores and as such may be seen as a possible threat to test relevance.

Risk-taking

Many studies have failed to either explicitly speak of risk-taking tendency, have remained within the realm of speculations about possible effects, or have employed limited and somewhat questionable instruments for measuring this tendency.

As a "personality variable". Some studies (Cronbach, 1946; Cureton, 1971; Ebel, 1965; Lyerly, 1951; Michael, 1968; Sherriffs & Boomer, 1954; Swineford & Miller, 1953; Votaw, 1936) have only mentioned the possible influence of "personality variables" but have implied that these variables are more properly labelled "risk-taking tendency". Using this imprecise term "personality variable", some studies claim that conventional and correction testing can not assess or reduce its influence on test scores (Cronbach, 1946; Cureton, 1971; Lyerly, 1951; Slakter, 1968ab; Swineford & Miller, 1953; Votaw, 1936). Indeed, some studies (Lyerly, 1951; Sherriffs & Boomer, 1954; Votaw, 1936) go so far as to suggest that the use of correction formulae introjects
personality variables into the scores. With regard to the probability testing and confidence weighting procedures, the literature is filled with conjecture and speculation and is generally lacking in experimental evidence. Shuford, Albert, and Massengill (1966) suggested that personality variables may be influential during the initial exposures to the probability testing procedures but claimed that this influence should decrease as the test-takers gained some familiarity with the systems. Ebel (1965) claimed that personality variables had contaminated his results but presented no data to support this claim. Michael (1968), in noting the "findings" of the Ebel (1965) study, recommended the use of a specific testing procedure to reduce these personality effects. Soderquist (1936), in his early work on confidence weighting, mentioned the possibility of the unwanted influence of the test-taker's personality. Wiley (1960) believed that the use of confidence weighting procedures favored the conscientious test-taker over the superficial and impulsive test-taker. One study (Coombs, Milholland, & Womer, 1956) suggested the need to investigate the influence of utility for risk but did not speculate about how this might be accomplished. Using a battery of personality tests which included the F-scale, dogmatism scale, rigidity scale, and various anxiety measures, Echtenacht, Boldt, and Sellman (1972) could find no replicable evidence for the influence of any underlying personality variables on probability testing scores.
As risk-taking tendency. Several studies have centered in on possible personality influences and have specified risk-taking tendency as the target of their investigations (Hansen, 1968, 1970; Jacobs, 1971; Krauft & Beggs, 1972; Swineford, 1938). The Choice Dilemmas Questionnaire (CDQ) (Kogan & Wallach, 1964) was used to measure risk-taking tendency in the two Hansen studies (1968, 1970) and the Krauft and Beggs study (1972). Hansen (1968, 1970) noted significant correlations between the scores on the Choice Dilemmas Questionnaire (CDQ) and test scores from a probability testing procedure. In disagreement with this finding are the results of the Krauft and Beggs (1972) study which found no significant correlations. Jacobs (1971) and Swineford (1938) used two related measures "confidence expression (CONF)" and "gambling tendency (G)". Both studies revealed the influence of risk-taking on confidence weighting test scores.

The existing experimental evidence seems to indicate that the penalty characteristic of confidence weighting and correction testing allows for the influence of risk-taking tendency whereas this tendency is not influential on non-penalty testing such as probability and conventional testing procedures. Thus a penalty by risk-taking interaction is predicted.

Validity of previous measures. Those studies which have scrutinized the role of risk-taking tendency may be assessed in terms of their theoretical implications only
after evaluating the measures used to uncover this tendency. It was beyond the purposes and scope of this study to attempt an evaluation of the test battery used by Echternacht, Boldt, and Sellman (1972). Furthermore, the utilization of such a battery could not be practically undertaken. There was, however, the possibility for evaluating single indices of risk-taking such as the CDQ. A pilot study was undertaken during the fall, 1974 semester. The CDQ was administered and the resulting scores were analyzed for possible relationships with scores from a confidence weighting procedure (Dressel & Schmid, 1953). There was no significant correlation between the test scores and the CDQ scores. This finding replicates a finding of Krauft and Beggs (1972) but does not provide evidence for the validity of the CDQ as a measure of risk-taking behavior. The relevancy of the CDQ was evaluated, in the main experiment, by comparing the CDQ scores with previously used measures of risk-taking tendency and a measure of risk-taking on objective examinations (RTOOE).

An alternative measure. Slakter (1967) proposed a new method for measuring risk-taking behavior on objective examinations. The method is based upon the manner in which examinees respond to "nonsense" questions that are imbedded within a group of legitimate questions. Legitimate questions are defined as questions that have one correct alternative and n-1 incorrect alternatives. A "nonsense" question is one that has no correct alternatives and no incorrect alter-
natives for a given population. Slakter (1967) stated that "risk-taking on objective examinations (RTOOE) can now be defined as guessing when the examinee is aware that there is a penalty for incorrect responses." The RTOOE measure has been found to be highly correlated with a guessing measure (Rz) developed by Ziller (1957). In order to compare the validity of the CDQ and RTOOE, both were administered to the students used in this study. It was predicted that the RTOOE would be more highly correlated with previously used measures of risk-taking, also derived from the testing behavior of the students, than was the CDQ.

Partial Knowledge

Weighting of alternatives has been proposed as a solution for the problem created by the inability of conventional testing scoring to represent different levels of knowledge. The value of weighting may be examined by comparing both the probability testing and confidence weighting procedures with conventional procedures but such a comparison would be trivial in nature in that it is conceded that conventional testing procedures do not represent any partial knowledge differentiations. A more interesting comparison would be to evaluate the effects of penalty within the context of weighting procedures. If the inclusion of penalty, in confidence weighting, causes the examinees to be more cautious in their responses and less willing to make complete certainty choices, then confidence weighting should reflect a greater number of partial know-
ledge responses than would probability testing. Beyond the question of whether tests exhibit degrees of partial knowledge, we may be interested in the degree to which tests can fully differentiate responses of the several classes previously outlined.

**Validity Criteria: Reliability**

The reliability of test scores derived from the four testing procedures provides a meaningful, but often misunderstood, method for evaluation and comparison. The degree to which an instrument can remain insulated from temporary and/or chance factors determines the reliability of that instrument. When considering multiple-choice questions, the effects of guessing behavior on test scores may be seen as a major threat to test reliability.

The results from studies dealing with correction testing seem to be equivocal but generally negative in that penalty tends to decrease test reliability. While it has been stated that "wild guessing" decreases test reliability (Michael, 1968), it has also been theorized that the use of a correction formula results in greater error variance which reveals itself as decreased reliability (Cureton, 1966; Frary, 1969; Glass & Wiley, 1964; Martois, 1972; Michael, 1968; Ruch & Stoddard, 1925). Four studies (Michael, 1968; Olagbade, 1971; Sherriffs & Boomer, 1954; Ruch & De Graff, 1926) reported equal reliability with conventional and correction testing.

