A NEW APPROACH TO MEASUREMENT OF PARTIAL KNOWLEDGE

APPROVED:

[Signatures]

Major Professor

Minor Professor

Chairman of the Department of Psychology

Dean of the Graduate School

The problem of this study is that of developing a testing procedure for multiple-choice tests which would increase the relationship between test scores and a criterion. The procedure investigated in this research was one in which subjects took a multiple-choice test but were required to continue responding on each item until the correct answer was obtained. The total number of responses was used as the score on the test.

The purpose of this research was to investigate the possibility of increasing predictability by changing the procedure of administering the test, rather than changing the test itself.

In order to determine whether the new method did significantly increase predictability, coefficients of correlation and t tests were used to determine significance or correlations and significant difference between correlations. Correlation coefficients were calculated for two sets of data. The first set of data consisted of subjects' accumulated grade point averages and conventional scores
on the test. The conventional score was determined by the number of correct responses made on the first attempt on each item. The second set of data consisted of the Ss' accumulated grade point averages and the experimental test scores. These two correlations were tested for significance and were found not significant at the .05 level. The correlations were also tested for significant difference. The t test for significance found the two correlations not to be significantly different.

These findings indicate that the experimental method did not enhance predictability. The experimental method is evaluated, and suggestions for further research are given.
A NEW APPROACH TO MEASUREMENT OF PARTIAL KNOWLEDGE

THESIS

Presented to the Graduate Council of the North Texas State University in Partial Fulfillment of the Requirements For the Degree of

MASTER OF SCIENCE

By

David E. Wagner, B. S.
Denton, Texas
August, 1973
A NEW APPROACH TO MEASUREMENT
OF PARTIAL KNOWLEDGE

In the field of testing there are many attempts being made to develop a testing procedure to capture partial knowledge that the examinee might not display on the conventionally administered multiple-choice test. This research was conducted in order to test a new administrative approach to testing that attempts to capture and fully utilize partial knowledge.

The decision to use the multiple-choice test was made after observation of the advantages and disadvantages of both objective and subjective testing procedures. Anastasi (1968) reports that the most commonly discussed advantages of subjective essay tests are that "they test the individual's ability to select, relate, and organize material, as well as his ability to express ideas clearly and accurately" (p. 160). She continues by saying that with the time limitations on essay tests, it becomes almost impossible for the individual to spend his time organizing and developing his ideas.

The essay test also is subjectively graded. The subjectivity in grading can prevent an accurate assessment of the individual's knowledge. While the essay test has
many problems, the objective test also has its shortcomings. Even though the objective tests can be more accurately scored, their ability to fully assess an individual's knowledge of a subject area is sometimes doubtful. The most common types of objective tests are true-false, matching, completion, and the multiple-choice tests. Of these types of tests, it was felt that the multiple-choice test gave the most objective evaluation of an individual's knowledge of a given subject.

Previous research in the area of measurement of partial knowledge found that the multiple-choice test was the most effective form for this type of testing (Willey, 1960).

The most commonly used form of the multiple-choice test for measuring partial knowledge of a particular subject area is the Three-Decision Test. The Three-Decision Test involves having the S mark his answer sheet in a manner designed by Willey (1960). The S records his responses on a five-option Programmed Multiple Response (P.M.R.) answer sheet. The examinee is told to record the answer he feels is correct by circling the corresponding box on the answer sheet. The S is then to place an x in the squares for the two options which he feels are definitely wrong. The answer sheet is then punched with a "bed of tacks" marking device. If the correct answer is circled, then there is no penalty. Correct answers
not marked by the $S$ are given one penalty point. If the correct answer is marked with an $x$, three penalty points are given.

