THE APPLICATION OF DECISION TIMES AND REACTION TIMES IN THE CONSTRUCTION OF LATENCY WEIGHTED TEST SCORES

by

Charles Philip Whitman

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:

R. B. Frary, Chairman

J. C. Arnold

E. S. Geller

J. C. Fortune

D. E. Hinkle

November, 1975

Blacksburg, Virginia

THE APPLICATION OF DECISION TIMES AND REACTION TIMES IN THE CONSTRUCTION OF LATENCY WEIGHTED TEST SCORES

by

Charles Philip Whitman

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:

_____
R. B. Frary, Chairman

_____ _____
J. C. Arnold E. S. Geller

_____ _____
J. C. Fortune D. E. Hinkle

November, 1975

Blacksburg, Virginia

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to the members of my committee. Their advice and assistance has been invaluable. I appreciate the cooperation of Scott Geller and the class of students who participated, and of Rosie Higdon who gave access to computer equipment necessary to conduct the study. I owe special thanks to Scott Geller for introducing me to the world of research. Finally, I want to thank my wife Cynthia for her patience with me during this and previous research projects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

THE PROBLEM

The general goal of the present research was to explore the use of response latencies to determine item weights in the construction of multiple-choice test scores. The theories which might lead one to suspect response latency to be an indicator of knowledge arose from prior laboratory studies which determined that confidence in a decision is negatively correlated with response latency. Knowledge about material tested by a particular item should be reflected in the degree of confidence in the correctness of the alternative selected. Thus, it might be feasible to construct several estimates of knowledge by weighting items by response latencies in various ways. The utility of such scores would be determined by their properties compared to traditional scores on the same test. Three specific objectives and research questions were as follows:

1. The theoretical basis for the use of latencies in score construction was that latencies are related to a student's confidence in the correctness of his choice. The basic research questions were: Given a valid measure of a student's confidence, can the relationships between other indicators of confidence and response latencies observed in

experimental situations be replicated in test-taking situations, and are the relationships strong enough to be of practical significance?

2. If response latencies were related to student confidence and correctness then the applied research question was: Can the latencies be used to determine item weights in the construction of test scores which yield estimates of knowledge at least as valid and reliable as those from conventional scoring procedures?

3. Given the successful construction of latency weighted scores, how do the these scores compare with those from conventional procedures with regard to bias due to personality factors?

An important aspect of these goals is the absence of direct attempts to estimate the effect of luck or correct scores for guessing. Many studies of multiple-choice test scores have considered factors of luck and have computed "corrected for guessing" scores by applying penalties for incorrect answers. Guessing was not the main object of the present research, but one implication of the confidence interpretation of the response latency weights was that answers given with little confidence were given smaller latency weights than other items.

CHAPTER II

LITERATURE REVIEW

The idea that response latency might be inversely related to strength of association can be traced to Clark Hull (1943). Thirty years later, psychologists continue to measure the latency between stimulus presentation and decision responses in order to study cognitive processing during learning, memory retrieval, information processing, and decision making. The proposed research is designed to expand the utility of latency measures by exploring the use of response latency as a predictor of a student's "true knowledge". A brief discussion of theory and findings in the areas of probability learning, choice reaction time, decision making, and paired associate learning will illustrate the presumed value of applying response latency measures to the study of multiple choice test taking behavior.

First it is instructive to distinguish between decision time and choice reaction time as dependent variables. Decision time refers to the time a subject spends making a choice when there is no mandatory response and subjects are not rushed. On the other hand, choice reaction time refers to the time a subject spends executing a response required to a particular stimulus as given to the subject at the

outset of the experiment (i.e., there is only one known correct response to a given stimulus alternative), and subjects are told to respond as fast as possible with a balance between speed and accuracy so that error rates range from 1 to 10 percent (Smith, 1968). Choice reaction times usually range from 0.5 to 1.5 seconds while the range for decision times is usually much greater (e.g., 0.5 to 25.0 seconds).

Probability Learning

When a subject is required to predict which of several stimuli will occur next in a random sequence of stimulus presentations, the frequency of prediction for each stimulus will approximate its frequency of occurrence (e.g., Estes, 1964). This phenomenon, called probability matching or probability learning, has been the subject of considerable research and mathematical modeling. The typical experimental procedure in the study of probability learning is: a signal initiates a trial, indicating that the experimenter is ready for the subject to predict which of several stimuli will occur next in the sequence; then the stimulus is presented to the subject a short time, and after a short time interval the signal indicates the start of the next trial. Some probability learning studies have studied decision latency as a dependent variable.

However, a major procedural problem with these probability learning paradigms has been a failure to adequately define and measure decision time. For example, Gerjoy, Gerjoy and Mathias (1964) presented a ready signal 9.2 seconds after the preceding stimulus to indicate that subjects should make a stimulus prediction. Decision latency was the time from the ready signal to the prediction. Unfortunately, subjects could have made their decision prior to the ready signal, thus making any interpretation of the results equivocal.

By using a self-paced probability learning task, Myers, Gambino, and Jones (1967) measured latency from the onset of the stimulus on one trial to the response for the following trial. The authors indicated that this measure included both decision time and reaction time (time to execute the response) but in fact it also included stimulus identification and processing time for the preceding stimulus. That is, when a stimulus is presented, some stimulus identification and information processing would necessarily occur prior to decision making or conflict resolution involved in determining a prediction. The important findings in this study were: 1) as subjects learned the stimulus probabilities, the decision latencies to predict the more frequent stimulus became shorter than latencies to predict the less frequent stimulus; and 2) decisions following correct stimulus predictions were made

faster than were decisions following incorrect predictions. While both of these facts could be interpreted as evidence that decision latencies are a meaningful correlate of learning, the findings do not adequately support such a conclusion because the latencies measured included the stimulus identification component of the choice reaction process. The results could be artifacts of stimulus probability and prediction outcome effects frequently observed in choice reation times.

Choice Reaction Time

In addition to latency to make a stimulus prediction, the latency to identify one of two stimuli in a probability learning paradigm has also been considered as a dependent variable in probability learning (eg., Geller, Whitman, Wrenn, and Shipley, 1971; Geller, Whitman, and Farris, 1972; Hinrichs, 1970). The general procedure followed in these studies was as follows: a subject predicted which of the two possible stimuli would occur next, a ready buzzer sounded, stimulus onset followed a .5 to 1.5 second variable interval, stimulus offset occurred with the subject's identification response. Subjects were encouraged to make one of two possible responses as quickly as possible. Choice reaction time was indicated by the duration of the

stimulus which was the interval between stimulus onset and response occurrence.

A common finding is that reaction time to correctly predicted stimuli is considerably shorter than reaction time to incorrectly predicted stimuli (eg., Geller et al., 1971, 1972; Hinrichs, 1970; Whitman and Geller, 1971,1972). Furthermore, reaction time is shorter to identify the more probable stimulus than to identify the less probable stimulus. This stimulus probability effect reaches an assymptote after 50 to 75 trials of experience with the probability schedule, suggesting that the probability effect is a function of learning. In the experiments reported by Geller et al. (1971, 1972) both probability learning and the stimulus frequency effect on choice reaction time indicated that subjects readily learned a reversal of stimulus probability from .70/.30 to .30/.70.

Expectancy constructs have been employed to explain the effect of stimulus frequency and prediction outcome on choice reaction time (cf. Geller, et al., 1971). Assuming that subjects expectancy for a particular stimulus is an increasing function of its probability of occurrence, subjects' preparation for a stimulus should be a direct function of its probability of occurrence. Likewise, the effect of prediction outcome on choice reaction time has also been explained by expectancy notions: a subject either predicts the expected stimulus or expects the predicted

stimulus with the result being more readiness to process and respond to correctly predicted stimuli than to incorrectly predicted stimuli.

Expectancy notions can differ in several ways which are difficult to contrast experimentally. For example, some investigators hypothesize that expectancy affects stimulus coding and processing time (stimulus anticipation) while other investigators favor the idea that expectancy influences readiness to respond (response anticipation). Another issue is whether expectancy for a given stimulus results in facilitation to react to that stimulus, inhibition to react to other stimuli, or both. Other questions studied by expectancy theorists involve the concern over the domain of the construct: is expectancy discrete, limited to two or a few distinct levels; or is expectancy continuous, varying over a range of certainty?

A continuous expectancy notion was adopted by Whitman and Geller (eg., 1971, 1972) to account for sequential effects of prediction outcome in choice reaction time. Latencies to identify stimuli were shorter when the preceding one or two predictions were correct than when previous prediction outcomes had been incorrect. It was hypothesized that an intermediate confidence mechanism could be responsible for these sequential effects of prediction outcome. Specifically, confidence was assumed to increase following correct predictions and decrease following

incorrect predictions. Expectancy for a predicted stimulus was thought to be a direct function of subjects' confidence in their prediction, where both confidence and expectancy were assumed to be continuous. Thus, according to this model, reaction time to a stimulus was inversely related to expectancy for that stimulus: shorter latencies occurred when subjects were confident and thus had a high degree of expectancy, and longer latencies occurred when subjects were less confident and had less expectancy for the predicted stimulus.

Since the existence of a confidence mechanism had been based on inferences several times removed from the observable reaction latencies, it was of interest to ask for trial by trial subjective estimates of confidence, and to determine the relationship between choice reaction time and subjects' reported confidence for predicted stimuli. Geller and Whitman (1973) observed that choice reaction time to correctly anticipated stimuli shortened as reported prediction confidence increased, a result predicted by the continuous expectancy notions. However, another expectancy hypothesis, that choice reaction time to incorrectly anticipated stimuli would lengthen as prediction confidence increased, was supported in only one of three experimental conditions studied. Thus, continuous expectancy notions best described the effects observed in choice reaction times to correctly predicted stimuli.

Two-Choice Decision Time

A confidence construct has also been used to account for sequential variations in decision latencies. For example, Geller and Pitz (1968) studied subjects' revisions of confidence in a traditional two-choice decision making paradigm in which event probabilities were dependent on which of two possible data generators was selected. The sequence of a subject's responses on each trial was: predict which of two chips (red or white) would be drawn from one of two bags, press a button to see which chip color occurred, push a toggle switch to indicate a decision concerning which bag had been sampled, set a pointer to indicate confidence, and return the decision switch and confidence pointer to neutral positions. Decision time, the latency between the second and third responses, was measured without the subject's knowledge. The study was primarily concerned with the Intertia Effect, or subjects' resistance to decrease decision confidence estimates in spite of disconfirming information. The changes in decision latencies were generally consistent with an expectancy hypothesis: subjects used less time to make a decision following confirming or predicted events and took more time to hypothesize which data generator had been used following disconfirming or nonpredicted events.

An earlier study of decision time (Crandall, Solomon, and Kellaway, 1955) manipulated event probabilities in a two choice task. The following events occurred on each trial: the experimenter chose one of ten decks of ten cards each, the subject was told how many of the ten cards were marked and how much money (zero, five, or twenty-five cents) they would win or lose if the top card was marked, the subject guessed whether or not it was marked, and the reinforcement was tabulated when the top card was marked. Subjects were never told whether a guess was right or wrong. The time subjects used to make a prediction decision was a direct function of uncertainty. That is, the shortest decision latencies occurred when the event frequency imbalance was greatest, and the longest decision latencies occurred when the two events were equally likely. These results are similar to those later observed in a probability learning paradigm (Myers et al., 1967), except that Crandall et al. (1955) did not analyze prediction latency as a function of the frequency of the predicted stimulus. However, prediction latencies were analyzed as a function of the reinforcement contingent on the top card being marked. Since the reinforcement was not contingent on the correctness of a subject's prediction, the observed effects on decision latency are not applicable to the testing situation where reinforcement is contingent on correct answers.

Another area in which response latency has been used to study decision making processes is same-different judgments on physical continua. The same-different paradigm usually involves many possibile stimulus alternatives, two of which are displayed on each trial. It has been established that subjects require more time to make a same-different judgment when the stimuli are more alike (Festinger, 1943). This result suggests that more difficult discriminations require more time and/or that conflict resolution takes longer when the stimuli are similar.

Another paradigm to study two-choice decision time is judgment of numerical inequality. Using the numerals 1 through 9, Moyer and Landauer (1967) required subjects to identify whether the left or right digit was larger. There are two particularly interesting results in the judgment of numerical inequality: 1) the greater the difference between the numbers, the shorter the latency to identify the larger digit, and 2) the greater the numerical difference between digits the fewer errors subjects made (ranging from seven percent for differences of one, to less than one percent for differences greater than four). What is suggested, according to Moyer and Landauer, is that similar mechanisms are involved in comparing the digits as in making same-different judgments of inequality on physical continua. Whether that conclusion is warranted or not, it is clear that in two-choice tasks easier decisions are made in less time.