The evidence concerning the reliabilities found when
using probability testing procedures is limited and must be approached with caution in that positive results are generally reported by the original authors and not in replication studies. Some replications (Ebel, 1968; Krauft & Beggs, 1973; Olagbade, 1971) report equal reliability with either probability testing, confidence weighting, or conventional testing. The one exception to this pattern of results was Michael (1968) who did find increased reliability when using the Split Response Technique (SRT).

Several studies have reported various degrees of improved reliability using confidence weighting when compared with either correction or conventional test scores (Dressel & Schmid, 1953; Ebel, 1961, 1965; Soderquist, 1936; Swineford, 1938). However, Swineford (1938) argued that this apparent increase in reliability was the result of a statistical artifact and warned against any close comparison of tests because of the differing configurations of questions answered by the test-takers.

The investigation and treatment of reliability in the previously mentioned studies are of questionable value in that improper statistical procedures were employed. The use of Kuder-Richardson formula 20 is not appropriate with scores that take on values other than one or zero. It follows that correction, confidence weighting, and probability testing should not be analyzed with this formula or any standard internal consistency formulae for reliability estimation.
With these restrictions on the interpretability of previous studies in mind, it was predicted that conventional testing would provide higher reliability estimates than would correction testing. It was also predicted that the weighting procedures (confidence weighting and probability testing) would provide higher reliability estimates than would the non-weighting procedures (conventional and correction testing).

Guessing Behavior

The imposition of penalty in both the correction and confidence weighting testing procedures is employed as a possible solution to the problem of guessing. The amount of suppression of guessing behavior directly relates to the reliability of the test. An evaluation of the effectiveness of this solution may be seen in a comparison of the four testing groups. The most obvious method for assessing the amount of guessing behavior would be to count the number of correct and omitted responses. The use of a penalty factor in testing has been shown to lower the average test score (Martois, 1972; Michael, 1968; Ruch & Stoddard, 1925; Sherriffs & Boomer, 1954). More specifically, the use of penalty has been shown to decrease the number of correct responses and increase the number of incorrect and omitted responses when compared with testing lacking the penalty factor (Quereshi, 1974). When examinees are encouraged to guess, the number of correct responses is increased (Cureton, 1966; Michael, Stewart, Douglas, & Rain-

In order to establish the effectiveness of penalty in reducing the number of guessing responses, it was predicted that penalty procedures (correction testing and confidence weighting) would exhibit less guessing than would non-penalty procedures (conventional and probability testing). It was expected that penalty procedures would produce a greater number of omitted responses and a lesser number of correct responses than would non-penalty procedures. The possible reduction in the number of correct responses is a desired result in that a number of such responses could be viewed as being the result of guessing behavior and as such would contribute to a diminished reliability estimate. With regard to the two weighting procedures, the experimental evidence is meager. Soderquist (1936) speculated that confidence weighting procedures should be effective in reducing the amount of guessing behavior. However, the results of a study by Olagbade (1970) show no reduction in guessing when confidence weighting is compared to probability testing. It was predicted that there would be a difference in the amount of guessing behavior with confidence weighting and probability testing procedures.

**Measuring guessing behavior.** While the number of omitted responses was seen as one means of measuring guessing behavior, it was understood that this method may only be acceptable when analyzing data from testing procedures having a penalty factor. Without this penalty factor,
there would be few, if any, omitted responses. Therefore the use of other measures became necessary. These other measures include Utility for Risk (Rz), Confidence Expression (CONF), and Gambling Tendency (G) which were previously introduced in this paper and will be further described in the Dependent Measures section. One measure was developed for use in the present study. The Average Confidence in Correct Responses (A) was calculated to determine if the imposition of penalty reduces confidence even in correct choices.

Additional Criteria

Time in Testing

An argument against the use of nonstandard multiple-choice examinations is that the systems are too complex, too expensive, and take too much time in testing (Ebel, 1968). Rippey (1968) in arguing against the use of the Admissible Probability Measurement Procedure (Shuford, Albert, & Massengill, 1966), claimed that increases in reliability were matched by increased time in testing. The amount of time required by each testing system may then be an important determinant in judging the value of using such systems. It was therefore hypothesized that the use of weighting, in confidence weighting and probability testing, takes more time than nonweighting procedures, as in conventional and correction testing. An alternative hypothesis is that the use of penalty causes hesitation in responding and therefore correction and confidence weighting
should take longer than conventional and probability testing procedures.

**Student Acceptance**

Stokes (1966) found that examinees learned to use the Split Response Technique (SRT) of probability testing with little effort and that they found the system to be less difficult, fairer, and more challenging than conventional testing procedures. However, Krauft and Beggs (1973) noted that students preferred conventional testing to SRT. In addition, those questioned stated that SRT was more difficult than conventional testing and caused them to do more guessing. It seems rather puzzling that they also responded that the SRT was fairer, permitted them to score higher, and better represented their knowledge than did conventional testing.

A questionnaire was completed by all examinees used in the pilot study. The same questions as those used by Krauft and Beggs (1973) were completed by all test-takers. A confidence weighting system was used instead of SRT. The data from this questionnaire showed 72% of the examinees preferring the conventional testing over the confidence weighting system. The one exception in preference was an even split on the question of better representation of knowledge.

It is possible that the level of student acceptance of any testing system may be determined by a combination of factors. The penalty factor may not be well received by
test-takers who are not well prepared by the examination. However, this tendency to dislike penalty may be more universal and it was predicted that the penalty procedures (correction testing and confidence weighting) would be less acceptable to students than would non-penalty procedures (conventional and probability testing). It is also possible that the risk-taking tendency of examinees may interact with the presence or absence of penalty conditions. It was predicted that the level of acceptance will depend upon risk-taking tendency for those students in penalty testing groups (correction testing and confidence weighting) but will not depend upon risk-taking tendency in non-penalty procedures (conventional and probability testing).

Dependent Measures

(1) Number of Correct Responses (R). The number of correct responses from the five quizzes was totalled for each student. The highest possible score of 100. With conventional testing, the marking of the correct alternative counted as one correct response. With correction testing, the tallying of correct responses preceded any adjustment for guessing. With probability testing, an inferred choice system was employed with a correct response defined as the assignment of the greatest number of points to the correct alternative. With confidence weighting, the selection of the correct alternative, regardless of the number of confidence points assigned, counted as one correct response.
(2) Number of Omitted Responses (O). For conventional and correction testing, the failure to choose any alternative counted as an omitted response. For probability testing, the failure to assign more than three (3) points to any alternative counted as an omitted response. For confidence weighting, the failure to assign points to the chosen alternative counted as an omitted response.

