DOCUMENT RESUME

ED 058 318 TM 001 034

AUTHOR	Echternacht, Gary J.; And Others
TITLE	An Evaluation of the Feasibility of Confidence Testing as a Diagnostic Aid in Technical Training.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Air Force Human Resources Lab., Wright-Patterson AFB, Ohio.
REPORT NO	AFHRL-TR-71-33
PUB DATE	Jul 71
NOTE	139p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Confidence Testing; *Cost Effectiveness; Educational Diagnosis; Evaluation; *Feasibility Studies; Grades (Scholastic); Guessing (Tests); Item Analysis; *Job Training; Multiple Choice Tests; Personality Assessment; Psychometrics; Questionnaires; Scoring Formulas; Student Attitudes; Student Opinion; Teacher Attitudes; *Technical Education; Testing Programs

ABSTRACT

The feasibility and the cost-effectiveness of using confidence testing as a diagnostic aid in technical training programs were studied. Two types of confidence testing, Pick-One and Distribute 100 Points, were developed for comparison to conventional multiple-choice testing. The criteria for feasibility included end of block examination grades, number of student remediational sessions, and both student and instructor attitudes. In addition, the relationship of various personality variables to confidence test scores was examined for both types of confidence testing. The major finding was that while scoring was somewhat more time consuming, end of block examination grades improved slightly and the number of remediations required declined slightly when either confidence testing method was employed. Other areas of investigation produced essentially null results. Copies of the Student Attitude Questionnaire and the Instructor Questionnaire are appended. (Author/MS)

AFHRL-TR-71-33

AN EVALUATION OF THE FEASIBILITY OF CONFIDENCE TESTING AS A DIAGNOSTIC AID IN TECHNICAL TRAINING

By

Gary J. Echternacht
Educational Testing Service

Wayne S. Sellman, Capt, USAF
Technical Training Division (AFHRL)

Robert F. Boldt
Educational Testing Service

Joseph D. Young, Capt, USAF
3345th Technical School

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado

July 1971

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

AFHRL-TR-71-33

July 1971

AN EVALUATION OF THE FEASIBILITY OF CONFIDENCE TESTING AS A DIAGNOSTIC AID IN TECHNICAL TRAINING

By

Gary J. Echternacht
Educational Testing Service
Wayne S. Sellman, Capt, USAF
Technical Training Division (AFHRL)
Robert F. Boldt
Educational Testing Service
Joseph D. Young, Capt, USAF
3345th Technical School

TECHNICAL TRAINING DIVISION
AIR FORCE HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND
Lowry Air Force Base, Colorado

FOREWORD

This research represents a portion of the exploratory development program of the Technical Training Division, Air Force Human Resources Laboratory. The work was documented under Project 1121, Technical Training Development; Task 112103, Evaluating Individual Proficiency and Technical Training Programs, and was completed during the period July 1970 through June 1971. Dr. Marty R. Rockway was the Project Scientist and Capt Wayne S. Sellman was the Task Scientist. The study was performed in cooperation with the 3345th Technical School, Chanute AFB, Illinois. The services of Educational Testing Service, Princeton, New Jersey were obtained under Contract F41609-70-C-0044 of which Dr. Robert F. Boldt and Dr. Gary J. Echternacht served as co-principal investigators. Capt Wayne S. Sellman was the Air Force technical monitor and Capt Joseph D. Young was the Chanute AFB project officer.

Included among the many individuals who contributed to the accomplishment of this study were Major B. J. Dunnington, MSgt J. R. Fitzpatrick, and Mr. J. E. Ross, Jet Engine Branch, Aircraft Maintenance Training Department; and SMSgt H. L. McKellip, MSgt R. E. Cosner, and Mr. C. H. Ervin, Aerospace Ground Equipment Branch, Department of Weapon Systems Support Training; 3345th Technical School, Chanute AFB, Illinois.

This report has been reviewed and is approved.

GEORGE K. PATTERSON, Colonel, USAF
Commander

ABSTRACT

This report describes a study to determine the feasibility and the cost-effectiveness of using confidence testing as a diagnostic aid in technical training programs. Two types of confidence testing, Pick-One and Distribute 100 Points, were developed for comparison to conventional multiple-choice testing. The study was carried out in two technical training courses, Aerospace Ground Equipment Repairman (AGE) and Jet Engine Mechanic (JEM), currently being taught at Chanute Air Force Base, Illinois. The criteria for feasibility included end of block examination grades, number of student remediational sessions, and both student and instructor attitudes. In addition, the relationship of various personality variables to confidence test scores was examined for both types of confidence testing. The major finding was that while scoring was somewhat more time consuming, end of block examination grades improved slightly and the number of remediations required declined slightly when either confidence testing method was employed. Other areas of investigation produced essentially null results.

SUMMARY

Echternacht, G. J., Sellman, W. S., Boldt, R. F., and Young, J. D. An evaluation of the feasibility of confidence testing as a diagnostic aid in technical training, AFHRL-TR-71-33. Lowry AFB, Colo: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Problem

The purposes of this study were: (1) to determine the feasibility of using confidence testing, where the student responds in terms of his degree of confidence in item alternatives, as a diagnostic evaluative aid to instructors in Air Force technical training courses; and (2) to determine the cost-effectiveness of confidence testing versus the conventional multiple-choice testing now used in Air Force technical training courses.

Approach

Two experimental forms of confidence testing, termed Pick-One and Distribute 100 Points, were developed for use in the experiment. These experimental forms of testing were used by students in two different courses as was traditional multiple-choice testing. In addition, a special type of student remediation was developed and used with each type of testing as was the standard remediation procedure.

The various types of testing and remediation were used with daily quizzes administered as diagnostic aids. Criterion data consisted of end of block examination scores and the number of remediations required of each student. Both students and instructors were also asked to indicate their attitude toward the confidence methods. Records were kept indicating the length of time required to score the confidence tests and the time required for their administration. Personality tests which might be related to any tendency to mark confidence in a manner unrelated to achievement were also administered.

Results

In the analysis of the end of block examination scores, significant interactions were found which did not allow the interpretation of overall differences between the types of testing. The effectiveness of the types of testing varied with the different training shifts involved in the experiment. In general, the group using multiple-choice testing obtained lower end of block examination grades than did either group using confidence testing though the size of this effect varied from shift to shift. Students in the group using multiple-choice testing also required more remediations on the average than students using confidence testing.

It was found that confidence testing required slightly more administration time than did multiple-choice and that it required about twice as much time to score. Students were found to be only slightly favorable to confidence testing while instructors tended to indicate that it was more precise than their needs required and disliked the required increase in scoring time. In contrast to some studies, no personality variables were found to be substantially correlated with the confidence test score when differences in the number right were controlled.

Conclusions

Two experimental methods of confidence testing were developed for use with diagnostic tests administered in technical training courses. These methods resulted in improved end of block examination scores in instances where differences in the types of testing were found to be significant and in fewer average remediations. No significant personality variables were found to be substantially related to the process of allocating one's confidence. On the negative side, the time required to administer and score the quizzes increased, especially the scoring time, and the instructors objected to this increased scoring time.

This summary was prepared by Wayne S. Sellman, Technical Training Division, Air Force Human Resources Laboratory.

TABLE OF CONTENTS

TABLES AND FIGURES

TABLES AND FIGURES (Continued)

SECTION I

Introduction

One of the primary tasks that faces instructors in technical training situations is that of accurately assessing student knowledge of course materials. Often daily quizzes are administered as "diagnostic" aids to identify areas of instruction in which students are strong or weak. Although the recent past has seen the development of numerous devices and techniques for improving instruction, little has been done to improve the methods of measuring student achievement and diagnosing their strengths and weaknesses.

One of the most popular methods of testing student achievement is through the use of multiple-choice test items where the examinee is presented a question and a number of alternatives from which he is to choose the correct answer. However, the notion of requiring an examinee to choose only one alternative from a fixed number has been subject to criticism. For many items the examinee is quite sure as to the correct choice and has no difficulty indicating it; on the other hand, he may be able to eliminate some of the alternatives and then be forced to guess between the rest. Knowledge is not an all-or-none proposition. It seems reasonable to assume that a student who can eliminate some alternatives has more knowledge or insight than one who can eliminate none, and a student who selects an answer and indicates his doubt as to its correctness has more knowledge or insight than one who is completely misinformed and yet certain of his answer.

One possible approach for providing diagnostic information to instructors is confidence testing. Confidence testing attempts to provide a means of determining a student's degree of confidence in his response to various tests and performance situations. How should an examinee indicate his degree of confidence when choosing responses in the face of uncertainty? Possible solutions to this problem of method of response and the corresponding scoring system have been appearing in the literature since the mid 1930's and more recently have been associated with the names of de Finetti (1965), Coombs, Milholland and Womer (1956), and Shuford, Albert, and Massengill (1966). A complete review of the literature can be found in Echternacht (1971).

In confidence testing two assumptions are usually made: the examinee must be interested in obtaining a high score, and the scoring rule must be known to the examinee. When an examinee using confidence testing encounters an item for which he is uncertain of the correct response, his answer should reflect his degree of belief (i.e., his subjective probability) about the correctness of the various alternatives. This can be accomplished in a number of ways. Although not based on subjective probability, Coombs, Milholland and Womer utilize a response method where the examinee crosses out all alternatives he believes to be false. Another system not based upon

subjective probability, developed by Ebel (1965), utilizes a five-choice true-false format. The subjective probability approach advocated by de Finetti requires the examinee to allocate five stars or points over the alternatives present in such a way as to reflect his degree of belief in the alternatives he believes possible. In another noteworthy subjective probability measure, developed by Shuford and Massengill (1969) the examinee uses a device termed SCoRule to show his degree of belief in each alternative.

In the scoring for the approaches not based on subjective probability, a somewhat arbitrary method is used, such as obtaining a certain score for each incorrect alternative crossed out and a certain penalty score for each correct alternative eliminated. The subjective probability approach utilizes a concept termed reproducibility in the development of scoring systems. Basically, a scoring system is termed reproducible if an examinee can only maximize his expected score with respect to his state of knowledge only by responding with his true subjective probabilities for each alternative. Shuford and Massengill use a logarithmic scoring function to this end, while de Finetti relies on an approximation to what he terms the continuous method.

Advocates of confidence testing believe that their procedures provide more information and yield "fairer" scores than conventional multiple-choice testing since measures of the level of student knowledge of each test item are acquired rather than a simple indication that the student was right or wrong. Instructors could thus identify the level of student knowledge, and consequently, more accurately ascertain how and what additional teaching should occur.

If, in fact, confidence testing does provide information concerning a student's level of knowledge beyond that provided by conventional multiple-choice tests, it would appear that its use in technical training courses would allow instructors to tailor course presentations to correct student weaknesses and make materials more meaningful to students, thus enhancing the training program.

For the purpose of this study feasibility was defined in terms of student course performance, student remediations, and student and instructor attitudes toward the applicability and practicality of confidence testing in the setting of technical training in the Air Force. Thus, confidence testing would be deemed feasible if students subjected to confidence testing in their diagnostic daily quizzes performed better and required fewer remediations in courses than students not so exposed. Confidence testing would also be considered feasible if students and instructors found the practice to be useful and not too time consuming.

One factor influencing the feasibility of using confidence testing in technical training was the relationship between confidence test scores and various personality variables. Swineford (1938) first demonstrated a relationship between early methods of confidence testing and examinee personality when she derived a gambling score for each examinee which was

uncorrelated with the total test score. Thus, she concluded a confidence test score was comprised of two parts, one for achievement, the other for willingness to gamble. This study attempted to reevaluate this relationship using modern methods of confidence testing after the subjects had practiced with the methods.

SECTION II

Sample

The setting for this study was the 3345th Technical School at Chanute Air Force Base, Illinois. Two courses, Aerospace Ground Equipment Repairman (AGE) and Jet Engine Mechanic (JEM), were chosen from the various courses available for participation due to the high flow of students entering these courses each week. Upon course entry students were assigned to a six-hour instructional shift in a random fashion. The AGE course was divided into four nonoverlapping shifts, while the JEM course utilized only two shifts. These shifts were designated "A, B, C, and D" in AGE and "A and B" in JEM. The instructional time of Shift A was from 0600 hours until 1200 hours; Shift B from 1200 hours until 1800 hours; Shift C from 1800 hours until 2400 hours; Shift D from 2400 hours until 0600 hours. Students entering the JEM course were further assigned to different instructors within their shift. Both of these courses were organized into a number of instructional blocks that were of either a one- or two-week time period.