Four-Choice Decision Time

Of more relevance to the multiple-choice testing situation are the studies which have considered decision time as a function of the number, complexity, and attractiveness of the alternatives in decision-making situations. Hendrick, Mills and Kiesler (1968) studied latencies in a four-choice decision making situation where male college students rated and chose neckties under the pretext of a consumer research study. Although the subject matter of the decision is somewhat whimsical, the experiment was designed and conducted in an ingenious way. Subjects rated each tie's degree of attractiveness prior to the actual decision-making. A subject's ratings determined the sets of ties that subject would decide among. Times between presentation of a set of ties and the subject's choice were measured for four decisions per subject. The design was a 2x2 repeated measure factorial with sets of four ties which differed on one dimension or on many dimensions, and which were either equally attractive or not. Based on the individuals prior ratings, the sets of ties had either four equally attractive ties or two equally attractive with two unattractive ties. The results of the necktie study were as follows: when decisions were complex because the ties varied on many dimensions, decision time was shorter for four equally attractive alternatives than for two equally

attractive ties paired with two unattractive ties. When decisions were relatively easy because the ties varied on one dimension, decision time was longer for four equal alternatives than for two equal with two inferior alternatives. The authors interpretation of these results were: 1) a decision among four equally complex alternatives took less time than a similar decision among simple alternatives because in the complex case subjects ceased comparing the alternatives and made an impulsive response, and 2) times to choose one of two equal alternatives after two had been eliminated were shorter when the alternatives varied on one dimension because the decision was easier and subjects did not give up as readily when faced with two complex alternatives compared to four complex alternatives.

Fortunately the necktie study was replicated and extended by Pollay (1970,a) in a decision making situation described as selecting a profitable research and development project from sets of four projects which varied on the dimensions of complexity and number of equally attractive alternatives. The results were very similar to those of Hendrick et al. (1968) that subjects took longer to choose among four alternatives when two were undesirable than when all four were equally desirable.

Paired Associate Learning

There have been many studies of recall latencies in verbal learning. The verbal learning tasks involve students memorizing a list of pairs of words, or consonant - vowal - consonant (CVC) trigrams. When one of the words from the list of pairs is presented as a stimulus, the subject is required to respond with the corresponding word. The present discussion will be limited primarily tc the findings reported by Judd and Glaser (1969). The experiments reported suggested that recall latencies in the acquisition phase of paired associate learning were not useful in studying learning. However, in the overlearning trails following the Trial of Last Error (TLE) reponse latencies were longer for more difficult items, difficulty being defined by the number of acquisition trials required to reach the TLE for those items. According to Judd and Glaser, latency on overlearning trials in a paired associate task is equivalent to choice reaction time, with latency decreasing over the ten post TLE trials due to practice. Of course Smith (1968) would not agree, because the task was not basically stimulus identification, there was some degree of response uncertainty, and there was insufficient practice.

Judd and Glaser (1969) suggest that computer assisted instruction (CAI) programs may take advantage of response latency as an indicator of how much additional practice a

student needs in order to learn a particular item or concept. A later report of similar findings (Judd and Glaser, 1970) was not so definite in its support of latency as a measure of strength of association, but it provided an interesting link to the previously discussed expectancy and confidence constructs with the following statement: "While these results do not rule out the possibilty that response latency is a function of associative strength, they do strongly suggest that at least some component of the latency is a function of the subject's confidence in the correctness of his response." (p. 20).

The notion of confidence appears in some form in many of the conceptual models of latency in studies of probability learning, reaction time, decision making, and paired associate learning. In general the confidence construct has been used to explain shorter response latencies when subjects expected a particular event by assuming that confidence levels influence degrees of expectancy which in turn determine both decision times and reaction times.

Another application of reaction time as an indicator of learning was studied by LaBerge (1974). LaBerge was studying information processing in reacting, particularly the process of encoding letters. In a reaction time study which spanned five days, subjects learned to recognize a set of four unfamiliar letters as fast as familiar ones. The act of

rapidly encoding a stimulus composed of letters was termed an "automatic perceptial process" and was one of two major components in an interesting theory of development of reading skills. The use of reaction time to guage human performance in reading subskills is an example of application of reaction time as an aid in estimation of knowledge.

Multiple-Choice Testing

The easiest scoring method for multiple choice tests is to count the correctly answered items. Such scores have been criticized because of the unknown bais due to good or bad luck at guessing. When a correction-for-guessing formula is used to score a test, students are advised not to guess unless they can eliminate at least one alternative. The usual correction-for-guessing formula counts the number right and subtracts the number wrong divided by n-1 where n is the number of alternatives. Thus, random guessing among all the alternatives would tend to be penalized while guessing after eliminating at least one alternative would tend to at least break even. There has been considerable debate over the appropriateness of such corrected scores (see Diamond and Evans, 1973, for a review).

An alternative multiple choice test taking procedure was introduced by Coombs (1953) and extensively studied by Coombs, Milholland, and Womer (1956). The Coombs mode test

works the opposite of the traditional method of marking the correct answer, in that students mark only those alternatives among the n alternatives that they can confidently eliminate. A score of one point is accumulated for each incorrect alternative correctly identified and a score of -(n-1) is added if the correct answer is incorrectly identified as a distractor. Thus, on a four choice item, the range of possible scores is negative three to three. The primary objective of using the Coombs mode is the elimination of guessing; instead of a chance contribution or loss to the score, the student has the opportunity to indicate partial knowledge by correctly identifying some of the distractors.

One problem with the Coombs mode score is that the degree of confidence required to eliminate an incorrect alternative may differ from one student to another. Coombs et al.(1956) called the student's required confidence level his standard of assurance. A criterion index for standard of assurance was the student's raw score minus a theoretical score which was defined as the number of correctly answered items with all incorrect alternatives eliminated plus the sum over the remaining items of the probability of guessing correctly among the alternatives not eliminated. Since this criterion can not be computed when only the Coombs mode responses are available, Coombs considered the number of correct alternatives eliminated as the basis of an estimator

of standard of assurance. Assuming that misinformation was not influencing students' responses, the standard of assurance should be an inverse function of the number of correct alternatives eliminated. The problem was that while the criterion was not related to ability, the estimator based on incorrect eliminations was negatively related to ability.

Several modified Coombs mode testing procedures have been studied. For example, Coombs et al. (1956) introduced an alternate form in which subjects were asked to rate three out of four of the choices (i.e., the distractors) from one to three indicating the order in which they would eliminate them. This method has the advantage of retaining the order of elimination which might provide insight useful in revising test items. Another variation of the Coombs response mode is to have students eliminate as many of the distractors as they can, and then to choose the correct answer from those remaining. Frary and Zimmerman (1970) reported a similar procedure, with grades being determined by the Coombs mode score. In Frary's doctoral dissertation the best guess answers were used alone and in conjunction with distractor eliminations to construct "guessing-free" scores, various scores with reduced guessing components, and traditional raw scores (Frary, 1968). Thus it was feasible to study the influence of partial information on score reliability and validity.

In a comprehensive review of test item and alternative weighting, Wang and Stanley (1971) discussed yet another approach to the problem of estimating partial information, namely, a response mode which requires students to indicate a subjective confidence in the correctness of each of the alternatives. A number of different scoring procedures can be used with this response mode. Of particular interest is the type of score called an "admissible probability measurement" in which the scoring method must maximize a student's score only when the reported confidences reflect the student's true confidences. To do this the confidences associated with every alternative must be included in the score. A broad class of such scoring procedures was developed by Shuford, Albert, and Massengill (1966) and was termed "reproducing scoring systems." Because of the complexity of computing such scores, Shuford et al. suggested a score which was based on the confidences reported for correct alternatives only. As discussed in the Wang and Stanley review, the problem with such a score is that to maximize his expected score, a student should report 100 percent confidence in the alternatives he was most sure were correct.

The confidence weighting and Coombs mode scores both fall into a general category of response-determined scoring because the alternative weights in confidence procedures and the item weights in Coombs mode scores are determined for

each individual by his own responses. This is in contrast to the classical weighting of items or alternatives where one set of weights is determined and applied to all individuals. A review of the historical development of weighting procedures (Stanley and Wang, 1970) indicates that classical weighting procedures offer little hope as improved scoring methods, but that some forms of response-determined weighting seem worth investigating.

CHAPTER III

DATA COLLECTION

In-Class Tests

The subject pool was a class of 136 undergraduate students enrolled in a sophomore course in personality theory. During the course four thirty-five item quizzes were administered, each of which represented about 15 percent of a student's grade. The in-class quizzes were taken in the modified Coombs mode. That is, students eliminated as many distractors as they could and indicated which of the remaining alternatives they believed was correct. The student's score on a quiz was the better of two letter grades, one based on the Coombs mode score and the other based on the number of correct choices (i.e., the raw score). The Coombs mode in-class quizzes served as practice for an experimental procedure for responding to test questions displayed on a Cathode Ray Tube (CRT) computer terminal. Students were motivated to take the computer administered parallel forms of the last two quizzes by at least two factors. They could improve on their in-class quiz grade for the corresponding quiz with either of the letter grades based on the Coombs mode or raw scores on the computer quiz. Students also recieved extra credit for participating in this and other experiments, which could add no more than four percent to their course grade. About

40 percent of the course grade was determined by the final exam which was a conventional multiple-choice exam scored on the basis of number right with no correction for guessing.

Personality Measures

As part of the in-class study of personality, students had the cpportunity to take four personality tests which measured six traits. The traits were: test taking anxiety, need for approval, manifest anxiety (drive), depression, tendency to lie, and internal-external control of reinforcement. A brief description of each of the tests follows.

The Questionnaire on Attitudes Toward Testing Situations (ATTS) was a 16 item subset of the 21 true-false items administered by Sarason (1958) in a study of learning performance and test taking anxiety. The test instructions ask students to indicate how they feel when taking course exams and intelligence tests. The questions ask about behaviors and physiological conditions that indicate state anxiety in testing situations. For example, the first question in Sarason's test and in the subset was: "While taking an important examination, I perspire a great deal."

The measure of need for approval was a true-false scale titled "Personal Reaction Inventory" (PRI) which was introduced in the literature as "The Marlowe-Crowne Social Desirability Scale" by Crowne and Marlowe (1960). The PRI

is a 33 item scale which measures a persons need for approval in their answers to questions which have a clear socially accepted answer. For example: "I am always courteous, even to people who are disagreeable.", or "I never resent being asked to return a favor."

The third test was an 88 item true-false instrument which included three scales developed from the Minnesota Multiphasic Personality Inventory. The test consisted of the Taylor Manifest Anxiety Scale (TMAS) (Taylor, 1953, 1956) with buffer items selected from the Depression and Lie scales of the MMPI. The TMAS asks about behaviors and physiological conditions that indicate trait anxiety presumably a level of anxiety that remains relatively stable over long periods of time and in various stimulus settings.

The last personality trait measured was the internal external control of reinforcement as indicated by the 29-item "Rotter Scale" (Rotter, 1966). The internal external scale (IE) indicates the extent to which a person believes that reinforcement depends on their behavior rather than luck. The IE construct has been popular in personality research involving subjective confidence, expectancy, decision making, risk taking, decision time, and reaction time. For more information, see reviews by Lefcourt (1966) or Joe (1971).

Computer Testing

A total of 48 students took at least one of the computer quizzes (41 took quiz 3 and 33 took quiz 4). Each student was tested individually in a procedure that required approximately 45 minutes. When a student came to take a quiz he was seated at a computer terminal and the instructions which appear in Appendix I were read to him. Five practice questions gave the student an opportunity to adjust to the procedure and the response mechanism which is illustrated in Figure 1. The questions were presented in upper case letters as in Appendix II. When a student finished the quiz, the Coombs mode and raw scores were computed for him and he was shown histograms of the computer quiz scores of students who had already taken that quiz.

The computer terminal was an IBM 3277 display console which is capable of the simultaneous presentation of over 1700 characters. The terminal was connected to an IBM 370 model 158 which was operating in the Virtual Machine environment. The Assembler language program which administered the quizzes executed under the Conversational Monitor System (CMS). Assembler language was used in order to have control over IBM 370 software and hardware features which support display of data at 3277 terminals with minimal system overhead and which allow access to the extremely accurate system clock.