(3) Time in testing (T). The total time required by each student to complete the five quizzes was recorded.

(4) Utility for Risk (Rz) (Ziller, 1957).
\[ R = \frac{c(n/n-1)W}{(n/n-1)W+U}, \]
where \( n \) is the number of alternatives per question, \( W \) is the number of incorrect choices, and \( U \) is the number of omitted responses. This measure can be used only with conditions in which omissions are likely to occur. The formula provides a weighted proportion of the number of incorrect responses to the sum of incorrect responses and omitted responses.

(5) Gambling Tendency (G) (Swineford, 1938).
\[ G = \frac{W^*}{W + (n-1/n)O}, \]
where \( W^* \) is the number of incorrect responses marked with maximum confidence, \( W \) is the total number of incorrect responses, and \( O \) is the number of the omissions. This measure was used to analyze data from the probability testing and confidence weighting procedures.

(6) Confidence Expression (CONF) (Jacobs, 1971).
\[ CONF = \frac{W^*}{W}. \]
This formula is a simplification of the \( G \) measure. Jacobs (1971) eliminated the omissions term of the denominator. When there are few omissions, this simple
version of the Swineford G has little effect. As with the G measure, CONF was used with the probability testing and confidence weighting procedures.

(7) Average Confidence in Correct Responses (A). This measure was developed by the author for use in this study. The data from probability testing and confidence weighting were analyzed.

(8) Testing Procedure Questionnaire (TPQ). This questionnaire, assessing degree of acceptance of the testing procedures, was developed for use in this study and is included in the appendix. The answers were scored in such a manner so that the higher scores represent greater approval of the testing system.

(9) Choice Dilemmas Questionnaire (CDQ). This measure is derived from the responses of examinees to a series of twelve hypothetical stories. Each of these stories portray situations where some decision is required as to levels of risk-taking to accept. It was theorized that people tend to project themselves into the hypothetical situations and identify with the fictional characters (Kogan & Wallach, 1964). To the extent that this identification process reveals the risk-taking tendency of the examinees, the measure is a relevant one. However, we must consider the degree to which the Choice Dilemmas Questionnaire deviates from the realities of the classroom testing situation.

(10) Risk-taking on Objective Examinations (RTOOE).
The RTOOE score is a proportion of nonsense items attempted to the total number of nonsense items on the test. It was felt that the use of nonsense questions would preclude the possibility of responses from misinformation and isolate all nonsense attempts as pure guesses (Slakter, 1967). The vocabulary test developed to test RTOOE in this study is presented in the appendix.

**Hypotheses**

The following represent the specific problems that were investigated in this study. A listing of the measures used to test each hypothesis is presented within parentheses.

1. RTOOE is more highly correlated with previously used measures of risk-taking tendency than is CDQ. (RTOOE, CDQ, Rz, CONF, G, A)

2. The possible effects of levels of risk-taking tendency on testing behavior depends upon the presence or absence of penalty conditions. (R, O, T, CONF, G, A)

3. A greater number of partial knowledge responses will be emitted by students using confidence weighting than by students using probability testing.

4. There will be greater reliability of scores with conventional testing than with correction testing. (Hoyt's analysis of variance reliability estimate)

5. There will be greater reliability with weighting procedures (confidence weighting and probability testing) than with non-weighting procedures (conventional and cor-
rection testing) (same measure as hypothesis four).

(6) There will be less guessing with penalty procedure (correction and confidence weighting) than with non-penalty procedures (conventional and probability testing). (R, 0)

(7) Within a framework of weighting procedures, there will be a difference in the amount of guessing behavior with probability testing and confidence weighting. (R, 0, CONF, G).

(8) There will be more confidence expressed in the correct responses by those students using probability testing than by those students using confidence weighting. (A)

(9) There will be more time in testing for weighting procedures (probability testing and confidence weighting) than for non-weighting procedures (conventional and correction testing). (T)

(10) There will be more time in testing for penalty procedures (correction and confidence weighting) than for non-penalty procedures (conventional and probability testing). (T)

(11) There will be a greater acceptance of non-penalty procedures (conventional and probability testing) than of penalty procedures (correction and confidence weighting). (TPQ)

(12) The level of acceptance will depend upon risk-taking tendency for those students in penalty procedures (correction testing and confidence weighting) but will not depend upon risk-taking in non-penalty procedures. (TPQ)
EXPERIMENT

Method

Subjects

The Ss were 40 introductory psychology students from California State University, Northridge, whose participation fulfilled a course requirement. There were 26 females and 14 males.

Procedure

Measuring risk-taking tendency. On the first day of class, all Ss were given a vocabulary test which had been composed following the suggestions of Slakter (1967) and as such constituted a measure of risk-taking on objective examinations (RTOOE). A copy of this test has been included in the appendix. On the second day of class, all Ss were administered the Choice Dilemmas Questionnaire (CDQ) of Kogan and Wallach (1964). The Ss were not informed of the true purpose of either of these two tests until after completion of the experiment.

Group assignment. The four testing groups were matched in terms of academic ability. The SAT verbal, SAT math, and high school grade point average for each S were obtained from official school transcripts during the first week of class. The obtained scores and averages were combined with regression coefficients calculated during the pilot study as a means of predicting ability level with regard to an introductory psychology course. The prediction scores were then ranked and divided into ten clusters of four Ss.
The Ss from each cluster were randomly assigned to the four testing groups. It was explained to the students that their final course grades would be determined by absolute criteria rather than a curve. There would be no comparison of any student's scores with any other student either within or between testing groups for the purposes of grading and evaluation.

The testing groups. The four testing groups were as follows:

(1) Conventional testing (no penalty, no weighting). The students in this group were told that their scores would equal the number of correct responses with each choice receiving one point. There would be no penalty for incorrect guesses. There was no specific mention of the omission of responses. As with all of the testing groups there was no time limit set.

(2) Correction testing (penalty, no weighting). The test scores were determined using the formula: \( S = R - \frac{W}{(n-1)} \). The incorrect responses were penalized in accord with this traditional method. The computational process was explained but the students were not directly instructed about the advisibility of guessing. Any effect in terms of a reduction in guessing would therefore be the result of their understanding of the scoring system rather than as a response to explicit instructions.