The Three Choice Test was developed, according to Coombs, Milholland, and Womer (1956), in order to investigate "the notion that while an individual may not know the correct answer to an item, he may know some of the things which are wrong. This is called partial knowledge" (p. 14). If the concept of partial knowledge is true, and a method could be devised to capture this partial knowledge, the significance in the approach would be a difference in the distribution of scores. Willey (1960) in his investigation of the Three-Decision Test, found that the distribution of scores for the Three-Decision Test did have a significant shift in the center of the distribution when it was compared to conventionally administered multiple-choice tests. Willey reported that the examinees with scores at either end of the distribution are less likely to change their rank, but there are striking shifts within the middle two thirds. Therefore, the below-average student, who might score relatively high on the conventional multiple-choice test by luck, would be most affected by the three-point penalty. The change in the distribution of scores brought about by the use of the Three-Decision Test is
reported by Willey (1960) to be due to the decrease of the scores of test-sophisticated, or good-guessing students. The increased range of the S's scores provided by this method results in a more heterogenous distribution, thus providing an opportunity for increased correlations between the test scores and the chosen criterion.

While this type of testing does provide more sensitivity to partial knowledge than the conventionally administered multiple-choice test, there are some practical drawbacks. Willey (1960) mentions the fact that there is an increased amount of time in scoring and administering the Three-Choice Test. The conventionally administered multiple-choice test requires that each item be scored as correct or incorrect. The Three-Decision Test, as discussed earlier, must be scored for three different error types; thus, the scoring time is greatly increased. A second problem arising with the use of the Three-Decision Test is that it does not fully utilize all the alternatives of each multiple-choice item. While the Three-Decision Test requires that the individual account for three alternatives, it does not require any type of decision to be made on the other alternatives. The Three-Decision procedure will not affect the scores of those Ss who know the correct answers to the items. However, individuals who guess randomly at an item, have
a greater probability of success with the Three-Decision procedure than if they were required to make a decision on each alternative until they find the correct answer.

Another form of the multiple-choice test now under investigation is called 100% mastery. The difference between 100% mastery and the Three-Decision Test is that it requires the individual to continue guessing on each item until the correct answer is obtained. The 100% mastery procedure allows for complete utilization of each alternative on each item. Each individual will not need to use each alternative; however, the individual who has no knowledge of a given topic will find that by selecting randomly, he will be required on some items to make as many selections as there are alternatives. 100% mastery is currently being used by educators as a teaching aid.

A study by Mount (1971) shows how 100% mastery is currently being utilized in the field of education. Mount's study was on the use of 100% mastery in the elementary psychology courses at North Texas State University. 100% mastery was used on chapter exams, and the students were graded according to how many chapter exams they mastered. The student was required to obtain the correct answer on the first attempt on eight of ten multiple-choice questions. If the student did not obtain the correct answer, he would have to continue his attempt until the correct answer was
obtained. In this way, he was learning as he was tested. If the student did not answer eight questions correctly on the first attempt, he was required to take the chapter exam again. However, the same set of questions was not given twice, so the passing of the test was not merely a process of short term memory. This use of 100% mastery is the typical application. The purpose of 100% mastery in this setting is as a criterion, rather than a predictor. The current literature reveals no utilization of 100% mastery as a predictor, yet its design would seem to lend itself to this purpose. The concept and design of the Three-Decision Test and 100% mastery cause speculation that a combination of the two methods could possibly be developed into a more sensitive application of the multiple-choice test.

A test using the methodology of 100% mastery and having the purpose of the Three-Decision Test may have the ability to fully utilize the multi-choice test. The testing procedure would be similar to the 100% mastery, in that it requires the s to continue his attempt on each item until he obtains the correct answer. The test would differ from 100% mastery in that it would be administered only once, and the total number of attempts to mastery would be the score. This approach would fulfill the purpose of the Three-Decision Test in that it would measure partial knowledge and would penalize for guessing. The partial knowledge would be
measured in that if the examinee had partial knowledge, he could eliminate known incorrect answers and thus lower the number of attempts necessary to obtain the correct answer. Guessing would be penalized in that the person who guesses will have to make more attempts. While the guessing examinee could possibly get some answers right on the first guess, he would not get all items correct due to the probabilities involved in guessing. On the questions where the S guesses incorrectly, this type of test design will more effectively indicate guessing. On a conventionally administered multiple-choice test, the S would be penalized only one point for guessing and missing; in this test design, he could lose as many points as there are options. When a S has completed a test, it is the total attempts which will be used as the score.