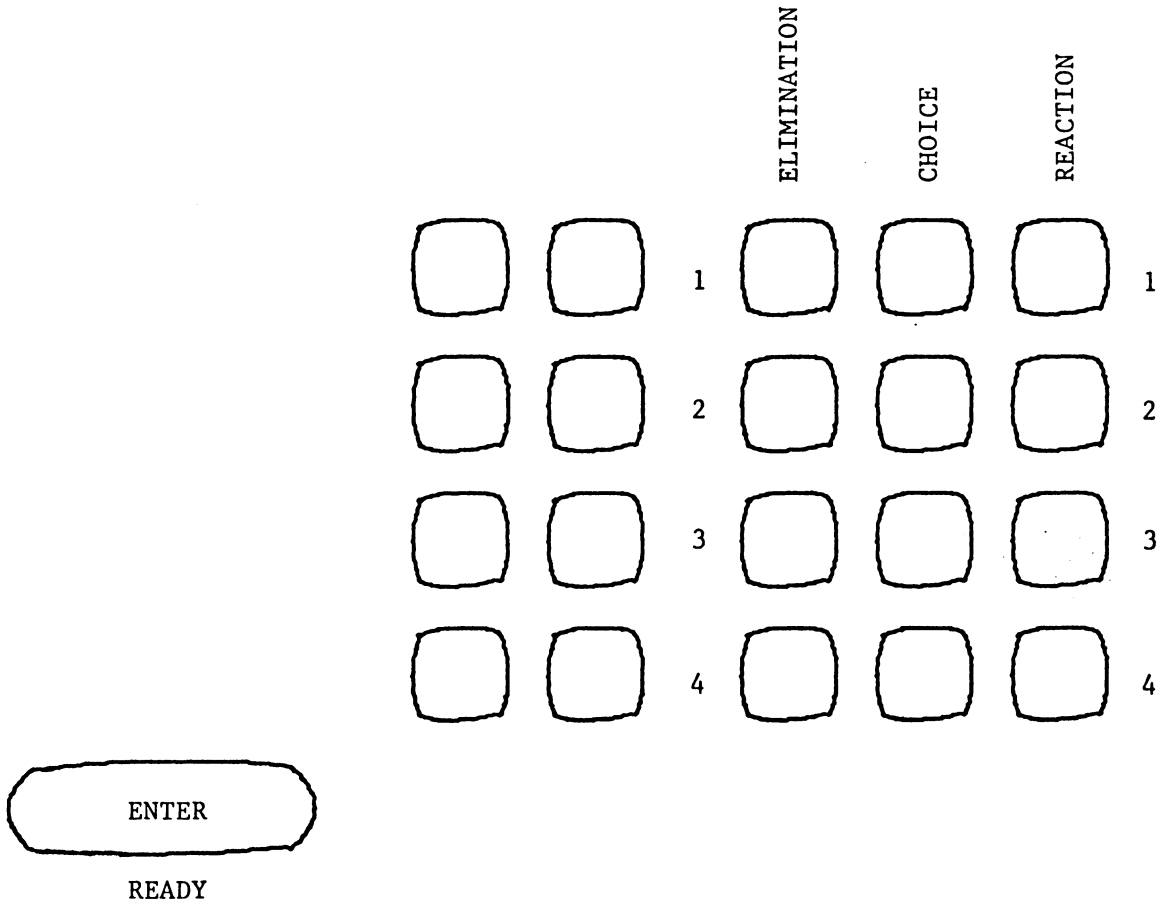

Figure 1. The Response Keyboard.

As indicated in the instructions, students were required to make several responses to answer each test item. Lists of the student responses and the time recorded by the computer appear in Figure 2. A question was started when the student pressed a "ready" key, and the entire stem of a question would appeared at once on the CRT. When the student pressed the "ready" key a second time the stem disappeared and the alternatives appeared. At this point the student pressed up to three keys in the "elimination" column. When the student had eliminated the distractors he was sure were wrong, he pressed a key in the "choice" column to indicate what he believed to be the correct answer. If the student decided he had made an incorrect elimination prior to pressing a choice key, he could press a key to restart that question. Once a choice had been made, there was no going back. Pressing a choice key caused the alternatives to disappear and a box of X's to appear around the center of the screen. The student then positioned his right forefinger over the column of "reaction" keys and pressed the "ready" key again causing the number of the correct answer to be displayed in the center of the screen. The choice reaction task was to press the corresponding reaction key as fast possible. The reaction response was the last event for that question and the student was free to press the "ready" key to continue.

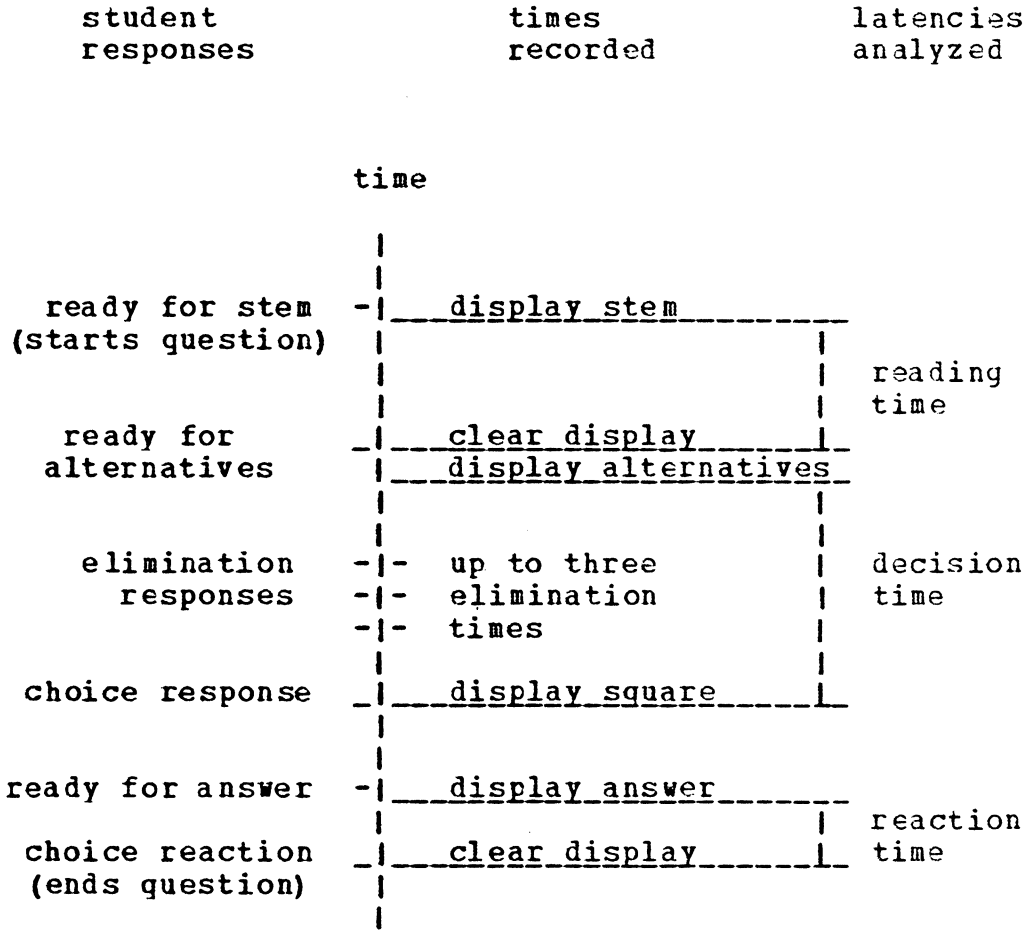

Figure 2. The Student's Responses and the Times Recorded.

Although the right side of Figure 2 indicates many times were recorded, three particular time intervals which proved to be of interest are illustrated in the left side of Figure 2. The reading time was the time the student used to process the stem of the question, and was defined as the time from the display of the stem until the response which indicated that the student was ready for the alternatives. When students made three eliminations, the first, second, and third elimination times were available, but were not comparable because the alternative information processing time was inseparable from the first elimination decision time. The decision interval of interest was the total decision time which included alternative information processing time, any elimination times, and the time from the last elimination to the choice response. Decision time was therefore defined as the time from the display of the alternatives to the pressing of a choice key. Finally, the choice reaction time was defined as the time from a display corresponding to the number of the correct answer until the pressing of the correct reaction key.

CHAPTER IV

DATA ANALYSIS AND RESULTS

Scores Based on Eliminations and Choices

An advantage of the modified Coombs' mode test procedure was the amount of information it provided. In addition to the Coombs' mode and raw scores that were used for grading purposes, two other scores previously studied by Lowry(1975) were computed. One was termed the students' "fair score" which was the number of items on which a correct choice was made from three or fewer alternatives. Another score, the students' "true ability", was computed as though only chance factors influenced a choice from the alternatives not eliminated. It was computed as the sum over the items which involved at least one elimination, of 1/A(i), where A(i) was one, two, or three, the number of alternatives not eliminated on the ith question. Lowry defined luck in the corrected for guessing response mode as the fair score minus the true ability score. The differences between the mean fair scores and the mean true ability scores for the in-class quizzes were 1.52, 0.97, 0.58, and 1.27. Table 1 persents the means and standard deviations of all the quiz scores.

Table 1. Means and Standard Deviations of Quiz Scores.

Score	Mean	Standard Deviation
In-Class Scores		
Quiz 1 N = 133		
Raw Score	25.4	4.5
Coombs Mode	60.3	17.9
True Ability	22.3	5.0
Fair Score	23.8	5.1
Quiz 2 N = 128		
Raw Score	23.5	5.1
Coombs Mode	52.3	20.3
True Ability	20.5	5.8
Fair Score	21.5	6.1
Quiz 3 N = 129		
Raw Score	22.3	4.8
Coombs Mode	49.9	19.5
True Ability	20.6	5.3
Fair Score	21.2	5.6
Quiz 4 N = 135		
Raw Score	24.5	4.9
Coombs Mode	60.1	20.7
True Ability	22.2	5.5
Fair Score	23.4	5.5
Exam N = 136	59.0	10.7
Computer quiz Scores		
Quiz 3 N = 41		
Raw Score	13.2	5.6
Coombs Mode	30.1	16.1
Quiz 4 N = 33		
Raw Score	10.9	3.3
Coombs Mode	19.8	14.6

A measure of the tendency to make eliminations, termed the student's standard of assurance, was defined by Coombs , Milholland, and Womer (1956) as the raw score minus a theoretical score which was the sum over all items of 1/A(i) where A(i) was the number of alternatives not eliminated on the ith question. Coombs et al. assumed that the chance component in the theoretical conventional score was equivalent to the chance component in the raw score. Hence, the standard of assurance measured the amount of information which influenced choices but was not sufficient to support eliminaticns. The standard of assurance means for each of the in-class quizzes were 2.79, 2.52, 1.41, and 2.03.

The intercorrelations of the raw scores and Coombs mode scores for the in-class quizzes, the computer quizzes and the conventional exam score appear in Table 2. Several observations can be made concerning the relative size of correlations among raw scores and Coombs mode scores. The highest correlations occurred between raw and Coombs mode scores for the same quizzes (.78, .86, .90, and .89 for the in-class quizzes and .92 and .88 for the computer quizzes). In 25 out of 30 pairs of score variables the correlation between a raw score and a Cocmbs mode score from different tests was less than or equal to the correlation between corresponding raw scores. Furthermore, the correlations between raw scores were consistently less than the correlations between corresponding Coombs mode scores.

Consistent with the above results, the correlations between the conventional exam and the quiz raw scores were in every case greater than the correlation between the exam and the corresponding Coombs mode score. Additional description of the raw and Coombs mode scores was reserved for use in the comparison of these scores with scores weighted by response latencies.

Table 2. Intercorrelations of Raw and Coombs Mode Scores.

Quiz	In-Class Quizzes Coombs Mode Scores 1	2	3	4	In-Class Quizzes Raw Scores 1	2	3	4
In-Class Coombs Mode Scores								
2	.69							
3	.66	.67						
4	.58	.57	.63					
In-Class Raw Scores								
1	.78	.61	.58	.51				
2	.61	.86	.65	.62	.61			
3	.61	.62	.90	.58	.63	.66		
4	.54	.59	.57	.89	.54	.66	.57	
Computer Coombs Mode Scores								
3	.61	.57	.50	.52	.57	.65	.49	.61
4	.59	.64	.59	.70	.38	.66	.54	.70
Computer Raw Scores								
3	.56	.53	.44	.46	.62	.62	.50	.62
4	.46	.53	.51	.51	.45	.57	.56	.61
Exam	.61	.66	.63	.60	.65	.69	.66	.69

Quiz	Computer Quiz Scores Coombs Mode 3	4	Raw 3	4
Computer Coombs Mode Scores				
4	.60			
Computer Raw Scores				
3	.92	.47		
4	.55	.88	.50	
Exam	.67	.68	.72	.70

Times as Dependent Variables

A demonstration that reading times, decision times or reaction times in the present testing environment were related to the number of elimination responses or to the correctness of the answer would indicate that some human performance measures might be useful in the estimation of knowledge. The usual approach to latency data in decision time and reaction time studies is to use within subjects variables to categorize hundreds of decision or choice reaction trials. The means of the latencies in the categories are taken for each subject as estimates of the true latencies for those conditions. Analysis of variance methods, usually repeated measures or mixed designs, are then used to evaluate the effects of the classification variables. A similar approach was followed in the analysis of the present data in spite of the fact that the computer quizzes had only twenty items.

Reading Time

The classification variables considered here were the number of alternatives eliminated and the student's choice (correct or incorrect). Hence, there were eight categories determined by zero, one, two, or three eliminations in conjunction with correct or incorrect choices. Each of the three latency measures was analyzed by taking mean latencies for each subject in the categories for which data was

available. Most students seemed somewhat confident in that they always made at least one elimination, while a few students appeared cautious in that they never made three eliminations.

Further loss of data occurred because some latency data was unacceptable. Reading times and decision times on items that a student restarted were discarded, as were those reaction times when students failed to press the correct key. In the analysis of reading times and decision times, the number of students contributing data to the zero and one elimination categories ranged from 23 to 30, while from 63 to 70 students contributed data to the two and three elimination categories. Because fewer than 20 percent of the students contributed to all eight categories, two by two repeated measures analyses of variance were computed for the two and three elimination categories using only those students who contributed data to those four categories. Since the sample analyzed was actually a subsample of the original volunteers, the analysis of variance results were interpreted as descriptive statistics rather than as hypothesis or significance tests.