(3) Probability testing (no penalty, weighting). The Split Response Technique (SRT) (Stokes, 1966) was used
with a kitty of ten points per item. The scores were then divided by ten to allow a maximum score of one point per question. A sample question was presented with various combinations of point assignment. The resulting scores from each assignment were fully explained. It was suggested that the scores would be maximized by a truthful representation of their knowledge in the assignment of points.

(4) Confidence weighting (penalty, weighting).
A modified form of the Dressel and Schmid (1953) testing procedure was used. The maximum assignment of points to the chosen alternative was five and the minimum was zero. The scores were then divided by five in order to provide a maximum score of +1 and a minimum score of -1 per item. As with the probability testing group, an example question was presented with all possible assignment of points to the correct and incorrect alternatives. The fact that the assignment of points to an incorrect alternative would cause a loss of points was stressed. The inadvisability of assigning five points when not sure of a correct response was emphasized. The possibility of a negative score on a quiz was examined to further stress the importance of a truthful display of confidence in knowledge.

Administering tests. During the second week of class, a practice test was administered in order to familiarize all Ss with the testing procedures. For the period of the next five weeks, one test was administered each week. The experiment was thus limited to five quizzes during the
first portion of the semester. This length of time was chosen in order to accomplish that necessary data analysis within a reasonable period of time. Although limited to five quizzes, this study represents an improvement over previously published studies, none of which exceeded an examination on two occasions. Each test contained twenty multiple-choice items with four alternatives for each item. The test questions were taken from a library of questions provided by the publisher of the textbook used in the class. The level of difficulty of the various questions appeared to be approximately equal and none of the questions contained "all of the above" or "none of the above" alternatives.

Testing procedure questionnaire. Following the fifth quiz of the experimental series, a questionnaire was completed by all Ss. This questionnaire measured attitudes toward the testing procedures. A copy of this questionnaire is included in the appendix. After completion of this questionnaire, the nature of the experiment was described and the students were then allowed to transfer to any of the four testing procedures for the remainder of the semester.

Results

A summary of the means of the dependent measures is presented in Table 3.

Relevance

Risk-taking tendency. In order to compare the relative merits of the CDQ and RTOOE as measures of risk-taking
Table 3
List of Means for Each Dependent Measure and Each Testing Procedure

<table>
<thead>
<tr>
<th>Measures</th>
<th>Testing Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td>71.40</td>
</tr>
<tr>
<td>Omit</td>
<td>0.0</td>
</tr>
<tr>
<td>Time</td>
<td>81.20</td>
</tr>
<tr>
<td>TPQ</td>
<td>20.60</td>
</tr>
<tr>
<td>CONF</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Rz</td>
<td></td>
</tr>
<tr>
<td>Rel. Coef.</td>
<td>.528</td>
</tr>
</tbody>
</table>
tendency, the scaling of the CDQ was reversed so that higher numbers represented greater risk-taking on both CDQ and RTOOE. A significant, negative correlation of -.27 (p<.05) between the students' CDQ and RTOOE scores suggested that two separate traits were measured. Neither CDQ nor RTOOE were significantly correlated with SAT verbal scores but both were significantly correlated (p<.01) with SAT math scores and high school grade point averages. The CDQ scores correlated negatively with SAT math scores (-.36) and with grade point averages (-.54). The RTOOE scores correlated positively with SAT math scores (.42) and grade point averages (.36). There were no significant sex differences for either CDQ or RTOOE. Hypothesis one predicted higher correlations of RTOOE with previous measures of risk-taking behavior than of CDQ with these measures. The Utility for Risk (Rz) measure correlated negatively with CDQ (-.11) and positively with RTOOE (.37). Using the data from the probability testing and confidence weighting groups, multiple correlations were calculated for both risk-taking measures with Rz, CONF, G, and A. The multiple correlations were .30 with CDQ and .59 with RTOOE. Therefore, since RTOOE was positively correlated with Rz and correlated more highly with Rz, CONF, G, and A than was CDQ, it was concluded that hypothesis one was supported and that RTOOE would be used in all subsequent analyses of this study.
Hypothesis two predicted an interaction between level of risk-taking tendency and the presence or absence of penalty conditions. To accomplish this analysis, the testing groups were subdivided on the basis of the RTOOE scores into high and low risk-takers. The analysis required two configurations in that different measures were available for individual testing groups. A 2 x 2 analysis of variance was performed using each of the six dependent measures of testing behavior from the probability testing and confidence weighting groups. The six measures were the number of correct responses, the number of omitted responses, time in testing, CONF, G, and A. The two independent variables were penalty (presence or absence) and risk-taking tendency (high or low) for each of the six analyses. Neither risk-taking as a main effect nor as an interactive agent with penalty were significant on any of the six analyses. Similarly, a 2 x 2 analysis of variance was performed on data from the conventional and correction testing groups. The dependent measures were number of correct responses, number of omitted responses, and time in testing. The independent variables were penalty and risk-taking. The analyses of number correct and time in testing showed no significant effects of either variable or interaction. The analysis of the number of omitted responses was complicated by the fact that no such responses occurred in the conventional testing group. Therefore, a separate analysis was performed for number of omitted responses using exact probability tables developed
by Mainland and Murray (1952). No significant association could be found between the number of omitted responses and level of risk-taking tendency. It was concluded that hypothesis two could not be supported as no interactive effect of risk-taking tendency and penalty could be found.

Partial knowledge. Hypothesis three predicted a greater number of partial knowledge responses with confidence weighting than with probability testing. Partial knowledge responses were defined as those responses receiving other than full or zero credit. Of the one thousand responses, confidence weighting and probability testing displayed 309 and 155 partial knowledge responses, respectively. This difference in the frequency of partial knowledge responses was significant, $\chi^2=65.96, p<.01$. Hypothesis three was supported. Both probability testing and confidence weighting allowed a differentiation of certain response classes in terms of the item scores. Those classes identifiable in both procedures were 1A, 1B, 2A+2C, 2B+2D. The responses of class 2A could not be distinguished from 2C and the same was the case with classes 2B and 2D. Confidence weighting clearly identified omissions whereas omissions were hidden within classes 2B and 2D with probability testing. The identification of omissions would be possible, in probability testing, only through examination of the answer sheets. Neither scoring procedure allowed for the identification of class 3 responses (random guesses). In terms of the frequency of various response classes, confidence weighting
produced only 86 responses of class 1B (complete misinform-
ation) while probability testing produced 216 such responses. In addition, confidence weighting produced approximately twice as many responses of class 2A+2C and class 2B+2D as did probability testing. The frequencies of the various classes are presented as percentages of the total number of responses in Table 4. When considering responses of classes 2B+2D and 2A+2C which are the partial knowledge responses, certain patterns of testing behavior appear. With confidence weighting, the distribution is strongly bimodal with peaks occurring at 3 and -3 with frequencies of 78 and 59 responses, respectively. There were few responses of 0, -1, +1, and -4 with frequencies of 17, 14, 10, and 13, respectively. With probability testing, the distribution was highly leptokurtic with a peak at 5 of 66 responses. There were very few responses of any other value but the distribution was symmetrical.