In order to administer a test of this design, a machine called a Mycom unit must be used. The Mycom is a pre-programmed teaching computer. It is designed so that a logistic program can be selected and keyed into it. When the program is keyed in, the S is given the test. The subject reads the item and selects the answer which he feels is correct. The item number is then found on the dial on the Mycom, and the dial is turned until the number corresponds to the arrow above the dial. The S then presses the response button, and if his selection is correct, a
green light flashes. If the selection is incorrect, the red light flashes. If the green light flashes, a response is recorded on the counter for correct response on the first attempt, and one attempt is recorded on the counter for total attempts. If the red light flashes, the S has one response recorded on the total attempts counter. The S must then select another answer and dial it in and press the response button. The S must continue responding until the correct answer is obtained. The only response which activates the first attempt counter is a correct response on the first attempt. The Mycom can be programmed so that the S must obtain the correct answer before progressing to the next question. The Mycom will record total number of responses correct on the first attempt. The count obtained on the first attempt counter corresponds to the score that would be obtained on a conventionally administered multiple-choice test. The count obtained on the total responses counter is the score of the experimental 100% mastery test. The counters attached to the Mycom unit eliminate any scoring time for the E. The two sets of scores allow the E to compare the experimental method to the conventional method of multiple-choice administration.

The purpose of this research is to investigate the following two questions: 1. If a test is administered in a manner requiring the Ss to obtain the correct answer on
each item before continuing to the next item, will the distribution of total number of responses have a greater range than the distribution of a conventionally administered multiple-choice test? 2. Will the increase in the range of scores cause the test to show an increase in its relationship with an external criterion?

The hypothesis of this research is that if Ss are required to obtain the correct answer on each item, the cumulative score will have a significantly higher correlation with the criterion than a conventional score.

Method

Subjects. Forty-six students enrolled in the second semester of Introductory Psychology courses at North Texas State University served as Ss. There were 21 females and 24 males. The Ss had a minimum of a sophomore standing, or 30 hours credit towards their bachelor's degree, and had completed the first semester of Introductory Psychology.

Instrument. The instrument used was a departmental exam normally given at the end of the first Introductory Psychology course. The items were written in multiple-choice form.

The correct answer on each item was positioned so that the order of the answers would correspond with the logistic program on the Mycom units. The test was administered on the Mycom units for purposes of scoring the number
of correct responses on first attempts and total number of attempts to 100% mastery.

**Criterion.** The Ss accumulated grade point average was used as the criterion. Each student was required to report their grade point average at the time of testing.

**Procedure.** The Ss were placed in a room containing three desks with a Mycom unit on each desk. Each S was placed at a desk and given a copy of the test. The front page of the test was an introduction page containing instructions on the use of the Mycom and sample questions. The test administrator read the instructions aloud and went through the sample questions and answered any question about the use of the Mycom unit. After all questions were answered, the Ss were told to turn the page and begin. Questions pertaining to the use of the Mycom were answered during testing, but the questions about item interpretations were ignored. When each S completed the test, the count on both counters was recorded along with the S's grade point average on the first 30 college hours attempted.

**Results**

The statistical analysis used in this experiment consisted of two Pearson product moment correlations and one converted t test. The t test was used to determine the significant difference between two correlation coefficients with a common x variable, or in this case, a
common criterion.

The reliability of the test had previously been measured at .87 (split-half). The population used for the reliability study was 200 students enrolled in the Introductory Psychology course in the fall semester of 1970 at North Texas State University.

The data gathered in this experimentation were first analyzed in two categories. The two categories consisted of conventional scores and experimental scores. Two sets of scores were both correlated with the Ss' grade point averages.

The grade point averages of the Ss, or the x variable, were the same in both correlations. The means, standard deviations, and ranges for x, y₁, y₂ were calculated and recorded in Table 1 below.