The analysis of variance of reading times for 47 students revealed an absence of significant differences, all $F < 1$. Of course, these results reflect great within category variation in reading times relative to the mean differences between categories. The standard deviations

within the reading time categories were about one half the mean latency, while the ratios of standard deviations of decision times and reaction times to the corresponding means were about one third and one fourth respectively.

Decision Time

Overall averages of students' mean decision times for correct and incorrect answers are graphed in Figure 3 as a function of the number of eliminations. In general, the decision latencies were shorter when the choice was correct than when the choice made was incorrect. Decision latencies also were shorter when zero or three eliminations were made than when one or two eliminations were made. Again, only 47 students provided data in the four categories of the analysis. The squares and triangles in Figure 3 represent means of the original sample with the indicated N supplying data, while the plus marks indicate cell means of the analysis of variance. The fact that the mean latencies for the students in the analysis were consistently two to four seconds shorter than the plotted means indicated that the data analyzed was representative of the original sample. The results of the analysis were reliable main effects of number of eliminations $F(1,46) = 143.03$, and correctness of choice $F(1,46) = 17.39$, without interaction $F < 1$. The R-square values for the main effects were .66 and .10 respectively.

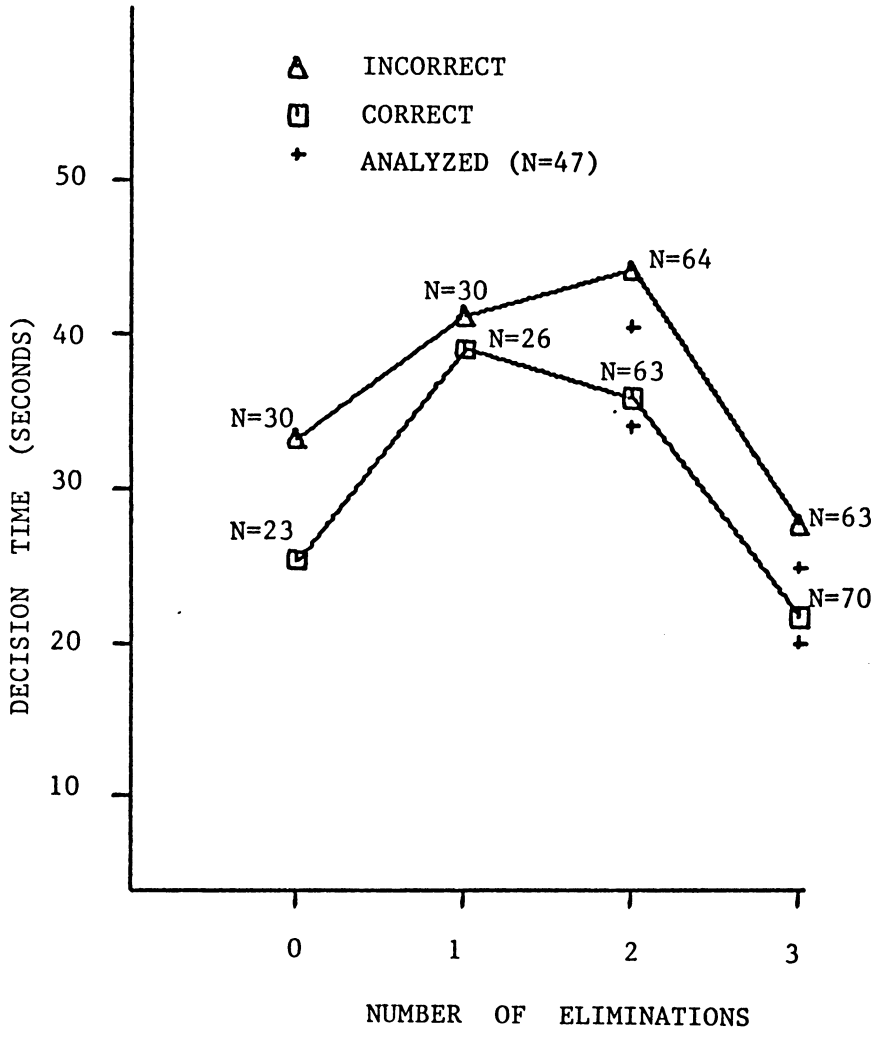

Figure 3. Mean Decision Time as a Function of the Number of Alternatives Eliminated and the Correctness of the Alternative Selected.

Reaction Time

A similar approach was taken in the analysis of choice reaction times. Figure 4 is a graph of reaction times to correctly and incorrectly anticipated answers as a function of the number of eliminations made. As the number of eliminations increased from one to three reaction time to correctly anticipated answers tended to be shorter, while reaction time to incorrectly anticipated answers increased considerably. The numbers of students with data in the eight categories were very similar to the numbers of students in the decision time analysis. As in Figure 3, the squares and triangles in Figure 4 represent the mean over all subjects with data available, while the plus marks represent the means over the 54 students who had data in the two and three elimination categories. The results of an analysis of variance were an interaction $F(1,53) = 7.12$, and an effect of correctness of choice $F(1,53) = 96.56$. The R-square for the interaction effect was .05 while the R-square for the main effect of correctness was .89.

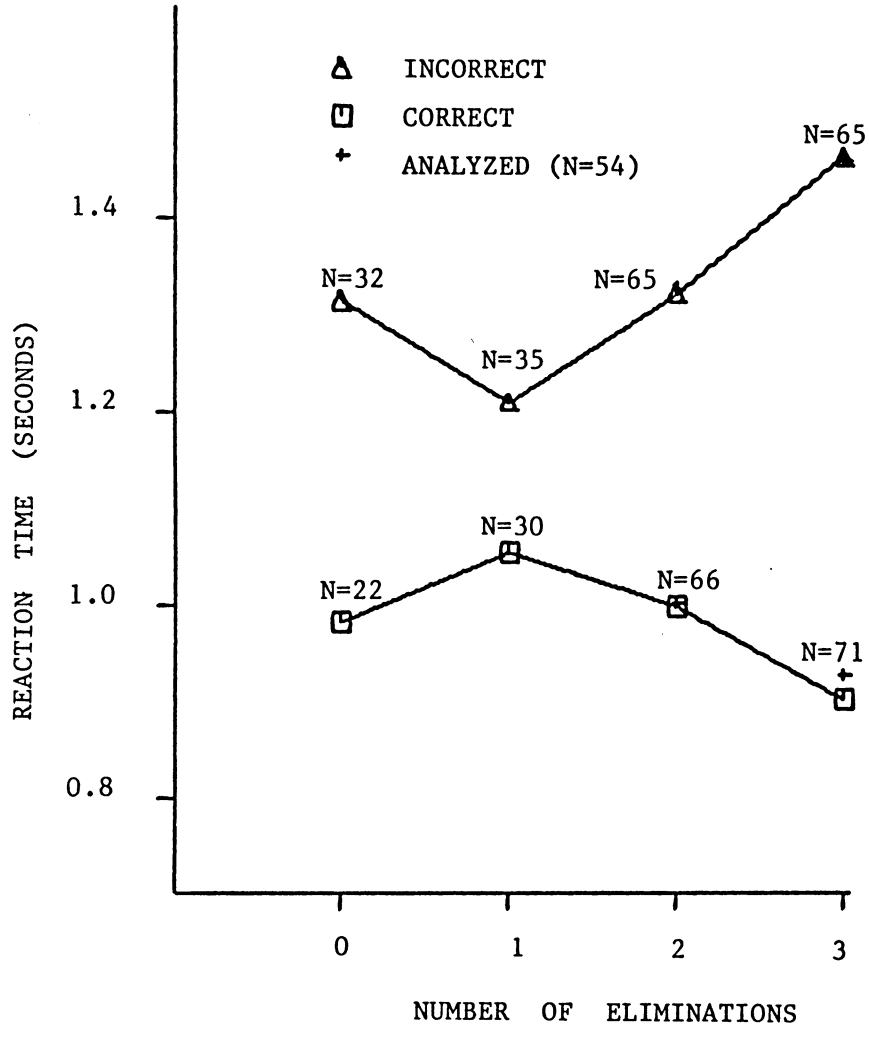

Figure 4. Mean Reaction Time as a Function of the Number of Alternatives Eliminated and the Correctness of the Alternative Selected.

Latency Weighted Scores

Four weighted score formulas were developed along two dimensions. While all scores included correct answers, two formulas added zero credit for incorrect answers, and two contrasting formulas subtracted penalties for incorrect answers. One each of these scores formed a pair which weighted each question by the number of alternatives eliminated on that question, and the remaining pair of scores did not weight questions by the number of eliminations. The latency weight was developed to avoid dependence on a student's absolute latency compared to other students' because the latency on a given question indicated confidence only in relation to other latencies for that student. Consequently the weight for a correctly answered item was one plus the subject's slowest time for any correctly answered question minus his time for that question. The weight for an incorrectly answered item, if a penalty was applied, was negative one plus the subject's shortest time for any incorrectly answered question minus his time for that question.

Exact descriptions of the latency weighted scores follow from these definitions:

let E(i) be one plus the number of eliminations made;

let C(i) be one if the answer was correct,

zero if the answer was incorrect;

let R(i) be defined for correctly answered questions as one plus the maximum latency when correct minus the latency for that item; and

let R(i) be defined for incorrectly answered questions as negative one plus the minimum latency when incorrect minus the latency for that item.

The four score types are then represented as the sum over the twenty questions of:

1) R(i) correct score with penalty for incorrect answers, unweighted by eliminations (CIUW).

2) C(i)R(i) correct score only, unweighted by eliminations (CUW).

3) E(i)R(i) correct score with penalty for incorrect answers, weighted by eliminations (CIW).

4) C(i)E(i)R(i) correct score only, weighted by eliminations (CW).

Reading times were not employed in score construction because there was no evidence of a relationship between reading time and the number of eliminations made or the correctness of the answer. Since each of the formulas was applied to both decision time data (DT prefix) and reaction time data (RT prefix), eight experimental scores were computed for each student.

Table 3. Means and Standard Deviations of Computer Quiz Scores.

Score	Mean	Standard Deviation
Computer quiz 3 N = 41		
Raw Score	13.2	5.6
Coombs MOde	30.1	16.1
Decision Time Scores		
DTCIUW	6.5	7.1
DTCUW	13.2	3.6
DTCIW	26.1	25.6
DTCW	46.8	15.4
Reaction Time Scores		
RTCIUW	13.1	14.1
RTCUW	21.8	11.1
RTCIW	50.9	52.4
RTCW	78.1	43.8
Computer Quiz 4 N = 33		
Raw Score	10.9	3.3
Coombs Mode	19.8	14.6
Decision Time Scores		
DTCIUW	1.5	6.2
DTCUW	10.7	3.1
DTCIW	9.6	20.6
DTCW	34.3	14.8
Reaction Time Scores		
RTCIUW	5.3	10.6
RTCUW	17.0	7.2
RTCIW	22.2	36.8
RTCW	54.4	29.9

The means and standard deviations of the computer quiz scores are presented in Table 3. Every one of the scores which penalized for incorrect answers, including the Coombs mode scores, had some persons with negative scores. The weighted scores were more variable than unweighted scores, with RTCIW having the greatest variance, followed by RTCW and DTCIW, with DTCW variances being comparable to Coombs mode score variances. A comparison of the score means and variances for the two computer quizzes revealed similar variances for corresponding scores and consistently higher scores on quiz 3 than on quiz 4. The size of the difference between quizzes was related to the variance of the scores, with the differences for RTCIW, RTCW, DTCIW, DTCW, and Coombs mode scores being 28.7, 23.7, 16.5, 12.5, and 10.3 respectively.

Cronbach's Alpha, an indicator of internal consistency that is a generalization of KR-20 for raw scores, was computed for each score for both the 35 item in-class quizzes and the 20 item computer quizzes. The Alpha values presented in Table 4 indicated that any of the item weighting schemes studied increased internal consistency compared to raw scores by increasing the variance of the test scores relative to the sum of the variances for the items. The largest Alpha value for a raw score was less than the smallest Alpha value for a Coombs mode score. The experimental scores had Alpha values consistently greater

than the Alpha of the raw score for that quiz and in the cases of RTCUW, RTCIW, and RTCW for quiz 3, and RTCW for quiz 4, the experimental scores Alpha values were greater than Alpha of the Coombs mode score for that quiz. RTCW had higher Alpha values than any other score for the same quiz, and quiz 3 RTCW has the highest Alpha value of any of the quiz scores, including all the in-class scores.

Table 4. Alpha Values for the In-Class and Computer Quiz Scores.

	20 Item Computer Quizzes		35 Item In-Class Quizzes			
	3	4	1	2	3	4
N	41	33	133	128	129	135
Raw	.714	.516	.737	.761	.742	.773
Coombs	.821	.783	.783	.803	.784	.838
DTCIUW	.718	.519				
DTCUW	.721	.518				
DTCIW	.747	.576				
DTCW	.790	.747				
RTCIUW	.807	.626				
RTCUW	.874	.722				
RTCIW	.823	.688				
RTCW	.884	.808				

The intercorrelations of the latency weighted scores appear in Table 5. For each quiz, the 12 correlations between pairs of scores based on the same response latency exceeded .91 with the exception of the quiz 4 correlation between RTCUW and RTCIUW which was .887. In every case, correlations between reaction time scores and decision time scores were higher for quiz 4 than the corresponding correlations for quiz 3. Correlations involving quiz 4 decision time scores and quiz 3 latency scores were consistently greater than the corresponding correlations between quiz 4 reaction time scores and quiz 3 latency scores. The correlations between quiz 3 and quiz 4 reaction time scores were smaller than any correlation between corresponding formula latency scores from quiz 3 and quiz 4. Correlations of decision latency scores between quizzes were generally greater than the correlation of raw scores between quizzes, and in about half the pairs they were greater than the correlation of Coombs mode scores between quizzes.