Reliability

Hypotheses four and five deal with the reliability coefficients obtained from the four testing procedures. Hoyt's (1941) analysis of variance approach for calculating internal reliability estimates was employed as the only appropriate measure for such a comparison. The average reliability coefficients for conventional, correction, probability testing and confidence weighting were .528, .505, .721, and .491, respectively. The coefficients for conventional, correction, confidence weighting are approxi-
Table 4
List of Frequencies of the Various Response Classes as Percentages of the Total Number of Responses

<table>
<thead>
<tr>
<th>Response Class</th>
<th>Testing Procedure</th>
<th>Probability Testing</th>
<th>Confidence Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td></td>
<td>62.9</td>
<td>58.8</td>
</tr>
<tr>
<td>1B</td>
<td></td>
<td>21.6</td>
<td>8.6</td>
</tr>
<tr>
<td>2B+2D</td>
<td></td>
<td>8.1</td>
<td>17.4</td>
</tr>
<tr>
<td>2A+2C</td>
<td></td>
<td>7.4</td>
<td>13.5</td>
</tr>
<tr>
<td>Omissions</td>
<td></td>
<td>*</td>
<td>1.7</td>
</tr>
</tbody>
</table>

* Identification of omissions requires examination of individual answer sheets.
mately equal. The small number of subjects (10) in each testing group prevented a rigorous test of any possible differences between the four groups. A test for differences between correlations showed no significant differences in reliability estimates. Therefore, hypotheses four and five were not supported by the data of the study. However, using the Spearman-Brown Prophecy formula, an Improvement Factor was calculated which allowed a comparison of each coefficient with that of probability testing. This Improvement Factor indicates the number of times a test would have to be increased in length in order to equal the reliability of probability testing. These Factors were 2.31, 2.53, and 2.67 for conventional, correction, and confidence weighting, respectively. The reliability coefficients and Improvement Factors for the four testing groups are presented in Table 5.

**Guessing behavior.** Hypothesis six predicted less guessing with penalty procedures (correction and confidence weighting) than with non-penalty procedures (conventional and probability testing). Using two available measures of guessing behavior, number correct and number omitted, the hypothesis predicted a decrease in correct responses and an increase in omitted responses with penalty procedures. Using the number of correct responses as the dependent measure, a 2 x 2 x 2 analysis of variance was performed. The independent variables were penalty, weighting and risk-taking. The mean number of correct responses were 71.43
Table 5
List of Reliability Coefficients and Improvement Factors for the Four Testing Procedures

<table>
<thead>
<tr>
<th>Testing Procedure</th>
<th>Reliability Measures</th>
<th>Average Reliability</th>
<th>Improvement Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td></td>
<td>.528</td>
<td>2.31</td>
</tr>
<tr>
<td>Correction</td>
<td></td>
<td>.505</td>
<td>2.53</td>
</tr>
<tr>
<td>Probability</td>
<td></td>
<td>.721</td>
<td>----</td>
</tr>
<tr>
<td>Confidence Weighting</td>
<td></td>
<td>.491</td>
<td>2.67</td>
</tr>
</tbody>
</table>
and 73.80 for the non-penalty and penalty groups, respectively. There were no significant main effects or interactions. In terms of number of correct responses, hypothesis was not supported. Using the same analysis of variance with number of omitted responses as the dependent variable, as presented in Table 6, the main effect for penalty was significant, \( F(1,32) = 8.32, p < .01 \). The mean number of omitted responses was .80 and 4.25 for the non-penalty and penalty groups, respectively. There was a significant penalty by weighting interaction, \( F(1,32) = 9.83, p < .01 \). The means for conventional, correction, probability testing, and confidence weighting were 0, 7.2, 1.3, and 1.6, respectively. In that conventional testing produced no omissions, the use of analysis of variance is somewhat questionable. Therefore, a test of the exact probabilities was performed using the Mainland and Murray (1952) tables. There was a significant association between the number of omissions and penalty conditions and there was a significant interaction between penalty and weighting \( (p < .05) \). Eliminating the conventional group from consideration, the differences between correction testing and probability testing and correction testing and confidence weighting were found to be significant \( (p < .05) \) using Duncan's Multiple Range Test. However, there was no difference between the probability testing and confidence weighting in the number of omitted responses. Hypothesis six was supported in terms of omissions but not in terms of correct items.
Table 6
Analysis of Variance on Mean Number of Omitted Responses

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (penalty)</td>
<td>119.025</td>
<td>1</td>
<td>119.025</td>
<td>8.32</td>
<td>.01</td>
</tr>
<tr>
<td>B (weighting)</td>
<td>46.225</td>
<td>1</td>
<td>46.225</td>
<td>3.23</td>
<td>N/S</td>
</tr>
<tr>
<td>C (risk-taking)</td>
<td>9.025</td>
<td>1</td>
<td>9.025</td>
<td>.63</td>
<td>N/S</td>
</tr>
<tr>
<td>AB (penalty x weighting)</td>
<td>140.625</td>
<td>1</td>
<td>140.625</td>
<td>9.83</td>
<td>.01</td>
</tr>
<tr>
<td>AC (penalty x risk-taking)</td>
<td>1.225</td>
<td>1</td>
<td>1.225</td>
<td>.09</td>
<td>N/S</td>
</tr>
<tr>
<td>BC (weighting x risk-taking)</td>
<td>4.225</td>
<td>1</td>
<td>4.225</td>
<td>.30</td>
<td>N/S</td>
</tr>
<tr>
<td>ABC (penalty x weighting x risk-taking)</td>
<td>15.625</td>
<td>1</td>
<td>15.625</td>
<td>1.09</td>
<td>N/S</td>
</tr>
<tr>
<td>Within (error)</td>
<td>458.016</td>
<td>32</td>
<td>14.313</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis seven predicted a difference in the amount of guessing with probability testing and confidence weighting. The four dependent measures available for this analysis of guessing behavior were the number correct, the number omitted, G, and CONF. A 2 x 2 analysis of variance was performed with penalty and risk-taking as the two independent variables with each of the four dependent measures. There were no significant differences among the groups in terms of number correct and number omitted. The mean CONF scores for confidence weighting and probability testing were 38.3 and 58.2, respectively. This difference was significant, \( F(1,16) = 4.60, p<.05 \). This analysis is presented in Table 7. The mean G scores for confidence weighting and probability testing were 37.10 and 56.10 and this difference was significant, \( F(1,16) = 4.54 \ p<.05 \). This analysis is presented in Table 8. The G and CONF data suggest less guessing with confidence weighting which has a penalty component than with probability testing which does not. Since the data from the number of correct and omitted responses failed to show a difference among the groups, support or nonsupport of hypothesis seven cannot be determined with the present data.