Since the experimenters were primarily interested in confidence testing as applied to a multiple choice format, the daily quizzes used in each course were examined to determine a period where most daily quizzes given were multiple choice in nature. After all daily quizzes were examined, blocks two and three were selected for further study from the JEM course, while Blocks six, seven, and eight were selected from the AGE course.

All students entering these phases of the courses between October 14, 1970 and November 18, 1970 were selected as subjects in the experiment. These students, who served as subjects in the experiment, were primarily young men having recently enlisted in the Air Force. Data were collected for 434 students, 180 in AGE and 254 in JEM. The average Airman Qualifying Examination (AQE) percentile ranks were approximately 70 for those students in AGE and 60 for those students in JEM. Further details regarding the two courses under study can be found in the Plan of Instruction for Jet Engine Mechanic and Aerospace Ground Equipment Repairmen (Air Training Command, 1970, (a) (b)).

SECTION III

Design

The effects of three different methods of daily quiz testing on course performance as measured by end of block examination scores were under study in this experiment. In addition, the effects of two types of remedial treatment and the interactions of the remediation type with testing type were of interest.

Of the three methods of testing under study, two were experimental confidence procedures while the third was a control procedure. The control procedure consisted of traditional multiple-choice testing with four alternative response items.

One confidence testing procedure, termed "Pick-One", required the examinee to first choose the alternative he believed to be correct, exactly as he would in a conventional multiple-choice test, and then indicate on a five-point scale his sureness in his response. This scale ranged from "very sure", indicating complete certainty on one end, to "not sure", indicating complete ignorance on the other end. The points on the scale were designed to correspond to various subjective probability levels for the chosen alternative. A scoring scheme was devised that was reproducible as far as the probability of the response chosen was concerned, though in the present experiment confidence was rated and the reproducibility property approximated. A complete description of this technique was given by Boldt (1971). The Pick-One confidence testing method was devised for both examinee and test administrator ease. It was felt that this method was the least demanding on both the responder and the scorer. Scoring was simple as there were only nine possible scores for an item and test administrators could remember these scores after a little practice.

A second type of confidence testing used in this study, termed "Distribute 100 Points", approximated the method devised by Shuford and Massengill (1969). Using this method, the examinee was first required to choose an alternative and record that as being his selected answer. He then indicated his subjective probability of each alternative's being correct by distributing 100 points over the various alternatives. A truncated logarithmic scoring function was used. This method differed from that devised by Shuford and Massengill only in that the examinees were asked to respond directly with their subjective probability rather than use a response device such as the SCoRule. Illustrations using both the Pick-One and Distribute 100 Points methods appear in Appendix III.

Two types of remediation were used in this study. A student was assigned to a remedial session of two hours following his scheduled class if he performed unsatisfactorily on the daily quiz (usually scoring below 70 percent), had poor performance in the previous block, showed weakness in practical performance, or missed class time due to sickness or leave. In each case the assignment of a student to a remediation session was left to the discretion

of the individual instructor. One type of remediation was the standard or control remediation procedure in use at the technical school. A special remediation was devised as an alternative method. This method was based on the notion that students responding incorrectly with high confidence should receive a different type of instruction than students responding incorrectly with low confidence. Students who were misinformed (wrong answer with high confidence) would go through a two stage remedial process, first being instructed as to why their responses were wrong and then why the correct answer was, in fact, correct. Students who were simply not informed (wrong answer with low confidence) would go through only a single stage remedial process, being instructed as to why the correct answer was, in fact, correct. In this manner, an initial step could be taken to allow instructors to tailor their remedial instruction to the needs of the students. Additional discussion of the remedial procedures can be found in Appendix I.

The two factors, method of testing and method of remediation, when taken in combination, produced six treatment combinations. These six treatment combinations were then assigned in a random order within each instructional shift to six classes as they entered the appropriate blocks under study. Once a particular class entered the experiment and was assigned a particular method of testing and remediation, it continued use of only that combination until it concluded its part of the experiment. In the JEM course where an entering class was subdivided and assigned to various instructors, everyone in that entering class received the same treatment combination regardless of his instructor, and continued using that treatment combination even though the composition of subjects assigned to an instructor within a shift changed from block to block.

The scheduling and the assignment of the various treatment combinations appears in Figures 1 and 2. The rows represent the weeks of the experiment. The columns denoted T_i, i=1, 2, ..., 6, represent the particular testing treatment combination used by the class entering the experiment at the jth week in one of the shifts. The types of testing were coded as: Multiple choice (Mult Ch), Pick-One (Pick-One), and Distribute 100 Points (Dist 100). Thus, the class entering Block 6 of AGE in the first week of the experiment in Shift A used Distribute 100 Points confidence testing with special remediations until they completed Block 8. Similarly, the class entering Block 6 of AGE in Shift B during the third week of the experiment used Pick-One confidence testing with special remediation. The assignment of the treatment combinations to the T_j was accomplished independently for each shift in each course using a table of random numbers.

Two pieces of data were collected for each student as he completed participation in the experiment: final end of block examination scores, three for AGE and two for JEM, and the number times each student was assigned to remediation. These records were obtained from the technical school student files.

Since Swineford (1938) was able to derive a gambling score from confidence responses orthogonal to the total test score, a secondary consideration undertaken in this study was a study of how various personality factors affected the

Figure 1. Schedule for Aerospace Ground Equipment Repairman

		T_1	T_2	T_3	T_4	T_5	T_6
	1	B 1 o					
	2	c k 6	B 1 o				
	3	Block 7	c k 6	B 1 o			
W	4	B 1 o	Block 7	c k 6	B 1 o		
E	5	c k 8	B 1 o	Block 7	c k 6	B 1 o	
E	6		c k 8	B 1 o	Block 7	c k 6	B 1 o
K	7			c k 8	B 1 o	Block 7	c k 6
	8				c k 8	B 1 o	Block 7
	9					c k 8	B 1 o
	10						c k 8

	Shift A		Shift B		Shift C		Shift D
T_1	Dist 100 Special	T_1	Mult Ch Special	T_1	Pick-One Control	T_1	Dist 100 Control
T_2	Dist 100 Control	T_2	Dist 100 Control	T_2	Dist 100 Special	T_2	Mult Ch Special
T_3	Pick-One Special	T_3	Pick-One Special	T_3	Mult Ch Special	T_3	Pick-One Control
T_4	Pick-One Control	T_4	Dist 100 Special	T_4	Mult Ch Control	T_4	Pick-One Special
T_5	Mult Ch Control	T_5	Mult Ch Control	T_5	Dist 100 Control	T_5	Mult Ch Control
T_6	Mult Ch Special	T_6	Pick-One Control	T_6	Pick-One Special	T_6	Dist 100 Special

Figure 2. Schedule for Jet Engine Mechanics

WEEK	T_1	T_2	T_3	T_4	T_5	T_6
1	B l o					
2	c k 2	B l o				
3	B l o	c k 2	B l o			
4	c k 3	B l o	c k 2	B l o		
5		c k 3	B l o	c k 2	B l o	
6			c k 3	B l o	c k 2	B l o
7				c k 3	B l o	c k 2
8					c k 3	B l o
9						c k 3

	Shift A		Shift B
T_1	Pick-One Special	T_1	Mult Ch Control
T_2	Mult Ch Special	T_2	Mult Ch Special
T_3	Pick-One Control	T_3	Pick-One Special
T_4	Dist 100 Control	T_4	Pick-One Control
T_5	Dist 100 Special	T_5	Dist 100 Special
T_6	Mult Ch Control	T_6	Dist 100 Control

confidence responses of the subjects taking the confidence tests. As of now, though, few studies have been made of the effects of personality on modern confidence response procedures. In order to examine the relationship between personality factors and confidence responses, a battery of personality tests was developed and administered to each class as it entered the experiment. This battery included the Clayton and Jackson F-scale (1961), Rokeach's Dogmatism scale (1956), the Alpert-Haber facilitating and debilitating test anxiety scales (1960), the Gough-Sanford rigidity scale (1957), Barratts' Impulsiveness scale (1959), and a self-sufficiency test developed previously by Educational Testing Service. In addition, modifications of Kogan and Wallach's (1964) risk-taking tests were used yielding five scores for five different gaming strategies. These tests may be found in Appendix II.

Each instructor was asked to return to the experimenters the last three multiple-choice daily quizzes for the instructional block he taught. For classes subjected to confidence testing, tests were scored using both the standard rights only scoring method and the appropriate scoring method for confidence testing. In addition, an attempt was made to obtain AQE General test scores as a measure of verbal ability. However, many of these test scores could not be found in either the technical school records for the two courses under study or from the personnel files of the Air Force Human Resources Laboratory, Personnel Division, Lackland Air Force Base. Therefore, this information was not used in the study of the personality variables.

At the conclusion of a class participation in the experiment, each student was given a questionnaire concerning his attitudes toward the course testing, in general, that takes place in the Air Force. Students who were subjected to confidence testing were also asked about the difficulties they experienced, their attitude toward the confidence tests, their response styles, and the method's aid during remediation and review. Some questions were common to all types of testing, some questions were common to both types of confidence testing, and some were specific to each type of testing. These questionnaires can be found in Appendix IV.

As part of a cost analysis, instructors were given a questionnaire which asked, among other things, how useful were the confidence responses, how much time was required for administration, and what their attitudes were toward the testing process in general. Also, instructors were asked to keep a log indicating the test administration time, number of answer sheets scored, and the test correction time for each type of testing they encountered. From this information, scoring time per answer sheet was calculated for confidence and multiple choice procedures. This questionnaire can be found in Appendix VII.

It was hypothesized that: (1) individuals subjected to confidence testing would have higher block examination scores than subjects using regular multiple-choice testing; (2) all block examination scores would respond in a similar manner to testing conditions since there was no reason to expect that blocks would respond differentially; (3) subjects using confidence testing and having the special remediation sessions would obtain higher block examination scores than the other combinations of testing and remediation type ; (4) confidence testing would require fewer remediations than multiple-choice.

With regard to the intercorrelations of the personality tests, it was hypothesized (1) that some unspecified personality variables would correlate significantly with the confidence scores, (2) that when the multiple-choice test variables were partialled out, significant correlations would remain between the same unspecified personality variables and the confidence test scores, and (3) that the confidence test score would correlate approximately one with the multiple-choice test score.

the standard deviations of the respective end of block examination scores and were used for the purpose of comparing group means.

Table 1

Correlations of the End of Block Examination Scores in AGE with Standard Deviations on the Diagonal

Variable	Block 6 Score	Block 7 Score	Block 8 Score
Block 6 Score	6.697	.399	.536
Block 7 Score		8.993	.534
Block 8 Score			7.665

End of block examination scores were recorded in terms of percent of items correct, and the standard deviations were in terms of percentage points. The correlations were all significantly nonzero with the correlation between Block 8 scores and Blocks 6 and 7 scores being roughly the same. The correlation between the Block 6 scores and Block 7 scores appeared to be slightly lower.

When the shift by testing type interaction was tested for significance, two of the three discriminant functions available were found significant with probabilities less than .05. This was interpreted to mean that the testing type effect depended upon the particular shift, block, and type of testing that was being examined; hence, no overall main effects were examined for AGE. In order to better understand this interaction, univariate one-way analyses were performed within each shift, with the testing type effects being calculated in each case. From this point on, the discussion of the analysis of end of block examination scores in AGE will be presented shift by shift.

The analysis for shift A indicated that there were no significant differences between the types of testing. Thus, it was concluded that in shift A each of the types of testing on daily quizzes rendered the same result.