Table 5. Intercorrelations of Latency Weighted Scores.

Quiz 3	Decision Time Scores				Reaction Time Scores			
	CIUW	CW	CIW	CUW	CIUW	CW	CIW	CUW
Quiz 3 Scores N = 41								
DTCW	.932							
DTCIW	.987	.962						
DTCUW	.999	.933	.985					
RTCIUW	.729	.717	.731	.730				
RTCW	.513	.580	.537	.516	.943			
RTCIW	.704	.723	.723	.705	.993	.959		
RTCUW	.510	.517	.514	.232+	.953	.984	.954	
Quiz 4 Scores N = 26								
DTCIUW	.545	.628	.612	.544	.525	.584	.590	.501
DTCW	.446	.640	.532	.454	.492	.639	.576	.489
DTCIW	.610	.718	.672	.614	.640	.707	.693	.622
DTCUW	.538	.620	.604	.538	.518	.577	.582	.495
RTCIUW	.415	.453	.461	.415	.338+	.366+	.397	.301+
RTCW	.287+	.414	.341+	.295+	.257+	.350+	.321+	.234+
RTCIW	.494	.578	.544	.498	.472	.516	.522	.435
RTCUW	.229+	.224+	.247+	.232+	.117+	.124+	.158+	.087+

Quiz 4	Decision Time Scores				Reaction Time Scores		
	CIUW	CW	CIW	CUW	CIUW	CW	CIW
Quiz 4 Scores N = 33							
DTCW	.911						
DTCIW	.960	.913					
DTCUW	.999	.910	.960				
RTCIUW	.868	.777	.821	.861			
RTCW	.736	.789	.721	.729	.917		
RTCIW	.843	.812	.868	.837	.960	.935	
RTCUW	.686	.594	.625	.680	.944	.915	.887

\+ Correlation <u>not</u> significantly different from 0, $p>.05$.

The correlations between the latency weighted scores and the raw and Coombs mode scores are presented in Table 6. With the exception of the four correlations between quiz 3 latency weighted scores and computer quiz 4 Coombs mode scores, correlations between decision time scores and raw or Coombs mode scores were greater than corresponding correlations between reaction time scores and raw or Coombs mode scores. Correlations involving raw scores were generally higher than correlations involving Coombs mode scores. Latency weighted scores correlated with raw scores at least as well as with Coombs mode scores in 62 out of 96 comparisons. Of the 34 comparisons in which correlations with Coombs mode scores exceeded correlations with raw scores, 17 involved DTCW or RTCW scores; and 25 involved either one of the in-class quiz 1 scores, or computer quiz 3 latency scores with computer quiz 4 scores, or computer quiz 4 latency scores with computer quiz 3 scores. In order to simplify the evaluation of latency weighted scores, correlations from Tables 2 and 6 were assembled in Table 7.

Of particular concern were the correlations of the various scores with the final exam raw scores. In six out of eight pairs compared, correlations between decision latency scores and the exam were greater than the correlations of raw scores from the same computer quiz and the exam. In contrast, none of the corresponding reaction

Table 6. Correlations of Raw and Ccombs Mode Scores with Latency Weighted Scores.

Quiz 3	Decision Time Scores				Reaction Time Scores			
	CIUW	CW	CIW	CUW	CIUW	CW	CIW	CUW
In-Class Scores N > 40								
Raw 1	.62	.63	.64	.63	.45	.34	.46	.31
Coombs 1	.57	.67	.60	.58	.53	.50	.54	.44
Rawn 2	.62	.70	.66	.63	.52	.45	.54	.40
Coombs 2	.53	.65	.57	.53	.49	.45	.51	.38
Raw 3	.53	.53	.53	.54	.35	.26+	.34	.24+
Coombs 3	.46	.56	.50	.47	.32	.30+	.34	.23+
Raw 4	.62	.66	.65	.62	.36	.26+	.38	.21+
Coombs 4	.46	.56	.50	.46	.28	.23+	.30+	.16+
Exam	.72	.73	.74	.73	.48	.33	.48	.30+
Computer Scores N > 24								
Raw 3	.99	.93	.98	.99	.73	.51	.71	.52
Coombs 3	.92	.96	.94	.92	.69	.56	.70	.51
Raw 4	.51	.60	.58	.51	.49	.55	.56	.47
Coombs 4	.46	.66	.55	.47	.53	.68	.61	.53

Quiz 4	Decision Time Scores				Reaction Time Scores			
	CIUW	CW	CIW	CUW	CIUW	CW	CIW	CUW
In-Class Scores N > 30								
Raw 1	.49	.38	.50	.49	.44	.28	.43	.32+
Coombs 1	.50	.53	.57	.50	.46	.43	.53	.32+
Raw 2	.59	.61	.66	.59	.51	.49	.60	.39
Coombs 2	.54	.60	.64	.55	.47	.48	.59	.34+
Raw 3	.54	.56	.60	.54	.44	.43	.51	.32+
Coombs 3	.49	.59	.55	.50	.40	.44	.48	.25+
Raw 4	.64	.67	.70	.65	.52	.46	.59	.35
Coombs 4	.53	.67	.59	.54	.46	.48	.54	.28+
Exam	.71	.65	.79	.71	.61	.44	.69	.38
Computer Scores N > 24								
Raw 3	.53	.45	.61	.53	.46	.30+	.51	.23+
Coombs 3	.57	.59	.65	.56	.50	.48	.62	.32+
Raw 4	.99	.91	.96	.99	.87	.75	.85	.69
Coombs 4	.88	.95	.93	.88	.77	.76	.84	.57

+ Correlation <u>not</u> significantly different from 0, $p > .05$.

Table 7. Summary of Correlations Compared in Evaluating Latency Weighted SCores.

Score	Exam	same quiz Raw	same quiz Coombs	corresponding quiz Raw	corresponding quiz Coombs
In-Class Scores	N > 128				
1 Raw	.65				
1 Coombs	.61	.78			
2 Raw	.69				
2 Coombs	.66	.86			
3 Raw	.66				
3 Coombs	.63	.90			
4 Raw	.69				
4 Coombs	.60	.89			
Computer Quiz 3	N = 41				
Raw	.72			.50	.44
Coombs	.67	.92		.49	.50
DTCIUW	.72	.99	.92	.53	.46
DTCW	.73	.93	.96	.53	.56
DTCIW	.74	.98	.94	.53	.50
DTCUW	.73	.99	.92	.54	.47
RTCIUW	.48	.73	.69	.35	.32
RTCW	.33	.51	.56	.26+	.30+
RTCIW	.48	.71	.70	.34	.34
RTCUW	.30+	.52	.51	.24+	.23+
Computer Quiz 4	N = 33				
Raw	.70			.61	.51
Coombs	.68	.88		.70	.70
DTCIUW	.71	.99	.88	.64	.53
DTCW	.65	.91	.95	.67	.67
DTCIW	.79	.96	.93	.70	.59
DTCUW	.71	.99	.88	.65	.54
RTCIUW	.61	.87	.77	.52	.46
RTCW	.44	.75	.76	.46	.48
RTCIW	.69	.85	.84	.59	.54
RTCUW	.38	.69	.57	.35	.28+

\+ Correlation not significantly different from 0, $p > .05$.

time scores correlated with exam scores as well as raw scores. However, quiz 4 RTCIW correlated .69 with the exam, which was nearly as high as the .70 correlation of the computer quiz 4 raw score with the exam. Furthermore, with the exception of quiz 4 DTCW, decision time scores had higher correlations with exam scores than any of the Coombs mode or raw scores.

Correlations of latency weighted scores with raw and Coombs mode scores from the same quiz were also useful in evaluating latency weighted scores. As evident in Table 7, raw scores correlated with the Coombs mode scores on the same quiz from .78 to .90 and the computer quizzes were similar with correlations of .92 and .88. With the exception of the CW scores the latency weighted scores correlated better with corresponding raw scores than with corresponding Coombs mode scores.

Another set of relationships which served to evaluate weighted scores were correlations of latency weighted scores with scores from the corresponding in-class quiz. The right hand columns of Table 7 summarize the result that quiz 3 decision time scores correlated with in-class quiz 3 scores better than raw or Coombs mode scores. This result was less consistent for quiz 4, with only DTCIW matching the .70 correlation between Coombs mode scores. Again, reaction time scores did not compare very favorably.

Personality Variables

The means and standard deviations of the six personality scales are presented in Table 8, and the intercorrelations involving personality scales appear in Table 9. The Taylor Manifest Anxiety Scale (TMAS) was related to all of the personality variables except the Personal Reaction Inventory (PRI). The Attitude Toward Testing Situations (ATTS) was positively related to all the personality variables except the PRI and the MMPI Lie scale items. It was observed that the ATTS score was slightly negatively correlated with most test scores and that the MMPI Lie items were positively correlated with computer quiz 3 raw, Coombs mode, and decision time scores. With the exception of TMAS and IE scores, correlations involving reaction time scores were smaller in absolute value than correlations involving similar decision time scores. The correlations between the TMAS and test scores was the most variable: r was between -.12 and .05 for in-class scores, -.32 and -.14 for computer quiz 3, and between .19 and .34 for computer quiz 4. Most other relationships between test scores and personality variables were more consistent across in-class and computer quiz scores.

Table 8. Means and Standard Deviations of Personality Scales. (N = 132)

Scale	Mean	Standard Deviation
ATTS	6.2	3.0
PRI	10.9	5.5
TMAS	19.6	8.2
LIE	10.7	1.5
DEP	15.4	3.3
IE	11.4	4.0

Table 9. Intercorrelations of Personality Variables and Test Scores.

	Personality Variable					
	ATTS	PRI	TMAS	LIE	DEP	IE
Personality Variables N > 125						
PRI	-.12					
TMAS	.52**	-.33**				
LIE	.09	-.55**	.14			
DEP	.40**	-.22**	.49**	.16		
IE	.28**	-.11	.30**	.05	.24**	
In-Class Scores N > 80						
Raw 1	-.35**	-.10	-.12	.09	-.15	-.21
Coombs 1	-.36**	-.03	.01	-.00	.02	-.17
Raw 2	-.33**	-.04	-.10	-.04	-.16	-.17
Coombs 2	-.33**	-.05	-.06	-.12	-.10	-.18
Raw 3	-.23*	-.09	.05	.15	-.07	-.00
Coombs 3	-.26*	-.05	.05	.09	-.08	-.13
Raw 4	-.34**	-.09	.02	.02	-.06	-.23*
Coombs 4	-.40**	-.11	-.00	.08	-.06	-.21
Exam	-.41**	-.10	-.12	.09	-.15	-.21
Computer Quiz 3 N = 26						
Raw	-.44*	-.12	-.21	.40*	-.22	-.10
Coombs	-.48*	-.19	-.14	.47*	-.25	-.06
DTCIUW	-.43*	-.15	-.19	.42*	-.19	-.08
DTCW	-.50*	-.16	-.20	.43*	-.21	-.09
DTCIW	-.46*	-.18	-.19	.45*	-.20	-.08
DTCUW	-.43*	-.15	-.19	.41*	-.20	-.09
RTCIUW	-.28	-.02	-.32	.09	-.19	.09
RTCW	-.23	-.01	-.31	-.01	-.15	.11
RTCIW	-.27	-.03	-.31	.09	-.17	.09
RTCUW	-.19	.01	-.31	-.05	-.15	.13
Computer Quiz 4 N = 23						
Raw	-.12	-.33	.24	.41*	.10	.14
Coombs	-.26	-.24	.20	.33	.10	-.02
DTCIUW	-.13	-.30	.22	.39	.09	.12
DTCW	-.21	-.19	.21	.30	.08	.01
DTCIW	-.23	-.31	.19	.34	.09	.05
DTCUW	-.14	-.29	.22	.38	.10	.11
RTCIUW	-.09	-.21	.30	.35	.02	.17
RTCW	-.13	-.10	.33	.25	-.04	.17
RTCIW	-.22	-.23	.28	.34	-.04	.15
RTCUW	.01	-.10	.34	.24	.01	.22

* $p < .05$
** $p < .0025$

The number of correlations in Table 9 suggested a simpler summary of the relationships via a multivariate approach. Two different approaches were made to the analysis, one being a canonical correlation between two sets of variables, and the other being a univariate correlation of factor scores from independent factor analyses of the two sets of variables. The personality scores were correlated with five sets of test scores: the nine in-class scores including the exam were one set, the other four sets were the four decision time or the four reaction time scores from either quiz 3 or quiz 4. The canonical correlation coefficient can be considered as the the maximum univariate correlation over all possible linear combinations of the variables in each group. Thus, the maximum correlation between the set of personality variables and each set of test scores was computed: r was .64 for the in-class tests, .81 for quiz 3 reaction time scores, .70 for quiz 3 decision time scores, .78 for quiz 4 reaction time scores, and .75 for quiz 4 decision time scores.