Hypothesis eight predicted a greater degree of Confidence Expressed in Correct Responses (A) for probability testing than for confidence weighting. A 2 x 2 analysis of variance had penalty and risk-taking as the independent variables. The mean A scores for probability testing and
### Table 7
Analysis of Variance on Mean CONF Scores

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (penalty)</td>
<td>1980.500</td>
<td>1</td>
<td>1980.500</td>
<td>4.60</td>
<td>.05</td>
</tr>
<tr>
<td>B (risk-taking)</td>
<td>1328.450</td>
<td>1</td>
<td>1328.450</td>
<td>3.09</td>
<td>N/S</td>
</tr>
<tr>
<td>AB (penalty x risk-taking)</td>
<td>174.050</td>
<td>1</td>
<td>174.050</td>
<td>.40</td>
<td>N/S</td>
</tr>
<tr>
<td>Within (error)</td>
<td>6887.200</td>
<td>16</td>
<td>430.450</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8
Analysis of Variance on Mean G Scores

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (penalty)</td>
<td>1805.000</td>
<td>1</td>
<td>1805.000</td>
<td>4.54</td>
<td>.05</td>
</tr>
<tr>
<td>B (risk-taking)</td>
<td>1584.200</td>
<td>1</td>
<td>1584.200</td>
<td>3.99</td>
<td>N/S</td>
</tr>
<tr>
<td>AB (penalty x risk-taking)</td>
<td>156.800</td>
<td>1</td>
<td>156.800</td>
<td>.39</td>
<td>N/S</td>
</tr>
<tr>
<td>Within (error)</td>
<td>6354.800</td>
<td>16</td>
<td>6354.800</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
confidence weighting were .969 and .897, respectively. This difference was significant, $F(1,16) = 16.02$, $p < .01$. The analysis is presented in Table 9. Hypothesis eight was supported.

Additional Criteria

Time in testing. Hypothesis nine predicted a greater time in testing for the weighting procedures (probability testing and confidence weighting) than for the non-weighting procedures (conventional testing and correction testing). Hypothesis ten predicted a greater time in testing for the penalty procedures (correction testing and confidence weighting) than for the non-penalty procedures (conventional and probability testing). A $2 \times 2 \times 2$ analysis of variance was performed with penalty, weighting, and risk-taking as the independent variables. The mean time in testing for the weighting groups was 84.4 minutes and the mean time for non-weighting groups was 84.5 minutes. The mean times for the penalty and non-penalty groups were 87.75 and 81.15 minutes, respectively. There were no significant main effects or interactions. Hypotheses nine and ten were not supported.

Student Acceptance. The degree of student acceptance of the testing procedures was measured by the TPQ. Hypothesis eleven predicted a greater acceptance of non-penalty testing than of penalty testing procedures. A $2 \times 2 \times 2$ analysis of variance was performed with penalty, weighting, and risk-taking as the independent variables. The mean
## Table 9
Analysis of Variance on Mean A Scores

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (penalty)</td>
<td>.026</td>
<td>1</td>
<td>.026</td>
<td>16.02</td>
<td>.01</td>
</tr>
<tr>
<td>B (risk-taking)</td>
<td>.004</td>
<td>1</td>
<td>.004</td>
<td>2.42</td>
<td>N/S</td>
</tr>
<tr>
<td>AB (penalty x risk-taking)</td>
<td>.001</td>
<td>1</td>
<td>.001</td>
<td>.31</td>
<td>N/S</td>
</tr>
<tr>
<td>Within (error)</td>
<td>.0256</td>
<td>16</td>
<td>.0016</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TPQ scores for the penalty and non-penalty groups were 14.30 and 21.15, respectively. This difference was significant, $F(1,32) = 51.00, p < .01$. The analysis is presented in Table 10. Hypothesis eleven was supported. Hypothesis twelve predicted a penalty by risk-taking interaction. The analysis failed to show any significant two-way interactions and therefore, hypothesis twelve was not supported. There was a significant penalty by weighting by risk-taking interaction, $F(1,32) = 5.025, p < .05$. The mean TPQ scores for low risk-takers in the conventional, probability testing, correction testing, and confidence weighting groups were 21.40, 20.40, 13.00, and 15.60, respectively. The mean TPQ scores for high risk-takers for the conventional, probability, correction, and confidence weighting groups were 19.80, 23.00, 15.20, 13.40, respectively. This interaction is illustrated in Figure 1.

Discussion

The purpose of this study was to evaluate modifications of the conventional multiple-choice question. Penalty and weighting were employed as the manipulated variables of a two-factor design. This experimental approach allowed the four testing procedures to be compared in terms of validity, practicality, and student acceptance. The question of validity assessment was examined by testing specific hypotheses proposed under the separate topics of relevance and reliability.

Validity: Relevance

Considered as relevance problems were the possible influence of risk-taking tendency on testing behavior and the differentiation of
Table 10
Analysis of Variance on Mean TPQ Scores

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Level of Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (penalty)</td>
<td>469.225</td>
<td>1</td>
<td>469.225</td>
<td>51.00</td>
<td>.01</td>
</tr>
<tr>
<td>B (weighting)</td>
<td>5.625</td>
<td>1</td>
<td>5.625</td>
<td>.61</td>
<td>N/S</td>
</tr>
<tr>
<td>C (risk-taking)</td>
<td>.625</td>
<td>1</td>
<td>.625</td>
<td>.07</td>
<td>N/S</td>
</tr>
<tr>
<td>AB (penalty x weighting)</td>
<td>1.225</td>
<td>1</td>
<td>1.225</td>
<td>.13</td>
<td>N/S</td>
</tr>
<tr>
<td>AC (penalty x risk-taking)</td>
<td>.625</td>
<td>1</td>
<td>.625</td>
<td>.07</td>
<td>N/S</td>
</tr>
<tr>
<td>BC (weighting x risk-taking)</td>
<td>.025</td>
<td>1</td>
<td>.025</td>
<td>.00</td>
<td>N/S</td>
</tr>
<tr>
<td>ABC (penalty x weighting x risk-taking)</td>
<td>46.225</td>
<td>1</td>
<td>46.225</td>
<td>5.025</td>
<td>.01</td>
</tr>
<tr>
<td>Within (error)</td>
<td>294.400</td>
<td>32</td>
<td>9.200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Mean TPQ Scores for low and high risk-takers in the four testing procedures.