In the analysis of testing type within shift B, one significant discriminant function was found. An examination of the univariate F-ratios indicated that a significant difference occurred only in Block 7. The effects, in terms of deviation of means, of the various types of testing are given in Table 2. The group using multiple-choice testing had the lowest average block examination score, while Distribute 100 Points had the highest. The difference between the multiple-choice mean and the Distribute 100 Points mean was 11.375, which was slightly more than one standard deviation and was considered to be large. No significant difference between the types of testing was found for the remaining blocks.

SECTION IV

Results

As stated earlier, the various treatment combinations were assigned at random to classes as they entered the experiment. Due to the large number of instructors involved in the experiment in relation to the number of students, it was necessary to confound the instructor variable. A description of each of the univariate analyses used in this section can be found in Fisher (1958). The following sections report the results of the analyses of (1) the end of block examination scores; (2) the number of remedial sessions; (3) student attitudes; (4) instructor attitudes, (5) personality variables, and, (6) the daily quiz administration and scoring time. This process was carried out identically for every shift in both AGE and JEM.

Type of Testing and End of Block Scores

It was assumed that students were randomly assigned to their respective shifts allowing a three-way factorial analysis with the independent variables. The three factors used as independent variables were type of testing, type of remediation, and shift. The dependent variables used in this analysis were the respective end of block examination scores earned while in the experiment. Since these dependent variables were correlated, a multivariate analysis of variance was used. The data and detailed analyses, including the univariate F-ratios, appear in Appendix V. Detailed expositions of multivariate analyses of variance can be found in Rao (1952), Morrison (1967), and Pruzek (1971).

In the AGE course three end of block examination scores served as criteria in the analysis. Under normal circumstances a three-way design incorporating three types of testing, two remediation types, and four shifts would utilize 24 cells in a factorial design. However, in the present case only 19 cells contained data, and two of these contained very little data. These missing cells resulted from some administrative confusion by the instructors involved with respect to the data collection system. In order to overcome this missing data problem, the factor of remediation type was deleted as a main effect and relegated to a nesting variable in a two-way design. This approach was deemed feasible since the instructors had indicated, informally, that the special remediation procedures were infrequently used. Since the procedures were used infrequently, it was assumed that the remediation type effect was negligible. Thus, a two-way factorial layout was conceptualized, with type of testing and shift serving as factors and type of remediation used to nest classes within the various treatment combinations. When this adjustment was performed, no empty cells remained.

As a part of the analysis, the correlations among the end of block examination scores within the error term of the design were calculated. These estimates appear in Table 1 and were the proper within-cell estimates of the population correlation coefficients. The values on the diagonal represent

Table 2

Effects (group mean - grand mean) of the Types of Testing in Block 7 of Shift B in AGE

Type of Testing	Effect
Multiple-Choice	-6.667
Pick-One	1.958
Distribute 100 Points	4.708

When shift C was analyzed, one discriminant function was found to be significant. Univariate analyses on the Block 6, 7, and 8 scores produced significant F-ratios on only the Block 6 scores. Table 3, indicates that multiple-choice testing was definitely the least effective method in this

Table 3

Effects (group mean - grand mean) of the Types of Testing in Block 6 of Shift C in AGE

Type of Testing	Effect
Multiple-Choice	-5.354
Pick-One	.521
Distribute 100 Points	4.833

block, while Distribute 100 Points was the most effective method. The difference between the means for multiple choice and Distribute 100 Points was 10.187, which is about one and one-half standard deviations. Such a difference was considered very large and substantial. Since no significant differences were found in Blocks 7 and 8, the testing types were concluded to be equally effective in these blocks.

No significant differences between testing types were found in shift D. Thus there was no particular advantage in using any of the types of testing on the daily quizzes in this shift.

In summary, of the four shifts and three blocks, twelve analyses in all, only two analyses resulted in significant types of testing differences. In

both of these cases multiple-choice testing had the lowest mean block examination score while Distribute 100 Points had the highest mean block examination score.

In the JEM course two variables served as criteria in a multivariate analysis of variance. These two variables were end of block examination scores received while in instructional Blocks 2 and 3. The design called for a three-way factorial with type of testing, type of remediation, and shift as factors. There were three types of testing as before, two types of remediation, and two shifts, designated A and B. An initial examination of the data showed that all 12 cells of the design contained data. The number of observations was not uniform, as two cells had 6 and 8 subjects, while the remaining cells contained between 15 and 32 observations.

The correlation between Block 2 and Block 3 examination scores was estimated to be .547, which was slightly higher than the estimates obtained in AGE. The estimated standard deviations were 6.883 and 6.096 for Blocks 2 and 3 respectively.

When the testing type by remediation type by shift interaction was tested, one discriminant function was found to be significant. In order to better understand this interaction, the analysis was divided so that the types of testing could be examined within the four combinations of shift and remediation type. From this point on the analysis will be discussed by these four groups.

One significant discriminant function was found when the types of testing were considered in Shift A for classes using special remediation. An examination of the univariate F-ratios yielded significant differences in both blocks. As this was an unusual finding in multivariate analysis of variance, an examination of additional statistics was undertaken in order to interpret this result. The correlations between the discriminant variable, the appropriate linear combination of block scores, and the block scores were examined. The correlation between the discriminant variable and Block 2 scores was found to be .997 indicating that the two were identical for all practical purposes. The correlation between the discriminant variable and Block 3 scores was found to be .61, which was low since this correlation could not be less than .547, the correlation between the two block scores. Thus, it appeared that the correlation between the block scores resulted in the univariate estimate of the effect of the type of testing that was found in Block 3.

The effects of the various types of testing are given in Table 4. It was apparent that multiple-choice testing was again low, while Distribute 100 Points testing was about two-thirds of a standard deviation higher. This difference was notable but not as large as those previously reported. Pick-One testing, on the other hand, differed greatly from multiple-choice testing in that the difference was about two standard deviations.

Table 4

Effects (group mean - grand mean) of the Types of Testing in Block 2 of Shift A Using Special Remediation in JEM

Type of Testing	Effect
Multiple-Choice	-5.618
Pick-One	6.942
Distribute 100 Points	-1.324

Two significant discriminant functions were found when the types of testing were examined within Shift A when control remediation was used. Also, the univariate F-ratios were significant for each instructional block. Thus it was concluded that there were significant testing type effects in each block. These effects are presented in Table 5. In Block 2 multiple-choice testing was again low, with Distribute 100 Points being about three-fourths of a standard deviation higher and Pick-One being about one standard deviation higher. In Block 3, however, both multiple choice and Pick-One were low while Distribute 100 Points was approximately one standard deviation higher than multiple choice.

Table 5

Effects (group mean - grand mean) of the Types of Testing in Shift A Using Control Remediation in JEM

Type of Testing	Block 2	Block 3
Multiple-Choice	-4.028	-2.225
Pick-One	2.741	-1.032
Distribute 100 Points	1.287	3.257

When the types of testing were analyzed in Shift B for classes using special remediation, two discriminant functions were found to be significant. When the univariate F-ratios were examined, significant F's were found in both blocks. The estimates of the treatment effects are given in Table 6. In each case the mean block grade was lowest for the group using multiple choice testing. Table 6 indicates that the difference between the means for the group using multiple-choice testing and either confidence procedure was about one standard deviation. There seemed little to choose between the Pick-One and Distribute 100 Points methods in Block 2, while the Pick-One method appeared to be superior to the Distribute 100 Points method in Block 3. The difference

between the multiple-choice and Distribute 100 Points methods was about one standard deviation while the difference between the multiple-choice and Pick-One methods was about two standard deviations and therefore quite outstanding.

Table 6

Effects (group mean - grand mean) of the Types of Testing in Shift B Using Special Remediation in JEM

Type of Testing	Block 2	Block 3
Multiple-Choice	-4.564	-5.561
Pick-One	2.191	5.466
Distribute 100 Points	2.373	.095

No significant testing type differences were found in Shift B when the control remediation type was used. Thus, it was concluded that there was no difference in the block scores for the groups using the three types of testing in Shift B when the control remediation was used.

In summarizing the results of the analysis of the end of block scores for the various types of testing used in this experiment, one conclusion stands out. Multiple-choice testing consistently resulted in the lowest block scores when compared to Pick-One and Distribute 100 Points confidence testing. There was some question whether Pick-One or Distribute 100 Points was superior as that seemed to depend upon the particular shift and type of remediation. The analyses did seem to favor the Pick-One for the JEM course as that type of testing appeared to be more often superior to Distribute 100 Points over all shifts and remediation types where significance was found.

Type of Testing and Number of Student Remediations

Another criterion for feasibility was the number of remediations required for each student using the various types of testing under study. If confidence testing could reduce the number of remediations required, it would be beneficial to technical training. Therefore, the number of remediations each student required was recorded for each block in each course used in the experiment.

The average number of remediations per student was calculated for each group defined by type of testing, shift, and block of instruction. Using these data, there appeared to be no appropriate statistical test for assessing the significance of any type of testing differences since there appeared to be no error term. The individual student data could not be used as a dependent variable in an analysis of variance since the distribution of these data were Poisson rather than normal. Therefore, interpretation was based upon the consistency of the rankings of testing methods with respect to average number

of remediations within block and shift. In AGE there were 12 such block-shift combinations, while in JEM there were four. The average number of remediations per student are presented in Table 7 for AGE and in Table 8 for JEM.

In the AGE course, students who used multiple-choice testing required more remediations on the average than did students using either of the confidence testing procedures in 9 of the 12 shift-block combinations. The difference between the two confidence methods was slight; the two methods were the same with respect to the number of times they were ranked lowest in the average number of remediations. Thus, in AGE, confidence testing appeared to result in a reduction of the total number of remediations required. The differences between Pick-One and Distribute 100 Points appeared to be slight.

Table 7

Average Number of Remediations Per Student in AGE

	Shift A	Shift B	Shift C	Shift D
Block 6				
Multiple-Choice	1.4	2.5	2.3	1.1
Pick-One	.4	1.5	.5	1.1
Distribute 100 Points	.1	1.5	.3	1.1
Block 7				
Multiple-Choice	.7	0.0	.8	1.4
Pick-One	.6	1.5	.2	.9
Distribute 100 Points	.1	1.6	.3	.2
Block 8				
Multiple-Choice	1.1	1.9	1.8	.5
Pick-One	0.0	.6	1.0	1.0
Distribute 100 Points	.5	.9	1.0	.3

In JEM similar results were found. Table 8 shows that students subjected to multiple-choice testing required more remediations than either confidence method in three out of four shift block combinations. As in AGE, there seemed to be little difference in the Pick-One and Distribute 100 Points confidence testing with respect to the average number of remediations.

Thus, it was concluded that either Pick-One or Distribute 100 Points reduced the number of remediations required when compared with multiple-choice testing. It should be noted that this conclusion was based on judgment rather than objective testing and therefore should be taken with caution.

Table 8

Average Number of Remediations Per Shift in JEM

	Shift A	Shift B
Block 2		
Multiple-Choice	1.0	.8
Pick-One	.5	.4
Distribute 100 Points	.8	.3
Block 3		
Multiple-Choice	2.2	.7
Pick-One	.1	.7
Distribute 100 Points	.1	.6

Student Attitudes

A third feasibility criterion was that of student attitudes towards confidence testing. If students favored one of the confidence methods over the multiple-choice method to a significant degree, the process would be considered feasible even though there were no real significant gains in student achievement. At the conclusion of the subjects' participation in the experiment, each subject was administered an attitude questionnaire that asked him about the testing he had encountered thus far in the Air Force. In addition, subjects using one of the confidence testing methods in the experiment were asked about their testing behavior when using it as well as their evaluation of the process.

Subjects were asked to respond to the questions on a five-point scale where two categories represented positive responses to the item, one a neutral response, and two categories a negative response. The student attitude questionnaire is shown in Appendix IV. Thus, although the responses were based on a five-point scale, the responses could be classified into three categories in order to compensate for low frequencies in response categories. Two-way contingency tables were constructed for each item in the questionnaire, with type of testing and response category serving as the two classification variables. The resulting data were analyzed by the use of the chi-square statistic. First, all five categories of response were used, but when an expected cell frequency was less than five, the number of response categories was reduced to three by pooling the two positive response categories and the two negative response categories. Where an expected cell frequency remained less than five, the response classification corresponding to that cell was deleted.