The factor analysis approach was intended to determine sets of factor scores (which are linear combinations of variables) which were more meaningful than the linear combinations derived in the canonical correlation analysis. Two principal components analyses were done using the scores of 72 students who had taken all the in-class tests and given their student numbers on all the personality scales.

Factors with eigenvalues greater than one were retained but not rotated. The analysis of the in-class scores yielded one factor which accounted for 70 percent of the variance and factor loadings which ranged from .78 to .87. In contrast, the analysis of personality scales yielded two factors which accounted for 63 percent of the variance. The first factor may be identified with the PRI because it loaded positively on PRI and negatively on all the other variables. The second factor loaded positively on PRI, ATTS, and IE, and negatively on the MMPI Lie items. The correlations of the in-class test factor scores and the two personality factors were .19 and -.22 which were not significantly different from 0, $p > .06$. A similar approach for computer quiz scores proved unsuccessful because of the small number of students for which both computer quiz scores and personality measures were available. The small sample caused factor loadings to be unstable.

CHAPTER V

DISCUSSION

Scores Based on Eliminations and Choices

The relationships between raw and Coombs mode scores showed: 1) that raw scores generally had smaller values for both Cronbach Alpha values and intercorrelations with in-class scores, and 2) that the correlation between raw and Coombs mode scores on the same test increased as the students practiced taking quizzes in the modified Coombs mode. The most heavily criticized aspects of raw scores have been the possible biasing factors of luck in guessing, skill in multiple choice test taking, and bias due to personality factors. Unfortunately, the Coombs mode scores had similar correlations with the test-taking anxiety measure and there is no way to assess guessing effects in either raw or Coombs mode scores.

One problem addressed by Coombs et al. (1956) was the variation between equally knowledgeable students in making eliminations. The Coombs et al. criterion definition of standard of assurance was identical to the definition of luck in the free-guessing mode suggested by Lowry (1975), that is, a raw score or conventional score minus a true or theoretical conventional score. The reasons for this paradox were: 1) that the Coombs approach minimized the

factor that students might have luck different from chance and emphasized the notion that the difference between raw and theoretical raw scores was due to information not sufficiently trusted to warrant eliminations; and 2) that Lowry emphasized the influence of luck to increase or decrease the difference. It would appear that the difference between raw and theoretical raw scores observed in the present study and by Coombs et al. (1956), Frary (1968), and Lowry (1975), includes both luck and standard of assurance factors. Neither raw scores or Coombs mode scores can really separate the effect of luck from the effect of information which was not indicated by eliminations.

Response Times

Three latencies were studied as a function of the correctness of answers and the number of alternatives eliminated. The intervals considered were the reading time to encode the stem of the question, the decision time to encode the alternatives and select an answer, and the choice reaction time to identify the number of the correct alternative. Each of these times included time for at least two processes: information processing and response execution. Only the decision time and reaction time measures included an additional choice or response selection component. The absence of reliable differences between

categories of reading times, and the presence of differences between categories of decision times and reaction times, suggests that the independent variables influenced primarily the decision or choice component of the process.

The present research extended the study of response times from laboratory studies of simplified tasks to classroom oriented studies of realistic complex tasks. Demonstration of typical laboratory results in the test-taking environment suggests consideration of theories supported by experimental findings. Both decision time and reaction time results in the present study replicated experimental studies which have been interpreted in terms of learning, information processing, conflict resolution, expectancy, and subjective confidence. The response latencies will have applied value only to the extent that they represent these aspects of test-taking behavior.

In the results of decision time analysis two independent effects were observed: the correctness of the choice made and the number of alternatives eliminated. The observation that decision time to make correct choices was shorter than to make incorrect choices was consistent with the observation of shorter times on easier than on difficult decisions in laboratory studies (cf., Crandall, Soloman, and Kellaway, 1955; Festinger, 1943; Moyer and Landauer, 1967). In these experimental tasks, easier decisions were defined by the amount of information available to reduce uncertainty

and conflict resolution. In the laboratory experiments, information came from two sources: the question or stimulus that was presented, and what the subject had learned about the stimulus or task. In the testing situation information also came from both the question and what the student had learned.

The other main effect in the decision time means was the inverted "U" shape of decision time as a function of the number of alternatives eliminated. This result was consistent with those of Hendrick, Mills and Kiesler (1968) and Pollay (1970) who observed longer decision latencies when two out of four alternatives could be eliminated than when none of the alternatives could be eliminated. One interpretation was that decision latencies were shorter when conflict resolution was low because the student had a lot of information (indicated by three eliminations) or when the student who had no information (indicated by zero eliminations) gave up and guessed. Decision latencies were longer when conflict resolution was high because the student had some information but not enough to know the answer (indicated by one or two eliminations). Thus, given the number of alternatives eliminated, one could apply decision time as an additional indicator of the amount of information the student had.

The students' mean reaction times to the number of the correct alternative conformed surprisingly well with the

hypothesized results suggested by confidence constructs (cf., Geller and Whitman, 1973, Whitman and Geller, 1971, 1972). Reaction times were shorter when the student was correct than when he was incorrect, and the difference between reaction time when incorrect and reaction time when correct increased as the number of alternatives eliminated increased. Assuming that a student's confidence in his answer was reflected in the number of alternatives eliminated, then the reaction time results are readily explained in terms of confidence and continuous expectancy constructs. As confidence in a choice increased, expectancy for the anticipated stimulus increased both facilitation of responses to that stimulus and inhibition of responses to other stimuli.

Several studies have supported confidence constructs as mechanisms which could account for the effects of predictions and stimulus frequecy on decision times (Geller and Pitz, 1968), and reaction times (e.g., Geller and Whitman, 1973). Confidence has also been identified as a determinant of of response latency in paired associate learning tasks (Judd and Glaser, 1970). Thus, it has been well established that high levels of subject confidence were associated with shorter decision and reaction latencies in several different experimental settings.

Latency Weighted Scores

The application of response latencies as weights for test items was based on the interpretation of differences in latencies between items as measures of relative confidence in the alternatives selected. The desired effect of item weighting by confidence was to emphasize items which were answered with near certainty while minimizing weight for items which involved much uncertainty. Latency weights were formulated independently for correct and incorrect answers. Ignoring sign, the minimum latency weight was one, and weights increased with shorter latencies on correct responses and with longer latencies on incorrect responses. The latency weight was a continuous measure of the student's confidence in his choice on that question relative to other questions. Both the latency weight and the number of eliminations can be interpreted as indicators of decision confidence. However, the latency weight contains more information because it is a continuous variable, while the number of alternatives eliminated is a discrete variable which has only four possible values.

There were two considerations in the design of the scores. One factor was to penalize (or not) incorrect choices and the other was to weight (or not) by the number of alternatives eliminated. The raw score is an example of no penalty for errors and no weighting according to

eliminations. The Coombs mode score is an example of both penalty and weighting. Four scores were constructed with decision time weights and four with reaction time weights so as to include all combinations of penalty for incorrect answers and weight for alternatives eliminated.

Decision time scores and reaction time scores had considerably different properties. For example, while the Cronbach Alpha values for the latency scores were higher than for the raw scores for the same quiz; the values for reaction time scores were always higher than for decision time scores and RTCW scores had the highest Alpha values for both computer quizzes.

Latency scores in general correlated better with the raw scores than with the Coombs mode scores, and decision time scores generally had higher correlations with the raw and the Coombs mode scores than did reaction time scores. Decision time scores were more like corresponding raw scores than were corresponding Coombs mode scores, and all the decision time scores except quiz 4 DTCW had higher correlations with the exam score than any of the Coombs mode or raw scores. Thus, there was considerably more support for decision time scores than for reaction time scores in terms of correlations with the exam scores.

Personality Scores

Of the patterns apparent in the correlation of personality scores with the test scores, the most obvious was the negative correlation between test-taking anxiety and performance on in-class test scores and computer quiz 3 raw, Coombs mode, and decision time scores. Persons who were anxious in test-taking situations did not score as well as those who did not report anxiety when taking tests. The higher correlations between test-taking anxiety and test scores for computer quiz 3 than for computer quiz 4 might have been due to the novelty or uncertainty of the situation during computer quiz 3. Recall that most of the students who took computer quiz 4 had taken computer quiz 3.

Reaction time scores had generally lower correlations with the personality scores than the Ccombs mode, raw, or decision time scores for the same quiz. The reaction time scores consistently had smaller correlations with the ATTS measure than the in-class scores. Unfortunately, the reversal between computer quizzes in the relationship between TMAS and all the computer quiz scores cast further doubt on the reliability and interpretation of the results involving personality scales. However, if one had to select the score which minimized bias due to personality factors, one might choose the score which had the minimum sum over the personality scores of the squared correlations.

However, the same score did not have the minimum sum of squared correlations for both quizzes. For computer quiz 3 RTCUW and RTCW had the smallest sums, while for computer quiz 4 DTCW and RTCW had the smallest sums. Thus, the criterion of minimizing the squared correlations between test scores and personality scales suggests use of RTCW scores.

Summary

In terms of the three research questions presented in Chapter I, the results of the present research were encouraging. With regard to the first objective, the results observed when decision time and reaction time were considered as dependent variables supported the interpretation of response latencies in terms of confidence constructs. Thus, there was established an adequate theoretical basis for the use of response latencies in constructing item weights. In addition, the differences between latencies on correctly versus incorrectly answered items appeared to be sufficiently robust to be of practical significance.

In the construction of latency weighted test scores, the confidence constructs determined the form of the weights as differences between latencies. The resulting scores had high reliability estimates, particularly the reaction time

scores; and high predictive validity estimates, particularly in high correlations between decision time scores and raw scores or the exam.

The third research question concerned the influence of personality factors on test scores. While the Coombs mode and raw scores were negatively correlated with a measure of test-taking anxiety, reaction time scores for both computer quizzes and decision time sccres for computer quiz 4 were not highly correlated with any of the personality measures.

Recommendations for Test Scoring

The observations and results of the present study indicated that there was little advantage in using the Coombs mode in evaluating students in the classroom. Neither raw scores nor Coombs mode scores circumvents the problems of guessing and personality effects. Thus the recommendation of raw scores for everyday classroom testing parallels that of Lowry (1975): 1) to use raw scores because they are adequate for class use, 2) the tests are easier to administer, score and interpret, and 3) are consequently more efficient.

Latency weighted scores offered an alternative measure of individual performance. The decision latency scores were too much like raw scores to be worth using: they correlated very strongly with the raw scores, and they had similar

correlations with test-taking anxiety, especially on the student's first computer quiz experience. The reaction time scores may be more useful in that their correlations with raw scores from the same quizzes were nearly were nearly as high as correlations between corresponding Coombs' mode and raw scores, but the personality variables were relatively unrelated to the reaction time scores. Furthermore, the Cronbach Alpha values were always greater for the reaction time scores than for raw scores. In particular, the RTCW scores compared favorably with other scores as they had the highest Cronbach alpha values, and the next-to-lowest sum of squared correlations with personality variables. Intuitively, the notion of not penalizing incorrect answers and weighting correct answers by both the number of alternatives eliminated and the relative reaction time is appealing.

The idea of weighting test items by confidence is not new. Shufford et al. (1966) presented a brief review of why confidence in each alternative provides more information about the student's knowledge than either raw or Coombs mode scores. The disadvantage of the reproducing scoring systems was the complexity of the rules which determined the optimum response given the information available. The advantage of the reaction time scores was that the confidence indicators were natural and required little conscious effort. It is possible that reaction time scores were a truer indication

of knowledge and confidence than were subjective confidence estimates. Further research may develop applications of latency weighted scores in very sensative testing situations where it is critical to know how well the student knows the material.

Another application of response latency involves the continuous evaluation of student progress. In computer assisted instruction (CAI) environments the clock and scoring programs are built into the hardware and software of the system. The student's response time could accurately indicate how much additional practice was necessary to develop a certain degree of proficiency with the constructs being studied. The present results supported the prediciton made by Judd and Glasser (1969) that CAI programs could take advantage of response latencies as an indicator of learning.

The concept of computers administering tests has recieved some attention in the form of Taylor aptitude testing. In such tests there are pools of test items of varying difficulty and a computer selects items according to performance on and difficulty of preceding items. Items are selected in an attempt to converge on the true ability of the examinee. One result of the item selection process is to increase the effective length of the test. Perhaps an application of response latency weights could aid in item selection by indicating the examinee's self assurance.

Recommendations for Futher Research

As in any experimental wcrk, there are scme aspects of the present study which would best be changed if a replication were to be conducted. First and perhaps most inportant, the taking of quizzes administered by the computer should be mandatory. This requirement would have the effect of reducing sample bias by cutting the volunteer subject effect and increasing the sample size. Furthermore, one aspect of the procedure should be changed so that the entire question, stem and alternatives, would be displayed at once because reading times proved to be uninteresting and restarts of questions were a nuisance factor in the data.