Conventional testing  (Cnv)
Correction testing  (Cor)
Probability testing  (Prb)
Confidence weighting  (Cwg)
classes of response as reflected in test scoring.

Risk-taking tendency. Hypothesis one represented a critical attitude toward the Choice Dilemmas Questionnaire which had been used in a number of studies (Echternacht, Boldt, & Sellman, 1972; Hansen, 1968, 1970; Krauft & Beggs, 1972). It was thought that the hypothetical nature of the questionnaire would not evoke adequate involvement to accurately assess risk-taking. The use of the Risk-taking on Objective Examinations (RTOOE) procedure was seen as a solution to this problem. The decision to use RTOOE rather than CDQ resulted from a comparison of the correlations of these measures with previously used measures of guessing behavior (Rz, G, CONF) and the newly-developed measure A. While the decision to use RTOOE seemed strongly supported by the initial data analysis, subsequent considerations made the choice somewhat more tenuous. Closer inspection of the measures G and CONF suggested that these measures were assessing both responses resulting from guessing and misinformation. This problem not only affects the choice of RTOOE but also the value of G and CONF as guessing measures for hypothesis seven. Fortunately, the choice of RTOOE can still be justified in terms of correlations with Rz and A. The average correlations of CDQ with Rz and A were -.33 and .20, respectively, while the average correlations of RTOOE with the same measures were .33 and .42.

The failure to support hypothesis two which predicted an interactive effect of risk-taking tendency and penalty and the lack of any main effect of risk-taking on testing behavior contradicted the findings of a number of studies (Ebel, 1965; Hansen, 1968, 1970; Jacobs, 1971; Swineford, 1938). The findings of the present study were not as expected but are in agreement with the results found by Echternacht,
Boldt, and Sellman (1972) and Krauft and Beggs (1972). The present study extends their findings in that a different risk-taking measure is used (RTOOE).

Partial knowledge. The superiority of confidence weighting over probability testing in the number of partial knowledge responses display was as predicted by hypothesis three. However, the magnitude of this difference was greater than was expected. While the number of complete confidence correct responses (1A) was comparable for the two groups, the number of complete confidence errors (1B) with probability testing was approximately two and one half times than with confidence weighting. A possible explanation is that the students in this study were familiar with conventional testing, having experienced it throughout their school. Probability testing had the capacity to be converted into conventional testing by the assignment of full points to one alternative on all test items. Such responses occurred 84.5% of the time with probability testing as compared with 67.4% with confidence weighting. Thus, it seems reasonable to assume that students perceived probability testing as an unattractive opportunity to examine the quality of their knowledge. On the other hand, confidence weighting required careful consideration of confidence level in that poor choices brought heavy penalties. It is understandable, but unfortunate, that coercion in the form of penalty imposition was necessary to elicit this critical self-examination. On those occasions when the quality of knowledge was considered, the distributions of responses suggested a lack of sensitivity with regard to the smaller differences in confidence expression. With probability testing, there was a predominance of scores of 5 which suggests a large number of 5-5 splits. The
relatively rare occurrence of scores of 4 or 6 suggests that few of the test-takers were sensitive to the difference between a split of 5-5 and 6-4. There were a somewhat greater number of scores 7 and 3 but very few, if any, scores of 1, 2, 8, and 9. With confidence weighting, major peaks occurred at 3 and -3. The students did not fully utilize either system. An alternative explanation for this phenomenon would involve the relatively low difficulty level of the test questions. Perhaps more difficult questions would produce a higher usage of the weighting mechanisms of these systems.

Validity: Reliability

A comparison of the reliability of the four testing procedures was attempted using two separate strategies. The reliability coefficients were calculated using Hoyt's (1941) analysis of variance method which was the only appropriate measure available for this study. The obtained coefficients were used to calculate Improvement Factors for each of the procedures. Unfortunately, the small number of subjects precluded a rigorous test of the differences among the coefficients of the four groups. Therefore, hypotheses four and five proved to be untestable although the Improvement Factor did strongly suggest the superiority of the probability testing over either conventional, correction, or confidence weighting. The second method of comparison was an analysis of the amount of guessing behavior occurring with each of the testing groups. The rationale for this indirect approach was that guessing introduces temporary and/or chance factors into the test scores thereby decreasing the test reliability. Hypothesis six, as a test of the findings of many studies (Martois, 1972; Michael, 1968; Ruch & Stoddard, 1925; Sherriffs & Boomer, 1954), predicted a decrease in the number of
correct responses and an increase in the number of omitted responses with penalty procedures as compared to non-penalty procedures. The predicted testing patterns would imply greater reliability for the penalty procedures. There were no significant differences in the number of correct responses for the different groups. There was, however, a significant difference in the number of omitted responses. This finding suggests partial support for the position that penalty decreases guessing behavior. Hypothesis seven compared the testing behavior of students using probability testing and confidence weighting in terms of number of correct and omitted responses as well as G and CONF. While there were no significant differences in the number of correct or omitted responses, the G and CONF measures showed less guessing with confidence weighting. Obviously, the previously mentioned reservations concerning the value of G and CONF make conclusions difficult to reach. However, hypothesis eight adds support to the conclusion that there was less guessing behavior with confidence weighting. Such a conclusion may lead to the prediction of greater reliability for confidence weighting than for probability testing. We are then faced with conflicting conclusions. One possible explanation would be that while guessing behavior is an important determinant of reliability, there are other factors involved. Probability testing may have greater reliability than confidence weighting and a greater amount of guessing behavior as well.

Additional Criteria

Time in testing. The amount of time required to complete tests is a measure of practicality useful in comparing the procedures. The somewhat surprising result of no difference in time needed for the four
procedures demonstrates that nonconventional testing procedures are as easily administered as the conventional procedure. This conclusion is at variance with the anti-probability positions of Ebel (1968) and Rippey (1968a). Closer examination of the data suggests that while penalty did not make a significant difference with this data, future study with smaller within group variance may show that greater time is required by the penalty procedures.

Student acceptance. Conventional testing and probability testing find about equal acceptance with students whereas correction testing and confidence weighting find about equal non-acceptance. The imposition of penalty makes a difference while the use of weighting does not. The preference for conventional testing over confidence weighting replicates the findings of the pilot study. The equal acceptance of conventional and probability testing is in disagreement with the conclusions of Krauft and Beggs (1973) and Stokes (1966). The significant three-way interaction revealed a difference in preference for high and low risk-takers. The low risk-taker preferred conventional testing and found correction testing most objectionable. The high risk-taker preferred probability testing and disliked confidence weighting.