The results are divided into two parts. The first presentation covers the attitude items answered by all subjects in the experiment. The second presentation is for the items answered by those students taking either Pick-One

or Distribute 100 Points confidence testing. The complete frequencies of response are given in Appendix VI with the items where significant differences were found noted.

Items common to each type of testing. In AGE three items out of six items were found to have significantly different response patterns for the three types of testing. Students using multiple-choice testing indicated; (1) more satisfaction with the testing they had been using, and (2) that their testing more satisfactorily demonstrated what they really knew. Students using Distribute 100 Points testing, on the other hand, expressed less satisfaction than under null conditions. When asked how satisfied they were with the testing that takes place in their classroom, the group using Distribute 100 Points testing indicated more satisfaction than expected and the multiple-choice group expressed less satisfaction. In each case the group using Pick-One testing answered with frequencies close to those expected under the null hypothesis. No significant chi-squares were found with questions pertaining to the advantages of classroom testing, satisfaction with testing as an aid to remediation, and test results as an aid to remediation.

A different pattern characterized the JEM course. Significance was found in two of six items, only one of which was found to be significant in AGE. These items dealt with one's satisfaction with the classroom testing used and the advantages of classroom testing. In both cases the group using Pick-One testing indicated a more favorable attitude toward the testing they had used than expected under the null hypothesis.

When the results for these two courses are put together, there seemed to be little basis for recommending any procedure over another. In JEM the Pick-One type of testing seemed to be favored, but in AGE no such preference can be seen as there was no clear-cut method standing out in AGE. Since no one method of testing appeared to be highly regarded in both courses, no clear-cut conclusions can be made.

One variable confounded in the analysis was that of the students' instructors. It was not known whether students were reacting to the type of testing they had been using or whether they were responding to their instructors' teaching of their classes.

Items common to the two types of confidence testing. Twenty-five questions were asked of students using the two types of confidence testing. These questions dealt with how the students responded in the face of uncertainty, their ease in marking their answers, and their evaluation of confidence testing as compared to multiple-choice testing.

In the AGE course no significant chi-square values were found, indicating that students subjected to the two types of confidence testing under study responded similarly to the items. In general, students subjected to confidence

testing felt it was important to score high on the daily quizzes, understood how the tests were graded (they were not explicitly told), were only fairly accurate in marking their confidence, felt comfortable with the procedure, and tended to be only neutral or slightly positive in their evaluation of confidence testing as compared to multiple-choice testing. Further, students acknowledged that confidence testing required greater thought before answering. Students using Distribute 100 Points testing expressed little difficulty in distributing their points in such a way that they summed to 100.

While no significant chi-squares were found in AGE, five items were found to have significant chi-squares in JEM. Students using Pick-One testing indicated they understood how the tests were graded and felt more comfortable with the procedure to a greater extent than students using Distribute 100 Points testing. Students using Distribute 100 Points testing believed that their testing identified a useful level of knowledge, required more thought before responding, and was a useful device for relearning material to a greater degree than did the students using Pick-One testing.

As in the AGE course, students in JEM using confidence testing indicated they thought it was important to score high on the daily quizzes, felt they understood the ways of marking their answers very well, were only fairly accurate in marking their confidence when uncertain, found it easy to make a decision on how to mark their confidence, tended to gamble sometimes in marking their confidence, and tended to be favorable to confidence testing as compared to multiple-choice testing. Students using Distribute 100 Points testing indicated little difficulty in allocating the 100 points in such a way that they summed to 100.

In summary, no one method emerged as favored over both courses. Since results were mixed, with each of the types of testing showing some promise in various situations, interpretation was difficult. Therefore, the only conclusion drawn from the student attitude questionnaire was that no method was preferred by the students over any other.

Instructor Questionnaires

A questionnaire was given to each instructor who taught a class that utilized a confidence testing procedure. This questionnaire contained open-ended questions, which were completed by 37 instructors. The answers given by the instructors were categorized into broad categories by the experimenters, and specific comments were noted. The questions, along with the frequencies, can be found in Appendix VII.

Typically, an instructor noted only a few students who placed large amounts of confidence on wrong alternatives, even though most students responded with the highest confidence marks. This result may have come about as most students tend to score high on daily quizzes, 70 percent being the passing mark. One instructor noted that students usually scoring low tended to have more variation in the confidence attributed to the chosen alternative.

Most instructors were not influenced by the confidence responses and did not utilize them in remediation. Instructors who did attempt to use the confidence scores stated they used them only for identifying the lowest ability students, so that they could concentrate on these students. The instructors expressed difficulty in scoring, noting especially the lengthy time required.

The instructors felt the students handled the testing situation easily, in that there seemed to be sufficient time available and the method for assigning confidence was not deemed difficult. One instructor noted that difficulty in assigning confidence occurred only at the beginning of the experiment, with the least competent students having the most difficulty. Some instructors reported that their students found the procedure easy because they usually assigned the highest confidence marks to every question.

One concluding remark was furnished by an instructor who typified the instructor's attitudes; he stated, "It reminds me of using a bulldozer to clear snow from a sidewalk--it's too good. Percentages work well enough for our tasks."

Personality Variables as Related to Confidence Testing

As previously stated, it was desired to relate certain personality variables to confidence test scores. Personality variables of interest were dogmatism (DOG), authoritarianism (AUTH), facilitating anxiety (FAS), debilitating anxiety (DAS), rigidity (RIGID), impulsiveness (IMP), self-sufficiency (PRI), and risk-taking. Five betting strategies--maximum gain (MG), minimum loss (ML), long shot (LS), maximum variance (MB), minimum deviation from one-half (HALF MD)--were taken as measures of risk-taking. The risk-taking measures were modeled after those found in Kogan and Wallach's chance bets instrument. Basically, this test consisted of 36 randomly ordered pairs representing all possible combinations of three probabilities of winning (1/3, 1/2, 2/3) and three stakes (15¢, 30¢, 60¢). All bets were of zero expected value. The five strategy indexes had three different bases, two based on a monetary amount, two based on probabilities, and one on a combination of money and probability. The maximum gain strategy involved choosing that alternative with the larger potential winnings, the minimum loss strategy involved choosing that alternative with the smaller loss potential, the long shot strategy involved choosing that alternative with the lower probability of winning, the minimum deviation from one-half strategy involved choosing that alternative with probability of winning that was closer to one-half, and the maximum variance strategy involved choosing that alternative with the greater variance. Each of these personality tests were administered to each subject as he entered the experimental phase of the course.

Each instructor was asked to return to the investigators all answer sheets for the last three daily quizzes administered as confidence tests in his instructional block. Counts of the number of answer sheets returned were obtained for each daily quiz in each instructional block and each type of confidence testing. In AGE a sufficient number of subjects could not be found

for any particular daily quiz since each shift used a different set of quizzes for the same subject matter. Therefore, analysis of personality variables was confined to the JEM course. Two daily quizzes were found in JEM for each type of testing with a sufficiently large sample size to merit further analysis. The respective numbers of students taking these daily quizzes were 106 and 105 for the two quizzes chosen in the Pick-One case and 83 and 73 in the Distribute 100 Points case.

Each of the daily quizzes under consideration was scored in two ways: (1) a count of the total number of items answered correctly, termed the rights-only score, was made; (2) a confidence score based on the specific scoring function for confidence testing was obtained. Preliminary correlations were calculated among the rights-only score, the confidence score, and the shift in which the instruction took place. Shift was used as a variable in this analysis since it was suspected to be correlated with the test scores. Such correlation matrices were calculated for each of the four tests under consideration. The results of these calculations indicated that there was a significant association between the rights-only score and the shift, thus implying the need to remove or partial the shift variable from any further correlations.

The correlations between the confidence test scores (CON), the rights-only score (RIGHTS), and the various personality variables are given in Tables 9, 11, 13, and 15. The correlations between confidence test scores and the personality variables having partialled the rights-only score are given in Tables 10, 12, 14, and 16. These correlations were the proper within shift estimates of the population correlations. Only correlations significant at the .05 level of significance are reported.

From Tables 9 and 11, it should be noted that for daily quizzes administered to those students using Pick-One confidence testing, the two ways of scoring the test correlated .92 and .89. These correlations are extremely high and indicate that it makes little difference how the tests are scored. Also, it may be seen by comparing Tables 9 and 11 with Tables 10 and 12, that the intercorrelations among the personality variables tend to remain stable whether or not the rights-only score was partialled out. This occurred in part because the correlations between rights-only and the personality scores were small in absolute value. However, the correlation between rights-only and the confidence score was sizeable, and partialling out the rights-only score affected the correlations of confidence scores with the personality variables. The resulting partial correlations are quite different when the results of Quiz 1 are compared with those of Quiz 2. Hence, unless there is some crucial difference between the two quizzes that produces these differences one would tend to attribute them to randomness in the system. It should be noted in this connection that many coefficients were generated and compared on these data, and some apparently significant results are likely to appear by chance. The significance tests are probably much less conservative than usual and it would not be uncommon to find "significant" relations that were only apparent in this situation. Note, for example, that authoritarianism correlated significantly with the confidence score for both quizzes but with opposite signs. The significance of this correlation helps substantiate the hypothesis that

assigning high confidence is associated with some personality trait. However, the fact that the correlation coefficients were opposite in sign conflicted with this notion.

Another approach to studying the relation between personality and confidence score was to estimate the contribution that personality test scores make to the prediction of confidence test scores over and above that made by shift and by the number right score. Shift was included as a predictor because of the possibility of events taking place in the shift which have an effect on the confidence score. The number right was included because it was supposed that the value of the confidence score lies in the fact that its deviations from the number right are meaningful. Clearly, meaningful variation of the confidence scores that was shared with the number right did not help one decide between them. Hence, the variation of interest is that which was not shared with the number right and further should not occur merely because of temperamental or personality dispositions. Thus, analyses were performed wherein the residual variation after the confidence test scores were predicted, using the personality variables, shift, and number right, was compared with the residual variation after only shift and the number right was used on the predictor side. The residual variation was larger when the fewer number of predictor variables was used, but not large enough. The test of the size of the residual variation was made with an F-ratio which takes into account the spurious accuracy achieved by fitting more predictors, and the test indicated that the additional accuracy added by the personality variables was well within that which might be expected by chance. Thus, it was concluded that personality variables were not related to the confidence score and that the scoring systems under study were so similar that it seems to make little difference which scoring system is used.

The results for the quizzes in the Distribute 100 Points format was similar to that of the Pick-One format. The estimates of the population correlations were given in Tables 13 and 15 and with the rights-only score partialled out in Tables 14 and 16. Although the intercorrelations among the personality variables appear to be stable over the two quizzes, the intercorrelations of the personality variables with the confidence scores were unstable when the rights-only score was partialled out. In order to test simultaneously the significance of the correlations between personality and confidence score for each daily quiz, multiple regression analyses were performed for each daily quiz, using the personality variables as predictors and the confidence test score as the dependent variable with shift and rights-only score serving as covariates. In each case, F-ratios were found to be not significant. Thus, it was concluded that the significant first order correlations obtained were a result of randomness.

In summary, no personality variables were found to relate to confidence test score when the influence of rights-only score was removed. The rights-only and the confidence test scores in both the Pick-One and Distribute 100 Points formats were found to be so highly related that there appeared to be no practical difference in the scoring systems other than changing the scale of measurement.

Table 9

Correlation Matrix for Personality Variables, Rights-Only, and Pick-One Confidence Test Score on Quiz 1

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP	PRI
DOG	*												
AUTH	*	.33											
MG	*	*	*										
ML	*	.15	*	-.65									
LS	*	*	.14	.87	-.25								
HALF MD	.17	*	*	-.17	.14	-.16							
MV	-.15	*	*	.66	-.52	.54	-.80						
DAS	*	*	*	.16	-.19	*	*	*					
FAS	*	.26	*	*	*	*	*	*	-.56				
RIGID	*	*	*	*	*	*	*	*	*	*			
IMP	*	*	.15	*	-.14	*	*	*	.27	-.14	*		
PRI	*	.18	*	*	*	*	*	*	*	.26	.16	.17	
RIGHTS	.92	*	*	-.16	.15	*	.17	-.17	*	*	*	*	*

*indicates $r < .137$.