In the construction of latency weighted scores there are several recommended changes. In the present analysis, the weights were determined for each student using the maximum time over correct choices and the minimum time over incorrect choices as reference points to determine the relative latency weight. Given a second opportunity some alternative reference points might be considered, perhaps median times. Another possibility, particularly if the quizzes were longer, would be to use more than two reference points, say eight in tests when from zero to three eliminations are combined with correct and incorrect choices. In the construction of decision latency scores it might be appropriate to explore other formulas for decision

latency weights since the present decision time scores were very highly correlated with raw scores.

There are several other variables which might best be manipulated in laboratory studies of response latencies to items in general intelligence tests. For example, there could be effects of instructional variables, such as telling the student that his score is dependent on how fast he can react, or effects of variation in testing procedures, such as feedback providing the reaction times to the examinee immediately after each is measured. More traditional variables of learning could also be studied to determine what desirable or undesirable characteristics were typical of latency weighted scores. For example, a test-retest experiment could determine if latency scores more accurately predicted retention of learned material. In summary, the present study should serve as a foundation for research efforts designed to examine the utility of latency measures as item weights to enhance estimation of an individual's true knowledge.

REFERENCES

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

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

Crandall, Vaughn, J., Solomon, Dan, and Kellaway, Richard. Expectancy statements and decision times as functions of objective probabilities and reinforcement values. Journal of Personality, 1955, 24, 192-203.

Crowne, Douglas P. and Marlowe, David. A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 1960, 24, 349-354.

Diamond, James and Evans, William. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.

Estes, W. K. Probability learning. In Categories of Human Learning, ed. A. W. Melton, 1964, 89-128, New York: Academic Press.

Festinger, Leon Studies in decision: I. Decision time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. Journal of Experimental Psychology, 1943, 32, 291-306.

Frary, Robert B. Elimination of the guessing component of multiple-choice test scores: Effect on reliability and validity and an evaluation of related item-weighting methods. Unpublished doctoral dissertation. Florida State Universtiy, 1968.

Frary, Robert B., and Zimmerman, Donald W. Effect of variation in probability of guessing correctly on reliability of multiple-choice tests. Educational and Psychological Measurement, 1970, 30, 595-605.

Gerjoy, I. R., Gerjoy, Herbert, and Mathias, Richard. Probability learning: Left-right variables and response latency. Journal of Experimental Psychology, 1964, 68, 344-350.

Geller, E. Scott, and Pitz, Gordon F. Confidence and decision speed in the revision of opinion. Organizational Behavior and Human Performance, 1968, 3, 190-201.

Geller, E. Scott, and Whitman, Charles P. Confidence in stimulus predictions and choice reaction time. Memory and Cognition, 1973, 1, 361-368.

Geller, E. Scott, Whitman, Charles P., and Farris, John C. Probability discrimination indicated by stimulus predictions and reaction speed: Effects of S-R compatibility. Journal of Experimental Psychology, 1972, 93, 404-409.

Geller, E. Scott, Whitman, Charles P., Wrenn, Richard F., and Shipley, William G. Expectancy and discrete reaction time in a probability reversal design. Journal of Experimental Psychology, 1971, 90, 113-119.

Hendrick, Clyde, Mills, Judson, and Kiesler, Charles A. Decision time as a function of the number and complexity of equally attractive alternatives. Journal of Personality and Social Psychology, 1968, 8, 313-318.

Hinrichs, J. V. Probability and expectancy in two-choice reaction time. Psychonomic Science, 1970, 21, 227-228.

Hull, C. L. Principles of Behavior. New York: Appleton-Century-Crofts, 1943.

Joe, Victor Clark. Review of the internal-external control construct as a personality variable. Psychological Reports, 1971, 28, 619-640.

Judd, Wilson A., and Glaser, Robert. Response latency as a function of training method, information level, acquisition and overlearning. Journal of Educational Psychology Monograph, 1969, 60, Part 2, 1-30.

Judd, Wilson A., and Glaser, Robert. Variability of response latency in paired associate learning as a function of the training procedure. Pittsburgh: Learning Research and Development Center, University of Pittsburgh: 1970. (Technical Report)

LaBerge, David, and Samuels, S. Jay. Toward a theory of automatic information processing in reading. Cognitive Psychology, 1974, 6, 293-323.

Lefcourt, H. M. Internal versus external control of reinforcements: A review. Psychological Bulletin, 1966, 65, 206-220.

Lowry, S. R. The effect of luck and misinformation on the discrepancy between multiple-choice test scores and true ability. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, 1975.

Moyer, Robert S., and Landauer, Thomas K. Time required for judgements of numerical inequality. Nature, 1967, 215, 1519-1520.

Myers, Jerome L., Gambino, Blase, and Jones, Mari R. Response speeds in probability learning. Journal of Mathematical Psychology, 1967, 4, 473-488.

Pollay, Richard W. The structure of executive decisions and decision times. Administrative Science Quarterly, 1970, 15, 459-471. (a)

Pollay, Richard W. A model of decision times in difficult situations. Psychological Review, 1970, 77, 274-281. (b)

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 1966, 80, No. 1 (whole No. 609).

Sarason, Irwin G. Interrelationships among individual difference variables, behavior in psychotherapy, and verbal conditioning. Journal of Abnormal and Social Psychology 1958, 56, 339-344.

Smith, E. E. Choice reaction time: An analysis of the major theoretical positions. Psychological Bulletin, 1968, 69, 77-110.

Taylor, J. A. A personality scale of manifest anxiety. Journal of Abnormal and Social Psychology, 1953, 48 285-290.

Taylor, J. A. Drive theory and manifest anxiety. Psychological Bulletin, 1956, 53, 303-320.

Whitman, Charles P., and Geller, E. Scott Prediction outcome, S-R compatibility, and choice reaction time. Journal of Experimental Psychology, 1971, 91, 299-304.

Whitman, Charles P., and Geller, E. Scott Sequential effects of stimulus probability and prediction outcome on choice reaction time. Journal of Experimental Psychology, 1972, 93, 373-378.

APPENDICES

APPENDIX I

INSTRUCTIONS

The scores for the quiz you are about to take will count as the third quiz for the class if you do better on it relative to the rest of the class. The questions are to be answered as on the in-class tests: you will eliminate alternatives you feel sure are incorrect and then chose the correct answer from the remaining alternatives. Two scores will be computed as usual: 1) the number of questions with the correct answer chosen, and 2) the sum of the number of incorrect answers eliminated minus three times the number of correct answers inadvertantly eliminated. Your participation in taking this test is voluntary and cannot hurt your grades.

First, let me introduce you to the computer terminal which will administer the test questions and record your responses. You will be using only thirteen keys on the terminal: the right-hand pad of twelve keys which is divided into three columns of four keys each--one for each of the alternatives on a question; and the key at the bottom right of the alphabet, marked ENTER.

There will be five practice questions which do not count, followed by twenty test items. For each test item the procedure will be as follows. When you are ready to view the first part or stem of a question, press the ENTER key once and it will appear. When you have read the stem

and understand it and can remember it for several minutes, press the ENTER key again to indicate that you are ready for it to be erased and the four alternatives to appear. If you can eliminate one of the answers as being incorrect, immediately press the corresponding ELIMINATION key. Keep eliminating incorrect answers one at a time until you want to guess among the remaining alternatives. As soon as you have determined your choice, inform the computer by pressing the corresponding CHOICE key in the middle column of four keys. When you have indicated your choice the alternatives will disappear and a box will be displayed in the middle of the screen. When you are ready, press the ENTER key and the number of the correct answer will be presented in the box for you to identify by pressing the REACTION key corresponding to the number displayed. I want you to react as fast as you can without pressing one of the wrong keys. When you are ready for the next question, press the ENTER key again.

If you make a mistake and press a key unintentionally eliminating the correct choice, you can restart the question by pressing the REACTION key 4. A question will also be restarted if you eliminate all four alternatives and make a choice.

Remember that the first five items are practice, and I will be available to answer any questions about procedure, but I do not know anything abcut the test material.

APPENDIX II

COMPUTER QUIZZES

COMPUTER QUIZ 3

1. ROGERS CONTENDS THAT PEOPLE WOULD NOT DEVELOP BEHAVIOR DISORDERS IF THEIR BEHAVIOR WERE GOVERNED BY THEIR:
   1. ORGANISMIC VALUING PROCESS.
   2. EMPATHIC FEELINGS.
   3. IDEAL SELF-CONCEPT.
   4. CONDITIONS OF WORTH.

2. TO THE PHENOMENOLOGISTS, PAST EVENTS ARE IMPORTANT ONLY TO THE EXTENT THEY:
   1. CAUSE PRESENT ACTIVITY IN THE MENTAL FORCES
   2. DETERMINE AN INDIVIDUAL'S PRESENT BEHAVIOR.
   3. ARE CONSISTENT WITH AN INDIVIDUAL'S "HERE AND NOW"
   4. EXERT INFLUENCE ON PRESENT PERCEPTIONS.

3. THE POTENTIAL FOR CHANGE IN A CONSTRUCT LIES IN ITS:
   1. PERMANENCE.
   2. PREDICTABILITY.
   3. PERMEABILITY.
   4. PRODUCTIVITY.

4. THE BASIC ASSUMPTION INVOLVED IN PHENOMENOLOGICAL PERSONALITY CHANGE IS THAT MODIFICATION OF BEHAVIOR IS PRODUCED BY CHANGING:
   1. UNCONSCIOUS IMPULSES INTO CONSCIOUS AWARENESS.
   2. AN INDIVIDUAL'S PERCEPTION OF HIS EXPERIENCES.
   3. OBJECTIVE VARIABLES THE THERAPIST BELIEVES ARE CAUSAL.
   4. THE ACTIONS CALLED FORTH BY THE SELF.

5. MASLOW'S RESEARCH ON THE SELF-ACTUALIZING PERSON INVOLVED:
   1. COLLECTED DATA ON HISTORICAL FIGURES.
   2. CONVENTIONAL SAMPLING OF SUBJECTS.
   3. EXPERIMENTAL MANIPULATION OF EVENTS.
   4. EVALUATIONS OF THE RELIABILITY OF HIS OBSERVATIONS.

6. WHICH OF THE FOLLOWING IS ANTITHETICAL TO SELF-ACTUALIZATION?
   1. PHENOMENOLOGY
   2. CONDITIONS OF WORTH
   3. UNCONDITIONAL POSITIVE REGARD
   4. ORGANISMIC VALUING PROCESS

7. TO KELLY, MAN'S ACTIVITIES ARE DIRECTED BY:
   1. THE CONSEQUENCES OF PREVIOUS BEHAVIOR.
   2. HIS PERCEPTION OF EARLIER CIRCUMSTANCES.
   3. HIS BIRTH INTO THE PSYCHOLOGICAL WORLD.
   4. THE ANTICIPATION OF FUTURE EVENTS.

8. ROGERS HAS FOUND THAT AS THE RESULT OF THERAPY:
   1. PEOPLE DEVELOP MORE CONDITIONS OF WORTH.
   2. PEOPLE'S PERCIEVED AND IDEAL SELVES BECOME MORE CONGRUENT.
   3. PEOPLE RAISE THEIR IDEAL-SELF EXPECTATIONS.
   4. PEOPLE BECOME MORE SELF-ACTUALIZING.

9. IN ORDER TO SHOW UNCONDITIONAL POSITIVE REGARD, A CLIENT-CENTERED THERAPIST:
   1. REPLIES ON AN UNTHREATENING, EMOTIONAL LEVEL.
   2. INFORMS THE CLIENT HE IS BEING UNDERSTOOD.
   3. ACCEPTS WITHOUT EVALUATION ALL REACTIONS EQUALLY.
   4. LISTENS BUT DOES NOT COMMENT ON THE CLIENT'S COMMENTS.

10. PHENOMENOLOGICAL PERSONALITY THEORIES GET AROUND THE QUESTION OF WHY MAN BEHAVES BY ASSUMING:
    1. "WHY" TO BE LESS IMPORTANT WHEN COMPARED TO "HOW" AND "WHEN."
    2. THE HIGHER FUNCTIONS OF MAN TO PRESUPPOSE BEHAVIOR.
    3. THAT MAN BEHAVES BECAUSE HE IS ALIVE.
    4. THAT ALL BIOLOGICAL ORGANISMS BEHAVE BECAUSE OF INNATE DRIVES.

11. IN FIXED-ROLE THERAPY, THE CLIENT IS OFTEN ASKED TO:
    1. ROLE-PLAY THE PERSON IN HIS SKETCH FOR A TRIAL PERIOD OF TIME.
    2. WRITE A SKETCH OF HOW HE WOULD LIKE TO BE PERCEIVED BY OTHERS.
    3. ROLE-PLAY A VARIETY OF ROLES TO SEE WHICH ONES WORK BEST FOR HIM.
    4. PLAY THE PART OF OTHERS IN HIS LIFE TO SEE THE WORLD FROM THEIR PERSPECTIVE.

12. THE MOST FUNDAMENTAL CONCEPT IN ROGERS' THEORY WOULD BE:
    1. ORGANISMIC VALUING PROCESS.
    2. SELF.
    3. ACTUALIZING TENDENCY.
    4. NEED FOR POSTIVE REGARD.