Some Conclusions

It is perhaps not possible to state with any level of assurance that one procedure is superior to the rest. It is possible to attempt some conclusions based upon the limitations of the present study and put forth some suggestions for future research. The evidence concerning the validity of the testing procedures does not present a clear picture. Certainly the fact that probability testing and confidence weighting reflect some amounts of partial knowledge leads one to accept
the notion that weighting procedures do increase test relevance. The superior display of partial knowledge responses in confidence weighting suggests that, within the framework of the present study, the highest relevance is achieved using confidence weighting. The data seem to suggest that probability testing procedures produce the highest reliability coefficients. The data on guessing behavior as well as the failure to directly test the reliability differences makes future study necessary. It is not possible to state whether probability testing or confidence weighting have the greatest test validity. The criterion of time in testing failed to discriminate between the four testing procedures. The data on student acceptance showed probability testing to be better received than confidence weighting. If a choice between confidence weighting and probability testing were left to students then probability testing would most assuredly be their choice. Although this criterion is of secondary importance to those of relevance and reliability, it is important to remember that the more acceptable a new development is, quicker will it be disseminated and utilized. Perhaps a modification of the probability testing system would be possible to insure a greater utilization of partial knowledge response alternatives and a greater suppression of guessing behavior.

**Future Studies**

The present study suggests a number of improvements necessary for future studies. The number of students used must be increased so as to properly test differences in reliability estimates. The use of a single classroom seems to be important but the use of matching procedures rather than simple randomization of subjects appears to be unwarranted. More research into the use of RTOOE would help to alleviate any doubts
as to its value as a measure of risk-taking tendency. The development of new measures of guessing behavior is particularly important with the discovery of the inadequacies of the G and CONF measures. Since most of the grading of quizzes and calculations of testing behavior measures were accomplished by hand and this necessarily limited both the length of the experiment and the size of the testing groups, it seems most appropriate to suggest the investigation of the use of computers is called for.

Contributions of the Present Study

The value of this study is not to be found in the number of definitive answers it provides but rather in the manner in which it approached the problem and in the questions it raises for future research. The following list identifies the several unique contributions made to the area of testing research:

1. The use of a factorial design.
2. The matching of subjects with regard to academic ability.
3. The use of five quizzes over a period of five weeks as opposed to previous use of only one or two testing sessions.
4. An examination of the Choice Dilemmas Questionnaire as a measure of risk-taking tendency.
5. The development of a usable RTOOE measure.
7. An examination of the appropriate use of reliability measures.
8. The development of a classification system for test responses.
References


Cureton, E. E. Reliability of multiple-choice tests is the proportion of variance which is true variance. Educational and Psychological Measurement, 1971, 31, 827-829.

Davis, F. B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests

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


Ebel, R. L. Confidence-weighting and test reliability. *Journal of Educational Measurement*, 1965, 2, 49-57. (a)


Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.


Rippey, R. A fortran program for scoring and analyzing probabilistic tests. Behavioral Science, 1968, 13, 424. (b)

Rippey, R. M. Rationale for confidence-scored multiple-choice tests. Psychological Reports, 1970, 27, 91-98. (a)

Rippey, R. M. A comparison of five different scoring functions for confidence tests. Journal of Educational
Measurement, 1970, 7, 165-170. (b)

Ruch, G. M., & DeGraff, M. H. Correction for chance and
"Guess" versus "Do not guess" instructions in multiple-
choice tests. Journal of Educational Psychology, 1926,
17, 368-375.

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

Samejima, F. A. A paradox in the knowledge of guessing
model for the multiple-choice item. Proceedings of the
Annual Convention of the American Psychological Assoc-
iation, 1972, 7, 61-62.

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

Shuford, E. H., Albert, A., & Massengill, H. E. Admissible
probability measurement procedure. Psychometrika,
1966, 31, 125-145.

Slakter, M. J. Risk-taking on objective examinations.

Slakter, M. J. The effect of guessing strategy on objective
test scores. Journal of Educational Measurement, 1968,
5, 217-221. (a)

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

Slakter, M. J., & Koehler, R. A. A new measure of risk
taking on objective examinations. Educational and


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

Swineford, F. Analysis of a personality trait. *Journal of Educational Psychology*, 1941, 438-444.


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

Appendix
Vocabulary Test

This test contains a list of twenty (20) difficult words. It is to your advantage to try to do as well as you can on this test. The students receiving the top 25% (10 people) will gain 10 bonus points toward their grade in this class. This number of points could be the difference between a grade of A or B in this class.

IMPORTANT - There is a penalty for guessing. Your score is determined by the number of correct minus the number wrong.

1. innate: a. reward b. fearless c. slave d. inherent e. insight

2. quintulent: a. feverish b. faltering c. acrid d. prurient e. decaying

3. covert: a secret b. possible c. asleep d. to scorn e. lawsuit

4. tamorin: a. shrew b. drum c. woodpecker d. urn e. associate

5. ardician: a. handyman b. suitor c. barber d. geologist e. undertaker

6. strident: a. harsh-sounding b. sleepy c. round d. cruel e. joking

7. palient: a. dim b. important c. sharp d. twinkling e. friendly

8. ventrescation: a. wound b. eloquence c. latticework d. window e. profanity

9. antipathy: a. assistant b. to warn c. dislike d. fortress e. convincing

10. cavil: a. to quibble b. aware c. deserter d. blame e. delay

11. hiln: a. cupboard b. handle c. meadow d. outhouse e. relation

12. slake: a. to allay b. winding c. filled up d. break e. pompous

13. rhustate: a. inflamed b. stopped up c. frustrated d. reddish e. ill-bred
14. garrulous: a. happy b. vicious c. soft d. smooth
e. talkative

15. brunnage: a. fog b. anger c. rigging d. darkness
e. effrontery

16. anneal: a. hatred b. self-confident c. praise
d. to toughen e. to enlarge

17. chilblain: a. inflammation from moist cold b. fortress
c. dark blue d. to trot e. slander

18. suscern: a. see b. dissociate c. suspect d. worry
e. repudiate

19. dolt: a. to frolic b. to stir up c. easy d. witty
   saying e. stupid person

20. walder a. meander b. lancer c. renege d. mason
e. mend

number correct _______
number wrong -- _______
SCORE _______
### Testing Procedure Questionnaire

<table>
<thead>
<tr>
<th>Question</th>
<th>very much</th>
<th>not at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How fair is your grading system?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. How difficult is your grading system?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. How well is your knowledge represented by the system?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. How much guessing did you do?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. How well do you like the grading system?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>