Table 10

Correlation Matrix for Personality Variables and Pick-One Confidence Test Score with Rights-Only Test Score Partialled on Quiz 1

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP
DOG	.27											
AUTH	.15	.33										
MG	*	*	*									
ML	*	.15	*	-.65								
LS	*	*	.14	.87	-.24							
HALF MD	*	*	*	-.15	*	-.15						
MV	*	*	*	.65	-.51	.54	-.79					
DAS	*	*	*	.16	-.19	*	*	*				
FAS	*	.26	*	*	*	*	*	*	-.56			
RIGID	*	*	*	*	*	*	*	*	*	*		
IMP	*	*	.15	*	*	*	*	*	.27	-.15	*	
PRI	-.15	.18	*	*	*	*	*	*	*	.26	.17	.16

*indicates $r < .137$.

Table 11

Correlation Matrix for Personality Variables, Rights-Only,
and Pick-One Confidence Test Score on Quiz 2

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP	PRI
DOG	*												
AUTH	-.24	.32											
MG	*	*	*										
ML	*	.16	*	-.64									
LS	*	*	*	.87	-.24								
HALF MD	*	*	*	-.17	.14	-.16							
MV	*	*	*	.65	-.51	.54	-.80						
DAS	*	*	*	*	-.15	*	*	*					
FAS	*	.26	.17	*	*	*	*	*	-.49				
RIGID	-.14	.15	*	*	*	*	*	*	*	*			
IMP	*	*	.14	*	-.14	*	*	*	.27	-.17	*		
PRI	*	.18	*	*	*	*	*	*	*	.26	.15	.17	
RIGHTS	.89	*	-.17	*	*	*	.17	*	*	*	*	*	*

*indicates $r < .139$.

Table 12

Correlation Matrix for Personality Variables and Pick-One Confidence Test Score with Rights-Only Test Score Partialled on Quiz 2

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP
DOG	*											
AUTH	-.19	.32										
MG	*	*	*									
ML	*	.15	*	-.65								
LS	*	*	*	.88	-.25							
HALF MD	*	*	*	-.18	.16	-.15						
MV	.15	*	*	.66	-.53	.54	-.80					
DAS	.17	*	*	*	-.15	*	*	*				
FAS	*	.27	.18	*	*	*	*	*	-.49			
RIGID	*	.15	*	*	*	*	*	*	*	*		
IMP	*	*	.15	*	-.14	*	*	*	.27	-.17	*	
PRI	*	.19	*	*	*	*	*	*	*	.26	.15	.17

*indicates $r < .139$.

Table 13

Correlation Matrix for Personality Variables, Rights-Only, and Distribute 100 Points Confidence Test Score on Quiz 1

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP	PRI
DOG	*												
AUTH	.19	*											
MG	*	*	*										
ML	*	*	*	-.60									
LS	*	*	-.23	.84	*								
HALF MD	*	*	*	*	*	-.24							
MV	*	*	*	.67	-.37	.61	-.71						
DAS	*	*	*	*	.24	*	*	*					
FAS	-.17	.17	*	*	*	*	*	*	-.65				
RIGID	*	.21	*	*	.18	*	.29	-.30	*	*			
IMP	*	.21	*	*	*	*	*	*	.17	-.21	.25		
PRI	*	*	.17	*	*	*	*	*	*	*	.30	.54	
RIGHTS	.99	*	.20	*	*	*	*	*	*	*	*	*	*

*indicates $r < .166$.

Table 14

Correlation Matrix for Personality Variables and Distribute 100 Points Confidence Test Score with Rights-Only Test Score Partialled on Quiz 1

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP
DOG	.25											
AUTH	*	*										
MG	*	*	*									
ML	*	*	*	-.60								
LS	*	*	-.21	.84	*							
HALF MD	-.26	*	*	*	*	-.24						
MV	*	*	*	.67	-.37	.60	-.71					
DAS	.19	*	*	*	.24	*	*	*				
FAS	-.21	.18	*	*	*	*	*	*	-.64			
RIGID	*	.21	*	*	.18	*	.29	-.30	*	*		
IMP	*	.21	.18	*	*	*	*	*	.18	-.22	.25	
PRI	*	*	*	*	*	*	*	*	*	*	.30	.55

*indicates $r < .166$.

Table 15

Correlation Matrix for Personality Variables, Rights-Only,
and Distribute 100 Points Confidence Test Score on Quiz 2

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP	PRI
DOG	*												
AUTH	*	.21											
MG	*	*	*										
ML	*	*	*	-.57									
LS	*	*	-.16	.84	*								
HALF MD	-.18	*	*	-.16	*	-.26							
MV	*	*	*	.67	-.39	.58	-.73						
DAS	*	*	.29	*	*	*	*	*					
FAS	*	.33	.18	*	*	*	*	*	-.66				
RIGID	*	*	*	*	*	*	.25	-.24	*	*			
IMP	*	*	*	*	*	*	*	*	.20	-.30	*		
PRI	*	*	.18	*	*	*	*	*	-.16	.16	.32	.44	
RIGHTS	.92	*	*	*	*	*	-.18	*		*	*	*	*

*indicates $r < .155$.

Table 16

Correlation Matrix for Personality Variables and Distribute 100 Points Confidence Test Score with Rights-Only Test Score Partialled on Quiz 2

	CON	DOG	AUTH	MG	ML	LS	HALF MD	MV	DAS	FAS	RIGID	IMP
DOG	-.21											
AUTH	*	.21										
MG	*	*	*									
ML	*	*	*	-.57								
LS	*	*	-.16	.84	*							
HALF MD	*	*	*	-.16	*	-.26						
MV	*	*	*	.67	-.40	.58	-.73					
DAS	*	*	-.29	*	*	*	*	*				
FAS	*	.33	.18	*	*	*	*	*	-.66			
RIGID	*	*	*	*	*	*	.25	-.24	*	*		
IMP	*	*	*	*	*	*	*	*	.21	-.30	*	
PRI	*	*	.18	*	*	*	*	*	-.16	.16	.32	.45

*indicates r < .155.

Quiz Administrative Time and Scoring

Each instructor involved in the experiment was asked to keep a log indicating the length of time required to administer and score the daily quizzes given during his instructional period. In general, instructors using conventional multiple-choice testing kept records for every quiz given, while instructors using one of the confidence procedures kept records only for the quizzes they actually scored using their particular confidence format. Distributions of reported testing and scoring times were obtained and appear in Table 17. In addition to the distributions, statistics for the mean scoring time (minutes) required to score an answer sheet were obtained for each type of testing under study in each course. Table 17 indicates that although the mean test administration times were significantly different statistically, the difference was only one or two minutes in the Jet Mechanic course and thus probably not significant in a practical sense. In AGE, on the other hand, tests administered as Pick-One confidence testing were found to require less time for administration than either the conventional multiple-choice or the Distribute 100 Points confidence test procedures. This finding must have occurred as a result of the rough estimates of testing administration time since the task required of an examinee using multiple-choice testing was only a part of what is required of an examinee using Pick-One confidence testing.

For both courses the table reflected extreme differences in the time required to score the daily quizzes. Roughly speaking, the time required to score a Distribute 100 Points confidence test was about twice that required for scoring a multiple-choice test. The Pick-One confidence test scoring required even longer.

The probable reason for the greater time required for the Pick-One confidence testing procedure than for the Distribute 100 Points testing was the use of a real number scale rather than the integer scale used for Distribute 100 Points. A modification of the scales used for hand scoring has been recommended elsewhere (Echternacht, Boldt, & Sellman, 1971). These modifications reduced the Pick-One scale to an integer format and reduced the number of possible scores in the Distribute 100 Points case. Had the scale for Pick-One confidence testing been in integer form, it was hypothesized that this method would have reduced the time for scoring Pick-One tests to a level less than that for Distribute 100 Points.

Table 17

Frequency of Testing and Scoring Times Reported by Instructors for Daily Quizzes Using Confidence and Multiple-Choice Testing

Time in Minutes	Test Administration Time						Scoring Time					
	JEM			AGE			JEM			AGE		
	Multiple Choice	Pick One	Distribute 100 Points	Multiple Choice	Pick One	Distribute 100 Points	Multiple Choice	Pick One	Distribute 100 Points	Multiple Choice	Pick One	Distribute 100 Points
40 and Above	--	--	--	4	--	10	3	7	1	4	13	14
35-39	--	--	--	3	1	6	--	7	--	--	5	10
30-34	3	3	3	14	3	13	--	6	5	2	4	10
25-29	4	6	0	4	15	6	1	8	4	2	2	7
20-24	22	6	3	13	5	8	22	6	11	5	10	8
15-19	92	10	16	16	7	13	20	3	1	16	2	11
10-14	131	17	13	6	5	7	83	7	4	15	--	3
5- 9	45	6	1	--	1	--	135	4	9	17	--	--
Below 5	--	--	--	--	--	--	25	--	--	--	--	--
Mean	12.5	15.4	14.8	23.8	21.4	27.1	9.0	25.6	19.8	14.4	33.6	29.5
S.D.	4.4	7.0	5.5	10.0	6.2	12.4	6.1	12.0	11.4	9.6	14.4	12.3

Average Scoring Time (Minutes) per Answer Sheet			
	JEM	Multiple Choice	.87
		Pick-one	2.30
		Distribute 100 pts.	1.69
	AGE	Multiple Choice	1.52
		Pick-one	3.36
		Distribute 100 pts.	2.88

SECTION V

Conclusions

Conclusions from this study may be grouped into four categories: (1) those involving directly measurable criteria, i.e., end of block examination scores and the average number of student remediations; (2) those involving student and instructor attitudes; (3) those involving the relationship between personality and confidence marking; and (4) those involving the quiz administration and scoring time.

1. Conclusions resulting from the analysis of the directly measurable criteria.

   a. The effectiveness of the two types of confidence testing under study, Pick-One and Distribute 100 Points confidence testing, is dependent upon the type of remediation used and the shift in which the procedures were used.

   b. When significant differences between types of testing occurred with respect to end of block examination scores, multiple choice testing was always low, and the difference between multiple choice testing and the confidence procedures was large.

   c. The effects of using either the Pick-One or the Distribute 100 Points method were mixed. Neither method appeared to be superior to the other.

   d. Multiple-choice testing resulted in more remediations being required, on the average, than either confidence testing procedure. Neither confidence testing procedure was superior to the other with respect to average number of remediations required.

2. Conclusions resulting from the analysis of student and instructor attitudes.

   a. Although items were found favoring each of the three types of testing under study, no one method emerged as being highly preferred. It is concluded that students are indifferent to the type of testing method to be used for daily quizzes.

   b. Instructors tended <u>not</u> to use the confidence marks in planning remediation.

   c. Instructors expressed difficulty with the scoring of the confidence testing and objected to the length of time required to score them.

3. Conclusions resulting from the analysis of the personality data.

   a. Tests scored with a rights-only scoring formula correlate so highly with confidence test scores, using either the Pick-One or Distribute 100 Points methods, that the use of a confidence score seems unnecessary from a psychometric point of view.

   b. Although some personality trait scores correlate with confidence test scores to a significant degree when rights-only score is partialled out, these correlations do not appear to be stable from one test to another.

   c. Various personality traits do not contribute significantly to the prediction of confidence test scores when the influence of rights-only score is eliminated.

4. Conclusions resulting from the analysis of quiz administration and scoring time.

   a. Although the difference is statistically significant, there is no practical difference in the time required to administer daily quizzes in technical training as confidence tests.

   b. The time required to score the Distribute 100 Points method of confidence testing is about twice that required for multiple choice. In the case for the Pick-One method, the time required is slightly more than that required for the Distribute 100 Points method.

   c. Simplified scoring tables could be developed that should yield Pick-One scoring times that are closer to the multiple choice times.

SECTION VI

Recommendations

1. Since the results of the study did not indicate an overwhelming advantage for confidence testing, it is recommended that the implementation of any confidence testing program be undertaken with caution.

2. Further work is required on simplifying the scoring procedure used by the instructors.

3. Confidence testing does merit consideration as a method for diagnostic testing in technical training since students using the procedure perform as well or better on end of block examinations than students using conventional multiple-choice testing, and the number of remediations required seems to decrease when confidence testing is used.