13. WHICH OF THE FOLLOWING IS NOT A CHARACTERISTIC OF ROGERIAN THERAPY?
    1. THE THERAPIST CLARIFIES THE CLIENT'S FEELINGS.
    2. THE THERAPIST RESTATES THE CLIENT'S STATEMENTS.
    3. THE THERAPIST EMPATHIZES WITH THE CLIENT'S FEELINGS.
    4. THE THERAPIST INTERPRETS THE CLIENT'S STATEMENTS.

14. MASLOW'S PERSONALITY THEORY DIFFERSS FROM THOSE OF ROGERS AND KELLY IN THAT HE:
    1. DEALT ONLY WITH ABNORMALITY INVOLVED IN HIS CLINICAL PRACTICE.
    2. CHOSE TO CONCERN HIMSELF WITH MAN'S HIGHEST FUNCTIONS.
    3. RELEGATES MAN TO AN INITIAL STATE OF BEING INACTIVE IN NATURE.
    4. FOCUSED ON MAN'S HEALTHY PERSONALITY.

15. KELLY'S DEFINITION OF AGGRESSION IS MOST CLOSELY RELATED TO HIS CONCEPT OF:
    1. EXTENSION OF A CONSTRUCT SYSTEM.
    2. DEFINITION OF A CONSTRUCT SYSTEM.
    3. CONTINUED VALIDATION OF A CONSTRUCT.
    4. DEVIANT ROLE BEHAVIOR.

16. ACCORDING TO ROGERS, THREAT COMES INTO EXISTENCE WHEN:
    1. CONDITIONS OF WORTH ARE EMBODIED IN THE SELF-CONCEPT.
    2. THE NEED FOR POSITIVE REGARD BECOMES A REGULATING FORCE.
    3. CONDITIONAL POSITIVE REGARD IS WITHDRAWN.
    4. EXPERIENCES ARE INCONSISTENT WITH THE SELF-CONCEPT.

17. ACCORDING TO BOTH ROGERS AND KELLY, UNCONSCIOUS PROCESSES BECOME GREATER WHEN:
    1. ONE HAS A HEALTHY PERSONALITY.
    2. BEHAVIOR GROWS MORE DEVIANT.
    3. ONE GROWS OLDER.
    4. UNDERGOING CLIENT-CENTERED THERAPY.

18. AN IMPORTANT LIMITATION OF MASLOW'S DEFINITION OF SELF-ACTUALIZING PEOPLE IS THAT:
    1. SELF-ACTUALIZATION CANNOT BE DEFINED.
    2. MASLOW'S DEFINITION IS CIRCULAR.
    3. COMMON DEFINITIONS PROVE MEANINGLESS WHEN APPLIED TO SELF-ACTUALIZERS.
    4. SELF-ACTUALIZING INDIVIDUALS ARE REALLY EXCEPTIONAL PEOPLE.

19. ACCORDING TO ROGERS, PSYCHOTIC BEHAVIOR IS INDICATIVE OF:
    1. LACK OF DEFENSES.
    2. AMBIVALENT IDEAL SELVES.
    3. UNCONDITIONAL POSITIVE REGARD.
    4. NOT ENOUGH SELF-ACTUALIZATION.

20. CLIENT-CENTERED THERAPY IS SIMILAR TO FREE ASSOCIATION IN PSYCHOANALYTIC THERAPY IN THAT IT ATTEMPTS TO:
    1. FIND SYMBOLIC MEANING IN A PERSON'S THOUGHTS.
    2. LOCATE CONFLICTS WITHIN THE PERSON'S SEXUAL DEVELOPMENT.
    3. SET CONDITIONS SO A PERSON WILL BE OPEN ABOUT HIS EXPERIENCES.
    4. FIND THE PHYSICAL VARIABLES MOST LIKELY TO BE CAUSING BEHAVIOR.

COMPUTER QUIZ 4

1. OPERANT BEHAVIOR IS _____ , WHILE RESPONDENT BEHAVIOR IS _____.
   1. CONSEQUENTIAL; AUTOMATIC
   2. EMITTED; ELICITED
   3. OPERATED ON; RESPONDED TO
   4. LEARNED; INNATE

2. IF YOU WERE STARTING TO TEACH A FRIEND HOW TO DRIVE A CAR, IT WOULD BE BEST TO START HIM ON A:
   1. CRF SCHEDULE.
   2. VI SCHEDULE.
   3. FI SCHEDULE.
   4. FR SCHEDULE.

3. NEGATIVE REINFORCEMENT:
   1. DECREASES FREQUENCY OF A BEHAVIOR BY REMOVING A STIMULUS AFTER THE BEHAVIOR
   2. INCREASES FREQUENCY OF A BEHVAIOR BY PRESENTING A STIMULUS AFTER THE BEHAVIOR
   3. INCREASES FREQUENCY OF A BEHAVIOR BY REMOVING A STIMULUS AFTER THE BEHAVIOR
   4. DECREASES FREQUENCY OF A BEHAVIOR BY PRESENTING A STIMULUS AFTER THE BEHAVIOR

4. THE HISTORICAL ROOTS OF THE BEHAVIORAL STRATEGY ARE FOUND IN:
   1. OBSERVATIONS OF HUMAN BEHAVIOR.
   2. STUDIES OF COMPONENTS OF HUMAN BEHAVIOR.
   3. EXPERIMENTS IN HUMAN RESPONSE.
   4. LABORATORIES INVESTIGATING LEARNING.

5. REINFORCEMENT IN CLASSICAL CONDITIONING IS SAID TO OCCUR WHEN A:
   1. CR IS PAIRED WITH A UCS.
   2. UCR IS PAIRED WITH A CR.
   3. CS IS PAIRED WITH A UCS.
   4. CS IS PAIRED WITH A CR.

6. THE NATURE OF PREDICTIVE INFERENCES MADE BY A STRICT BEHAVIORAL PSYCHOLOGIST IS FROM:
   1. PAST BEHAVIOR TO FUTURE BEHAVIOR.
   2. PAST PERFORMANCE TO PRESENT ABILITY.
   3. PRESENT BEHAVIOR TO FUTURE LEARNING.
   4. PRESENT LEARNING TO ABILITY.

7. THE BASIC DEPENDENT VARIABLE IN OPERANT CONDITIONING IS:
   1. TIME TO CRITERION.
   2. LATENCY OF THE OPERANT RESPONSE.
   3. RATE OF EMISSION OF THE OPERANT RESPONSE.
   4. NUMBER OF TRIALS TO CRITERION.

8. ACCORDING TO BANDURA AND WALTERS, ACQUISITION IS PRIMARILY INFLUENCED BY:
   1. MOTIVATION.
   2. EXPOSURE.
   3. PUNISHMENT.
   4. REWARD.

9. IN MOWRER AND MOWRER'S WORK ON ELIMINATING BED WETTING BY CLASSICAL CONDITIONING, A BELL SERVED AS THE:
   1. UCS.
   2. UCR.
   3. CS.
   4. CR.

10. FROM THE OBSERVER'S POINT OF VIEW, DESIRABLE OUTCOMES WHICH OCCUR CONTINGENT UPON A MODEL'S BEHAVIOR ARE CALLED:
   1. SECONDARY REINFORCEMENTS.
   2. POSITIVE VALENCIES.
   3. VICARIOUS CONSEQUENCES.
   4. VICARIOUS REWARDS.

11. PUNISHMENT DIFFERS FROM NEGATIVE REINFORCEMENT IN THAT PUNISHMENT:
    1. DEALS WITH AVERSIVE STIMULI.
    2. PRODUCES A SUPPRESSION OF RESPONDING.
    3. CAN PRODUCE MORE GENERALIZATION OF RESULTS.
    4. MAKES CONTINGENT USAGE UNNECESSARY.

12. THE TECHNIQUE KNOWN AS "SHAPING" INVOLVES:
    1. GENERALIZATION OF SUCCESSIVE APPROXIMATIONS.
    2. CONDITIONING OF SECONDARY RESPONSES.
    3. REINFORCEMENT OF SUCCESSIVE COMPONENTS.
    4. LEARNING OF SEQUENTIAL TASKS.

13. WHICH OF THE FOLLOWING IS A PARTICULARLY USEFUL TOOL TO USE ALONG WITH PUNISHMENT?
    1. SELECTIVE REWARD FOR A COMPETING RESPONSE.
    2. POSITIVE REINFORCEMENT.
    3. AVOIDANCE TRAINING.
    4. DIFFERENTIAL REINFORCEMENT OF A NONCOMPETING RESPONSE.

14. MODELING REFERS TO BEHAVIOR OF PERSONS IN A:
    1. FAMILY CONTEXT.
    2. INDIVIDUAL CONTEXT.
    3. GROUP CONTEXT.
    4. SOCIAL CONTEXT.

15. PRESENTING A FEARED OBJECT TO A SUBJECT WHILE HE IS DOING SOMETHING PLEASANT IS CALLED:
    1. POSITIVE FEAR REDUCTION.
    2. RECIPROCAL INHIBITION.
    3. VICARIOUS REWARD.
    4. ORGANISMIC REORIENTATION.

16. WHICH LIST OF REINFORCEMENT SCHEDULES IS IN ORDER OF INCREASING RATE OF RESPONDING IN HUMANS?
    1. CONTINUOUS, FIXED RATIO, FIXED INTERVAL.
    2. FIXED INTERVAL, FIXED RATIO, CONTINUOUS.
    3. FIXED RATIO, FIXED INTERVAL, CONTINUOUS.
    4. CONTINUOUS, FIXED INTERVAL, FIXED RATIO.

17. WHICH OF THE FOLLOWING IS NOT AN UNDESIRABLE SIDE EFFECT OF PUNISHMENT?
    1. VERY RAPID EXTINCTION.
    2. EMOTIONAL RESPONSES.
    3. DISCRIMINATED OPERANT.
    4. GENERALIZATION OF FEAR.

18. RAZRAN'S "LUNCHEON TECHNIQUE" DID NOT INVOLVE:
    1. FREE LUNCH.
    2. OPERANT CONDITIONING.
    3. ATTITUDES.
    4. CLASSICAL CONDITIONING.

19. THE BEST EXAMPLE OF RESPONDENT BEHAVIOR IS:
    1. DRIVING A CAR.
    2. PLAYING A TENNIS MATCH.
    3. BLINKING EYES IN BRIGHT LIGHT.
    4. STUDYING FOR A MAJOR EXAMINATION.

20. A SEAMSTRESS IN A FACTORY IS PAID $1.50 FOR EACH SHIRT SHE COMPLETES. SHE IS PAID ON A:
    1. VARIABLE-INTERVAL SCHEDULE.
    2. VARIABLE-RATIO SCHEDULE.
    3. FIXED-INTERVAL SCHEDULE.
    4. FIXED-RATIO SCHEDULE.

The vita has been removed from the scanned document

THE APPLICATION OF DECISION TIMES AND REACTION TIMES IN THE CONSTRUCTION OF LATENCY WEIGHTED TEST SCORES

by

Charles Philip Whitman

(ABSTRACT)

The use of response latencies to determine item weights in the construction of multiple-choice test scores was investigated. Three times were measured on each test item by the computer system which administered two tests. The times recorded were: the reading time which the student used to read the stem of the question; the decision time which the student used to read the alternatives, to eliminate the incorrect choices he could identify, and to indicate the correct choice; the third time recorded was the choice reaction time which the student required to identify the number of the correct alternative.

Each of the three response times was analyzed as a function of two independent variables: the number of alternatives eliminated, and the correctness of the answer selected. Reading time was not significantly related to either of the independent variables. Decision time was shorter when correct answers were selected than when incorrect answers were selected. Furthermore, decision time was shorter when zero or three alternatives were eliminated than when one or two alternatives were eliminated. Reaction time was also shorter when correct answers were selected

than when incorrect answers were selected. In contrast to the decision time results of two independent main effects, the reaction time analysis indicated an interaction between correctness of the student's choice and the number of alternatives eliminated: when the decision was correct, reaction time became shorter as the number of eliminations increased; when the decision was incorrect, reaction time became longer as the number of eliminations increased. Both decision time and reaction time results were consistent with those of laboratory studies.

Item weights were constructed as the differences between item response latencies for each student so that between student differences in absolute response time were eliminated. Given a confidence construct popular in decision time and reaction time research, the latency item weights were formulated to maximize weight of items answered with relative certainty and to minimize weight of items answered with relative uncertainty. Test scores used in evaluating the latency weighted scores included raw scores (the number correct) and the Coombs mode scores (the number of alternatives correctly eliminated minus three times the number of alternatives incorrectly identified, since all items had four choices), and several personality trait scores. The reaction time scores had higher validity estimates than either the Coombs mode or raw scores from the same test, but did not correlate with corresponding raw

scores as well as the Coombs mode scores. In contrast, the decision time scores had validity estimates higher than raw scores and comparable to the Coombs mode scores, but were very highly correlated with the corresponding raw scores. In addition, decision time scores correlated with the exam raw scores moreso than any other measure. Finally, the effect of personality traits on each of the test scores was investigated. Both reaction time and decision time scores were less correlated with a measure of test-taking anxiety than either the Coombs mode or raw scores.