Table 1
Means, Standard Deviations, and Ranges of the Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>S.D.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.P.A.</td>
<td>2.69</td>
<td>.18</td>
<td>2.34</td>
</tr>
<tr>
<td>Conventional Scores (y₁)</td>
<td>71.00</td>
<td>22.79</td>
<td>63.00</td>
</tr>
<tr>
<td>Experimental Scores (y₂)</td>
<td>312.00</td>
<td>31.00</td>
<td>123.00</td>
</tr>
</tbody>
</table>

The Pearson r calculated for the conventional scores and S's grade point average was .53. The Pearson
r calculated for the experimental score and S's grade point average was -.40. A third r was calculated between \( y_1 \) and \( y_2 \), and the r was -.71.

A t test to determine significance for each of the correlations was calculated. The t tests showed \( r_{x_1y_1} \) and \( r_{x_1y_2} \) not significant. The correlation of \( y_1 \) and \( y_2 \) was significant at the .05 level. The t test between \( r_{x_1y_1} \) and \( r_{x_1y_2} \) had a t score of 1.1930 (df = 43). The difference was not significant at the .05 level.

**Discussion**

The statistical analysis determined no significant difference between the two administrative approaches. The correlation between the two sets of predictors showed them to be closely parallel. Neither set acted as a successful predictor, as indicated by the t test scores showing them to not be significantly correlated with the criteria. A further analysis reveals that the range of the two sets of scores did differ. The difference in the range can be accounted for by the scoring methodology. The experimental scores were cumulative response scores, thus automatically increasing the possible range of scores. The difference in the ranges was as predicted; however, the difference in S.D.'s was not great enough to indicate a significant shift in the scores. If the hypothesis of this research had been valid, the experimental scores would have had
significantly higher correlations with the criterion scores regardless of the overall validity of the predictors. From this analysis, it must be concluded that for these particular sets of variables, the hypothesis of this research was not supported.

The fact that the new scoring method used in this experimentation was not successful with the selected variables should not stop future research utilizing this new administrative approach. The E observed during experimentation that there could have been an uncontrolled variable hindering an accurate portrayal of the new approach's validity. The variable in question was the extent of actual effort used by Ss in order to find the correct answer after they had guessed incorrectly on the first attempt. The problem was that the Ss were conditioned to being rewarded as the consequence of their first attempt at an item. Therefore, any attempt made after the first attempt had no conditioned reinforcement. If the assumption that later attempts were not reinforcing is true, then it could possibly explain random selection behavior observed in the later stages of the test. The observation was made that speed with which the button on the Mycom unit was pressed increased as the S drew nearer the end of the test. In the case of many Ss, the button press was as rapid as the alternate selector could be moved. In the
early stages of the test, the speed of selection was much slower and seemed to indicate that the Ss were carefully observing each alternative after an incorrect guess. As the test progressed, the selection of new alternate answers increased, indicating that the Ss were using less effort to choose another alternative. It would seem logical then, that by placing proper consequences on good and poor performance, as determined by the experimental method, this variable could be controlled. With the proper controls, the 100% mastery technique could be a very helpful tool in the field of testing.

If this procedure is valid, the areas of testing in which it could be helpful are manifold. One area in which this procedure might be useful is pre-employment testing. The current emphasis for equal employment practices had caused much industrial testing to be outlawed. It is now necessary to develop valid procedures for screening. With the current level of high employment and the ever-present need to find trainable employees, a test that could detect applicants with partial knowledge in the employees area of work could be of great value to the employer. The 100% mastery method could also be beneficial in developing tests which do not illegally discriminate. If “test wiseness” is directly related to
the level of education, then many tests are inadvertently discriminating. This discrimination is brought about because non-whites typically drop out of school before whites. Thus, the possibility arises that through education, whites have obtained more "test wiseness." This "test wiseness" (as opposed to actual ability) enables them to score higher on conventionally administered multiple-choice tests. The new 100% mastery test procedure theoretically could help eliminate this discrepancy. The possibilities for the use of this approach are plentiful, and this would lead one to believe that additional research time in this area would be very beneficial.
REFERENCES