4. Further work is required on developing systems for using confidence responses in remediation.

COMMENT: Since it appears that confidence testing affects subsequent performance, it may be that it should be used to a greater extent in remediation procedures. It may also be that the results of confidence testing would be used in remediation more than they were here IF more use of intermediate levels of certainty by the examinees could be brought about (possibly through the use of the modified Pick-One procedure and scoring). Such a project would require more adjustment of the teaching procedures, particularly remediation procedures, than was possible in the present study.

REFERENCES

Air Training Command, Plan of instruction: Jet engine mechanic. POI 3ABR43230. 3345th Technical School, Chanute AFB, Ill., 1970. (a)

Air Training Command, Plan of instruction: Aerospace ground equipment repairman. POI 3ABR42133. 3345th Technical School, Chanute AFB, Ill., 1970. (b)

Alpert, R. & Haber, R.N. Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 1960, 61, 207-215.

Barratt, E. S. Anxiety and impulsiveness related to psychometric efficiency. Perceptual and Motor Skills, 1959, 9, 191-198.

Boldt, R. F. A simple confidence testing format. AFHRL-TR-71-31. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Clayton, M. B. & Jackson, D. N. Equivalence range, acquiescence, and over-generalization. Educational and Psychological Measurement, 1961, 21, 371-382.

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

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 13, 87-123.

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

Echternacht, G. J., Boldt, R. F., & Sellman, W. S. Users handbook for confidence testing as a diagnostic aid in technical training. AFHRL 71-34. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Echternacht, G. J. The use of confidence testing in objective tests. AFHRL 71-32. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Fisher, R. A. Statistical method for research workers. (13th ed.) New York: Hafner, 1958.

Gough, H. G. Manual for the California Psychological Inventory. Palo Alto, California: Consulting Psychologists Press, 1957.

Hevner, K. A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359-362.

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

Morrison, D. F. Multivariate Statistical Methods. New York: McGraw-Hill, 1967.

Pruzek, R. M. Methods and problems in the analysis of multivariate data. Review of Educational Research, 1971, 41, 163-190.

Rao, C. R. Advanced Statistical Methods in Biometric Research. New York: Wiley, 1952.

Rokeach, M. Political and religious dogmatism: An alternative to the authoritarian personality. Psychological Monographs: General and Applied, 1956, 70, 18.

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

Shuford, E. H. & Massengill, H. E. Confidence testing at the officer training school, Lackland Air Force Base, September 1968. Lexington, Mass.: Shuford-Massengill Corporation, 1969.

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

APPENDIX I

Procedures and Scoring

CONFIDENCE TESTING AS A DIAGNOSTIC AID IN TECHNICAL TRAINING

1.0 Introduction: During the next several weeks, classes in Aerospace Ground Equipment Repairman and Jet Engine Mechanic courses will take part in a study of confidence testing as a diagnostic aid. Blocks 6, 7, and 8 have been selected for study in the AGE course and blocks 2 and 3 for the Jet Engine course. This is the study which a briefing was given to the instructors on 13 August.

2.0 Objectives: The objectives of this study are two in number. First is to determine the feasibility of using confidence testing as diagnostic evaluative aids to instructors in AF technical training courses. This means that the study is designed to determine how well confidence testing helps you find out exactly what your students know, and how well, and just what they are not sure of or think they know when actually they don't. The second purpose is to determine the cost effectiveness of confidence testing vs that of conventional multiple choice testing practices currently in use in AF technical training courses. Feasibility will be evaluated in terms of student performance, attitudes toward the applicability and practicability of confidence testing, and a cost effectiveness analysis.

3.0 Tasks to be performed: The design is set up so that there will be six different kinds of classes for each shift of each course. You as instructors will be responsible for a number of tasks.

3.1 Personality tests: There will be three personality tests that need to be administered as soon as possible. Test 1 asks the student to express agreement or disagreement with a number of opinions. Test 2 is a test of willingness to take risks. Test 3 consists of a number of statements about feelings, tendencies, and preferences of the student which either characterize him or are uncharacteristic of him. Every student should respond to every item on the tests; if a student cannot take the tests at the same time as the others, you should make arrangements for him to take these tests as soon as possible. The best possible time for these tests to be administered is the first day of class in block 6 of the AGE course and block 2 of the Jet Engine course. This will be the only time these tests will be administered. It is absolutely necessary for every student to answer all the items in the personality tests once. Make sure that each student places his name and social security number in the upper right hand corner of each test.

3.2 The types of testing groups: You will be asked to conduct all your daily multiple choice quizes in one of three ways for a given group.

3.2.1 Conventional testing: Some groups will use the method that is now in use. In this method the student simply marks the answer that he thinks is most likely to be correct. Directions for the students are included. There should be no writing on the directions as these are to be used again for each testing period.

3.2.2 Pick-one confidence testing: In this method the student is asked to choose the answer that he thinks is most likely to be correct as in 3.2.1 and then indicates how sure he was that the answer he marked was in fact the correct one. This is done on a five point scale that appears on the right of the answer sheet. Directions for the students are included. There should be no writing on the directions as these are to be used again for each testing period.

3.2.3 Distribute 100 points confidence testing: In this method the student first indicates the answer that he thinks is most likely to be correct and marks that one, then he shows his feelings about the possible alternatives by distributing 100 points over the alternatives placing the most points on the answer that he has marked and a lesser number on any of the alternatives that he feels might be correct. Directions for the students are included. There should be no writing on the directions as these are to be used again for each testing period.

3.3 The types of remediation groups: In addition to using the above mentioned types of testing, you will conduct your remediation according to two different types of remediation.

3.3.1 Control remediation: Control remediation refers to that remediation that you are now using. You should assign people to remediation exactly as you do now and conduct remediation as you have in the past. An important point here is not to adjust your remediation in light of the special remediation that is described next. The goal of the study is to compare the present method with the method that follows which makes it necessary for you to conduct your remediation exactly as you do now when using control remediation.

3.3.2 Special remediation for conventional testing:

1. The first step in this method is to decide just who is to attend remedial. This should be done as you have in the past.
2. After you decide who is to attend remedial, make a list of all the items that every student missed in the remedial group. These items are the common group of items missed by everyone in remedial. During the remediation, you must explain why these answers they marked were wrong and why the correct answer was right.
3. You should make a second list of questions that only some of the people in the remedial group missed. Since some in this group answered correctly, have the students who answered correctly explain why the others were wrong and why the correct answer was right during remediation.
4. The basic principle involved in that every student in remediation should know why his wrong answers were wrong and why the right answers were correct.

3.3.3 Special remediation for confidence testing:

1. The first step here is once again to decide who is going to have to attend the remedial sessions. This should be done as you have in the past.
2. Once you have decided who is to attend remedial, make a list of all the questions that every student assigned to remediation missed.
3. Then, look at the confidence they placed in their wrong answers to these questions. If they placed a large amount of confidence in their answers, a great deal of time must be spent explaining why their answers were wrong and then why the correct answer is right. If they placed a small amount of confidence in their answer, less time may be spent on explaining why their answers were wrong. The question of how much is a large amount of confidence and a small amount of confidence should be decided by yourself. For general purposes though we can say that a large amount of confidence is 60 or more in the system where 100 points are distributed and when either of the top two confidence responses are checked in the system where you indicate confidence only in your answer. A small amount of confidence is between 25 and 40 in the one system and when either of the lower two responses are checked in the other system.
4. A list of questions should be made that only some of the remedial group missed. In this case the people who answered the question correctly should explain.
5. The general principle to be followed here is the same as previous, that being, the more confidence an individual places in a wrong answer, the more time is required to show that student why his answer is wrong.

3.4 The keeping of cost effectiveness records: You will be required to keep a record of both the time you spend correcting the daily quizes and in administering the daily quizes. This information will be provided by you on a special record form. This record identifies you, the course and block titles, shift, the dates of testing, the time spent on correcting the tests, the length of time the class took to complete the tests, and any comments you might have. The time spent correcting the tests should begin the moment you begin to look at the test papers and end when you have completed the grading and decided who, if anyone, must attend remedial.

3.5 Special scoring for confidence testing: When you are teaching a class that is using confidence testing, you will be required to score your daily quizes using a special scoring rule on three different occasions. These sessions will take place on the last three days of multiple choice daily quizes. The directions for the confidence testing indicate that the test will be scored using this special format at all testings. They should think that this is true for all testings although you will only score in this special way the last three times. You must encourage them to answer honestly and under no circumstance tell them that their confidence marks do not count in grading.

3.5.1 Special scoring for the pick-one confidence testing:

1. The first task that you must do is to examine the choices to the left. Circle any of those answers that are incorrect.
2. Next, look up the score on the scoring table for the answers that have been marked correct and add these together for the individual.
3. Next, look up the scores on the scoring table for the circled answers and add these together for that same individual. To obtain the total score, take the scores from 2 and decrease that score by the amount determined in 3.

3.5.2 Scoring for the distribute 100 points confidence testing:

1. First, circle the number of points that the student has given the correct answer. You can ignore the answer on the left hand side of the paper.
2. Next, look up on the scoring table the score that corresponds to the number of points that is given to the correct answer.
3. Add these numbers together to obtain the total score. When a student places less than 10 points on the correct answer, notice that his score is negative.

3.6 Student questionnaires: At the end of block 8 in the AGE course and block 3 in the Jet Engine course on the last class day, you are to administer one of three questionnaires to the students. The form of the questionnaire that you administer to the particular class depends on the kind of testing that group has been using over the past weeks. The first form is used for students that use the conventional testing, the second form for the pick-one confidence testing, and the third form for the distribute 100 points confidence testing.

3.7 Instructor questionnaires: The questionnaire that you are asked to fill out is of the open ended variety. That means that you are to respond freely and as completely as possible. The more information you can give us the better we will be able to make recommendations for future applications. These questionnaires should be filled out immediately after the last group that is involved in the experiment finishes your block.

COURSE _____ BLOCK _____ BLOCK TITLE _____

INSTRUCTOR _____ SHIFT _____

Date	Number Answer Sheets Scored	Type of Scoring		Time for Scoring (Mins.)	Time for Test Administration	Comments
		Regular	Special			

SCORING TABLE FOR DISTRIBUTE 100 POINTS CONFIDENCE TEST

Points on the Correct Answer	Score	Points on the Correct Answer	Score
0	- 100		
1	- 100	51	71
2	- 70	52	72
3	- 52	53	72
4	- 40	54	73
5	- 30	55	74
6	- 22	56	75
7	- 15	57	76
8	- 10	58	76
9	- 05	59	77
10	00	60	78
11	04	61	79
12	08	62	79
13	11	63	80
14	15	64	81
15	18	65	81
16	20	66	82
17	23	67	83
18	26	68	83
19	28	69	84
20	30	70	85
21	32	71	85
22	34	72	86
23	36	73	86
24	38	74	87
25	40	75	88
26	41	76	88
27	43	77	89
28	45	78	89
29	46	79	90
30	47	80	90
31	49	81	91
32	51	82	91
33	52	83	92
34	53	84	92
35	54	85	93
36	56	86	93
37	57	87	94
38	58	88	94
39	59	89	95
40	60	90	95
41	61	91	96
42	62	92	96
43	63	93	97
44	64	94	97
45	65	95	98
46	66	96	98
47	67	97	99
48	68	98	99
49	69	99	100
50	70	100	100

SCORING TABLE FOR PICK-ONE CONFIDENCE TESTING

	Score
If the answer is <u>correct</u> and the confidence is	
Very Sure	1.00
Sure	.89
Fairly Sure	.56
Not Very Sure	.36
Not Sure	0
If the answer is <u>wrong</u> and the confidence is	
Very Sure	-1.67
Sure	- .89
Fairly Sure	- .33
Not Very Sure	- .17
Not Sure	- 0

The following is an example of a 15 item multiple-choice test administered as a confidence test using both Pick-One and Distribute 100 Points formats. This test served as a pretest of the confidence systems as it was given to approximately 20 ETS employees not associated with the project. The questions involve mostly information known to many people in the Princeton, New Jersey geographic area. An attempt was made to include questions of varying difficulty.

Two answer sheets, one for each type of testing, follow. The answer sheets show how an examinee might respond to the given quiz.

EXAMPLE

1. The colors on a New Jersey license plate are

   (A) black and white
   (B) blue and yellow
   (C) black and cream*
   (D) brown and cream

2. The 1969 population of Princeton, New Jersey was estimated to be

   (A) 11,890*
   (B) 981
   (C) 13,060
   (D) 56,550

3. The capitol of the state of Washington is

   (A) Seattle
   (B) Tacoma
   (C) Spokane
   (D) Olympia*

4. The largest city in New Jersey is

   (A) Jersey City
   (B) Newark*
   (C) Camden
   (D) Paterson

5. The largest city in South Dakota is

   (A) Watertown
   (B) Souix Falls*
   (C) Aberdeen
   (D) Rapid City

6. The 12th President of the United States was

   (A) Truman
   (B) Taylor*
   (C) Madison
   (D) Pierce

7. The 1969 population of Trenton was estimated to be

   (A) 101,000
   (B) 104,000
   (C) 2,189
   (D) 102,000*

8. Which automobile model is usually considered to be most expensive?

   (A) Cadillac*
   (B) Ford
   (C) Rambler
   (D) Plymouth

9. The colors on a 1970 New Mexico license plate are

(A) yellow and black
(B) black and white
(C) yellow and red*
(D) red and white

10. Which state does not border Tennessee?

(A) Missouri
(B) South Carolina*
(C) Virginia
(D) Georgia

11. Which city is closest to the Pocono Mountains?

(A) Philadelphia
(B) Bayonne
(C) Scranton*
(D) Trenton

12. Educational Testing Service is located in which township?

(A) Lawrence*
(B) Princeton
(C) Ewing
(D) Hopewell

13. Which town is not located in Mercer County?

(A) Princeton
(B) Ringoes*
(C) Harbourton
(D) Hightstown

14. On the New Jersey Turnpike, what is the number of the New Brunswick Exit?

(A) 8
(B) 8A
(C) 9*
(D) 10

15. Which day is most likely to be pay-day at ETS?

(A) Wednesday*
(B) Friday
(C) Monday
(D) Thursday

A/S Example

1. C	A. 0	
	B. 0	
	C. 75	
	D. 25	
2. C	A. 50	
	B. 0	
	C. 50	
	D. 0	
3. A	A. 70	
	B. 0	
	C. 0	
	D. 30	
4. B	A. 100	
	B. 0	
	C. 0	
	D. 0	
5. B	A. 20	
	B. 40	
	C. 20	
	D. 20	
6. D	A. 0	
	B. 30	
	C. 30	
	D. 40	

7.	D	A. 30 B. 35 C. 0 D. 35
8.	A	A. 100 B. 0 C. 0 D. 0
9.	A	A. 25 B. 25 C. 25 D. 25
10.	A	A. 40 B. 30 C. 30 D. 0
11.	C	A. 30 B. 0 C. 70 D. 0
12.	B	A. 0 B. 100 C. 0 D. 0

13. C

A. 0
B. 50
C. 50
D. 0

14. C

A. 25
B. 25
C. 25
D. 25

15. A

A. 100
B. 0
C. 0
D. 0

1. C

_____ very sure
__X__ sure
_____ fairly sure
_____ not very sure
_____ not sure

2. C

_____ very sure
_____ sure
__X__ fairly sure
_____ not very sure
_____ not sure

3. A

_____ very sure
__X__ sure
_____ fairly sure
_____ not very sure
_____ not sure

4. B

__X__ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

5. B

_____ very sure
_____ sure
_____ fairly sure
__X__ not very sure
_____ not sure

6. D

____ very sure
____ sure
X fairly sure
____ not very sure
____ not sure

7. D

____ very sure
____ sure
____ fairly sure
X not very sure
____ not sure

8. A

X very sure
____ sure
____ fairly sure
____ not very sure
____ not sure

9. A

____ very sure
____ sure
____ fairly sure
____ not very sure
X not sure

10. A

____ very sure
____ sure
____ fairly sure
X not very sure
____ not sure

11. C

_____ very sure
__X__ sure
_____ fairly sure
_____ not very sure
_____ not sure

12. B

__X__ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

13. C

_____ very sure
_____ sure
__X__ fairly sure
_____ not very sure
_____ not sure

14. C

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
__X__ not sure

15. A

__X__ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

APPENDIX II

Personality Tests

TEST 1

OPINION INVENTORY FORM PQEP

Directions

In this inventory, you will find a number of statements expressing opinions with which you may or may not agree. Following each statement are six boxes labeled as follows:

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
☐	☐	☐	☐	☐	☐

You are to indicate the degree to which you agree or disagree with each statement by checking the appropriate box.

You may notice an occasional statement with which you neither particularly agree nor particularly disagree. If so, do the best you can by checking the box that seem most appropriate.

Please consider each statement carefully, but do not spend too much time on any one statement. Do not skip any items.

There are no "right" or "wrong" answers--the only correct responses are those that are true for you. This inventory is being used for research purposes only and your responses will be kept strictly confidential.

TURN THE PAGE AND BEGIN.

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. Sex crimes, such as rape and attacks on children always deserve more than mere imprisonment; such criminals must be publicly whipped or worse.	☐	☐	☐	☐	☐	☐
2. The true American way of life is disappearing so fast that force is absolutely necessary to preserve it.	☐	☐	☐	☐	☐	☐
3. The businessman and the manufacturer are undoubtedly more important to society than the artist and the professor.	☐	☐	☐	☐	☐	☐
4. Nowadays everyone is prying into matters that must remain personal and private.	☐	☐	☐	☐	☐	☐
5. Someday it will certainly be shown that astrology can explain a lot of things.	☐	☐	☐	☐	☐	☐
6. All of our social problems would be solved if we got rid of the immoral, crooked, and feebleminded people.	☐	☐	☐	☐	☐	☐
7. No sane, normal, decent person could ever think of hurting a close friend or relative.	☐	☐	☐	☐	☐	☐
8. Every person should have complete faith in his own independent judgment, not in some supernatural power whose decisions he obeys without question.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
9. Nowadays, since democracy demands that people of widely different background and station mix together a person should never be finicky about catching a disease from any of them.	☐	☐	☐	☐	☐	☐
10. When a person has a problem or worry he should always drop everything and concentrate upon it until the solution appears.	☐	☐	☐	☐	☐	☐
11. We are certainly bound to admire and respect a person if we get to know him well.	☐	☐	☐	☐	☐	☐
12. An insult to our honor should always be overlooked, for "whatsoever shall smite thee on thy right cheek, turn to him the other also."	☐	☐	☐	☐	☐	☐
13. Every truly mature person outgrows childish feelings of submissive respect and of excessive love and gratitude for his parents.	☐	☐	☐	☐	☐	☐
14. All attempts to divide people into two distinct classes of the weak and the strong are doomed to failure.	☐	☐	☐	☐	☐	☐
15. Science has its place, but there are probably things that might not be understood by the human mind.	☐	☐	☐	☐	☐	☐
16. A person who had bad manners, habits, and breeding would probably find it hard to get along with decent people.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
17. Seldom do weaknesses or difficulties hold us back if we have enough will power.	☐	☐	☐	☐	☐	☐
18. Wars and social troubles may someday be ended by an earthquake or flood that could destroy the whole world.	☐	☐	☐	☐	☐	☐
19. The wild sex life of the old Greeks and Romans was probably tame compared to some of the goings-on in this country, even in places where people might least expect it.	☐	☐	☐	☐	☐	☐
20. What this country probably needs is a few courageous, tireless, devoted leaders in whom the people can put their faith.	☐	☐	☐	☐	☐	☐
21. Some youth probably need the qualities of strict discipline, rugged determination, and the will to work and fight for family and country.	☐	☐	☐	☐	☐	☐
22. The urge to jump from high places is probably learned, not inborn.	☐	☐	☐	☐	☐	☐
23. The rebellious ideas that young people sometimes get should probably be encouraged and developed to guarantee mature citizenship in adulthood.	☐	☐	☐	☐	☐	☐
24. Probably few people have learned important things through suffering.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
25. A love of freedom and complete independence may be important virtues for children to learn.	☐	☐	☐	☐	☐	☐
26. Because human nature is improving, war and conflict may someday be eliminated.	☐	☐	☐	☐	☐	☐
27. Homosexuals may not be criminals and probably should not be punished as such.	☐	☐	☐	☐	☐	☐
28. If people occasionally talked things over and didn't work so hard, some others would probably be better off.	☐	☐	☐	☐	☐	☐
29. In times like these it is often necessary to be more on guard against ideas put out by people or groups in one's own camp than by those in the opposing camp.	☐	☐	☐	☐	☐	☐
30. Man on his own is a helpless and miserable creature.	☐	☐	☐	☐	☐	☐
31. A group which tolerates too much difference of opinion among its own members cannot exist for long.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
32. Unfortunately, a good many people with whom I have discussed important social and moral problems don't really understand what's going on.	☐	☐	☐	☐	☐	☐
33. Most people just don't know what's good for them.	☐	☐	☐	☐	☐	☐
34. Of all the different philosophies which exist in this world there is probably only one which is correct.	☐	☐	☐	☐	☐	☐
35. Most people just don't give a "damn" for others.	☐	☐	☐	☐	☐	☐
36. It is only natural for a person to be rather fearful of the future.	☐	☐	☐	☐	☐	☐
37. In the history of mankind there have probably been just a handful of really great thinkers.	☐	☐	☐	☐	☐	☐
38. The worst crime a person could commit is to attack publicly the people who believe in the same thing he does.	☐	☐	☐	☐	☐	☐
39. My blood boils whenever a person stubbornly refuses to admit he's wrong.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
40. The main thing in life is for a person to want to do something important.	☐	☐	☐	☐	☐	☐
41. A person who gets enthusiastic about too many causes is likely to be a pretty "wishy-washy" sort of person.	☐	☐	☐	☐	☐	☐
42. If given the chance I would do something of great benefit to the world.	☐	☐	☐	☐	☐	☐
43. A man who does not believe in some great cause has not really lived.	☐	☐	☐	☐	☐	☐
44. It is only when a person devotes himself to an ideal or cause that life becomes meaningful.	☐	☐	☐	☐	☐	☐
45. In the long run the best way to live is to pick friends and associates whose tastes and beliefs are the same as one's own.	☐	☐	☐	☐	☐	☐
46. Most of the ideas which get printed nowadays aren't worth the paper they are printed on.	☐	☐	☐	☐	☐	☐
47. There are two kinds of people in this world: those who are for the truth and those who are against the truth.	☐	☐	☐	☐	☐	☐

	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
48. In this complicated world of ours the only way we can know what's going on is to rely on leaders or experts who can be trusted.	☐	☐	☐	☐	☐	☐

TEST 2

INSTRUCTIONS: In this task, you will be shown pairs of dice bets that vary in terms of the chances of winning and losing, and the amounts of money that can be won or lost. I would like you to choose, in each pair, the bet that you would prefer to play. Indicate your decision by making a check in either box A or B in the right hand column below. Consider each pair separately -- do not let your decision in one case influence your decision in another. Later you will have the opportunity to actually play the bets that you now choose. You will play them in a dice game for the amounts of money described in the bets. So be sure that you choose now the bets that you actually will want to play, because you will be held to them.

The chances of winning and losing are written as fractions: Thus, 1/3 means 1 chance in 3, 1/2 means 1 chance in 2, 2/3 means 2 chances in 3.

Check the box on the right to indicate which bet you choose to make. The box marked A refers to the left side bet, the box marked B refers to the right side bet.

				A	B
1.	1/3 to win \$1.20 2/3 to lose \$ .60	vs.	1/2 to win \$ .30 1/2 to lose \$ .30	☐	☐
2.	2/3 to win \$ .15 1/3 to lose \$ .30	vs.	1/2 to win \$ .90 1/2 to lose \$ .90	☐	☐
3.	1/2 to win \$ .30 1/2 to lose \$ .30	vs.	2/3 to win \$ .30 1/2 to lose \$ .60	☐	☐
4.	1/2 to win \$ .90 1/2 to lose \$ .90	vs.	1/3 to win \$1.20 2/3 to lose \$ .60	☐	☐
5.	1/3 to win \$1.80 2/3 to lose \$ .90	vs.	2/3 to win \$ .45 1/3 to lose \$ .90	☐	☐
6.	2/3 to win \$ .15 1/3 to lose \$ .30	vs.	1/2 to win \$ .30 1/2 to lose \$ .30	☐	☐
7.	1/3 to win \$1.20 2/3 to lose \$ .60	vs.	1/2 to win \$ .60 1/2 to lose \$ .60	☐	☐
8.	2/3 to win \$ .15 1/3 to lose \$ .30	vs.	1/3 to win \$1.20 2/3 to lose \$ .60	☐	☐
9.	2/3 to win \$ .15 1/3 to lose \$ .30	vs.	1/3 to win \$ .60 2/3 to lose \$ .30	☐	☐
10.	1/3 to win \$1.80 2/3 to lose \$ .90	vs.	1/2 to win \$ .30 1/2 to lose \$ .30	☐	☐

				A	B
11.	2/3 to win $ .45 1/3 to lose $ .90	vs.	2/3 to win $ .15 1/3 to lose $ .30	☐	☐
12.	1/3 to win $1.20 2/3 to lose $ .60	vs.	2/3 to win $ .30 1/3 to lose $ .60	☐	☐
13.	1/2 to win $ .90 1/2 to lose $ .90	vs.	2/3 to win $ .45 1/3 to lose $ .90	☐	☐
14.	1/2 to win $ .90 1/2 to lose $ .90	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐
15.	1/3 to win $ .60 2/3 to lose $ .30	vs.	1/2 to win $ .30 1/2 to lose $ .30	☐	☐
16.	2/3 to win $ .45 1/3 to lose $ .90	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐
17.	2/3 to win $ .45 1/3 to lose $ .90	vs.	1/2 to win $ .30 1/2 to lose $ .30	☐	☐
18.	1/2 to win $ .30 1/2 to lose $ .30	vs.	1/2 to win $ .90 1/2 to lose $ .90	☐	☐
19.	2/3 to win $ .45 1/3 to lose $ .90	vs.	1/3 to win $1.20 2/3 to lose $ .60	☐	☐
20.	1/2 to win $ .90 1/2 to lose $ .90	vs.	2/3 to win $ .30 1/3 to lose $ .60	☐	☐
21.	1/3 to win $1.20 2/3 to lose $ .60	vs.	1/3 to win $ .60 2/3 to lose $ .30	☐	☐
22.	2/3 to win $ .15 1/3 to lose $ .30	vs.	1/3 to win $1.80 2/3 to lose $ .90	☐	☐
23.	1/2 to win $ .90 1/2 to lose $ .90	vs.	1/3 to win $1.80 2/3 to lose $ .90	☐	☐
24.	2/3 to win $ .30 1/3 to lose $ .60	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐
25.	1/3 to win $1.20 2/3 to lose $ .60	vs.	1/3 to win $1.80 2/3 to lose $ .90	☐	☐
26.	2/3 to win $ .15 1/3 to lose $ .30	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐

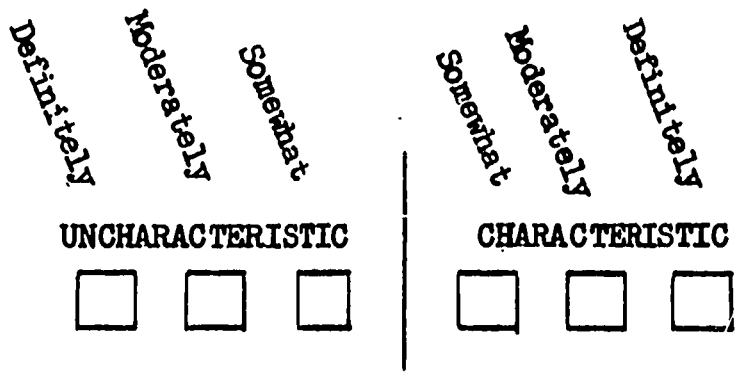
				A	B
27.	2/3 to win $ .15 1/3 to lose $ .30	vs.	2/3 to win $ .30 1/3 to lose $ .60	☐	☐
28.	1/2 to win $ .60 1/2 to lose $ .60	vs.	1/3 to win $1.80 2/3 to lose $ .90	☐	☐
29.	1/3 to win $ .60 2/3 to lose $ .30	vs.	2/3 to win $ .45 1/3 to lose $ .90	☐	☐
30.	1/2 to win $ .90 1/2 to lose $ .90	vs.	1/3 to win $ .60 2/3 to lose $ .30	☐	☐
31.	1/3 to win $ .60 2/3 to lose $ .30	vs.	1/3 to win $1.80 2/3 to lose $ .90	☐	☐
32.	1/3 to win $1.80 2/3 to lose $ .90	vs.	2/3 to win $ .30 1/3 to lose $ .60	☐	☐
33.	2/3 to win $ .45 1/3 to lose $ .90	vs.	2/3 to win $ .30 1/3 to lose $ .60	☐	☐
34.	2/3 to win $ .30 1/3 to lose $ .60	vs.	1/3 to win $ .60 2/3 to lose $ .30	☐	☐
35.	1/2 to win $ .30 1/2 to lose $ .30	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐
36.	1/3 to win $ .60 2/3 to lose $ .30	vs.	1/2 to win $ .60 1/2 to lose $ .60	☐	☐

TEST 3

PERSONAL INVENTORY--FORM PQEP

Directions

This inventory consists of a number of statements about feelings, tendencies, and preferences that may or may not be characteristic of you. Following each statement are six boxes labeled as follows:



Notice that there are three boxes on the left labeled Uncharacteristic with three gradations of difference ranging from Somewhat through Moderately to Definitely Uncharacteristic. Likewise there are three boxes on the right labeled Characteristic with three gradations of difference ranging from Somewhat through Moderately to Definitely Characteristic. You are to indicate the degree to which each statement is characteristic of you by checking the appropriate box.

You may notice an occasional statement that is neither particularly characteristic nor particularly uncharacteristic of you. If so, do the best you can be checking the box that seems most appropriate.

Please consider each statement carefully, but do not spend too much time on any one item. Do not skip any items.

There are no "right" or "wrong" answers--the only correct responses are those that are true for you. This inventory is being used for research purposes only and your responses will be kept strictly confidential.

TURN THE PAGE AND BEGIN

	UNCHARACTERISTIC			CHARACTERISTIC		
	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
1. During exams or tests, I block on questions to which I know the answers, even though I might remember them as soon as the exam is over.	☐	☐	☐	☐	☐	☐
2. The more important the examination, the less well I seem to do.	☐	☐	☐	☐	☐	☐
3. Time pressure on an exam causes me to do worse than the rest of the group under similar conditions.	☐	☐	☐	☐	☐	☐
4. I find that my mind goes blank at the beginning of an exam, and it takes me a few minutes before I can function.	☐	☐	☐	☐	☐	☐
5. Nervousness while taking an exam or test hinders me from doing well.	☐	☐	☐	☐	☐	☐
6. I find myself reading exam questions without understanding them, and I must go back over them so that they will make sense.	☐	☐	☐	☐	☐	☐
7. When I don't do well on a difficult item at the beginning of an exam, it tends to upset me so that I block on even easy questions later on.	☐	☐	☐	☐	☐	☐
8. In a course where I have been doing poorly, my fear of a bad grade cuts down my efficiency.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
9. When I am poorly prepared for an exam or test, I get upset, and do less well than even my restricted knowledge should allow.	☐	☐	☐	☐	☐	☐
10. I am so tired from worrying about an exam, that I find I almost don't care how well I do by the time I start the test.	☐	☐	☐	☐	☐	☐
11. While I may (or may not) be nervous before taking an exam, once I start, I seem to forget to be nervous.	☐	☐	☐	☐	☐	☐
12. I look forward to exams.	☐	☐	☐	☐	☐	☐
13. The more important the exam or test, the better I seem to do.	☐	☐	☐	☐	☐	☐
14. When I start a test, nothing is able to distract me.	☐	☐	☐	☐	☐	☐
15. Although "cramming" under pre-examination tension is not effective for most people, I find that if the need arises, I can learn material immediately before an exam, even under considerable pressure, and successfully retain it to use on the exam.	☐	☐	☐	☐	☐	☐

	UNCHARACTERISTIC			CHARACTERISTIC		
	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
16. In courses in which the total grade is based mainly on one exam, I seem to do better than other people.	☐	☐	☐	☐	☐	☐
17. I enjoy taking a difficult exam more than an easy one.	☐	☐	☐	☐	☐	☐
18. I work most effectively under pressure, as when the task is very important.	☐	☐	☐	☐	☐	☐
19. Nervousness while taking a test helps me do better.	☐	☐	☐	☐	☐	☐
20. I am often one of the first to give up trying to do a thing.	☐	☐	☐	☐	☐	☐
21. I dislike work that requires a great deal of attention to detail.	☐	☐	☐	☐	☐	☐
22. I am seldom methodical in the things that I do.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
23. I often don't finish tasks I start, sometimes even if they are very important.	☐	☐	☐	☐	☐	☐
24. It doesn't bother me to change my plans in the midst of an undertaking.	☐	☐	☐	☐	☐	☐
25. I do not work and study following a strict schedule.	☐	☐	☐	☐	☐	☐
26. Occasionally, I have done something dangerous just for the thrill of it.	☐	☐	☐	☐	☐	☐
27. I do not believe that promptness is a very important personality characteristic.	☐	☐	☐	☐	☐	☐
28. I enjoy having to adapt myself to new and unusual situations.	☐	☐	☐	☐	☐	☐
29. I am not always careful about my manner of dress.	☐	☐	☐	☐	☐	☐
30. I seldom become so wrapped up in something I am doing that I find it difficult to turn my attention to other matters.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
31. My interests tend to change quickly.	☐	☐	☐	☐	☐	☐
32. I am inclined to go from one activity to another without continuing with any one for too long a time.	☐	☐	☐	☐	☐	☐
33. I think it is usually wise to do things in a conventional way.	☐	☐	☐	☐	☐	☐
34. I find it easy to stick to a certain schedule, once I have started it.	☐	☐	☐	☐	☐	☐
35. I often find myself thinking of the same tunes or phrases for days at a time.	☐	☐	☐	☐	☐	☐
36. I always put on and take off my clothes in the same order.	☐	☐	☐	☐	☐	☐
37. I usually check more than once to be sure that I have locked a door, put out the light, or something of the sort.	☐	☐	☐	☐	☐	☐
38. I usually find that my own way of attacking a problem is best, even though it doesn't always seem to work in the beginning.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
39. I try to follow a program of life based on duty.	☐	☐	☐	☐	☐	☐
40. I prefer to stop and think before I act even on trifling matters.	☐	☐	☐	☐	☐	☐
41. I usually maintain my own opinions even though many other people may have a different point of view.	☐	☐	☐	☐	☐	☐
42. I never miss going to church.	☐	☐	☐	☐	☐	☐
43. There is usually only one best way to solve most problems.	☐	☐	☐	☐	☐	☐
44. I usually dislike to set aside a task that I have undertaken until it is finished.	☐	☐	☐	☐	☐	☐
45. I am usually able to keep at a job longer than most people.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
46. I scan newspapers rather than read them carefully.	☐	☐	☐	☐	☐	☐
47. I let myself "go" at a party.	☐	☐	☐	☐	☐	☐
48. As a youngster I enjoyed taking part in reckless stunts.	☐	☐	☐	☐	☐	☐
49. I like a great deal of variety in my work.	☐	☐	☐	☐	☐	☐
50. My friends consider me to be happy-go-lucky.	☐	☐	☐	☐	☐	☐
51. I like being where there is something going on all the time.	☐	☐	☐	☐	☐	☐
52. I like work that has lots of excitement.	☐	☐	☐	☐	☐	☐
53. I change my plans often.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat		Somewhat	Moderately	Definitely	
	UNCHARACTERISTIC				CHARACTERISTIC			
54. I like to take a chance just for the excitement.	☐	☐	☐		☐	☐	☐	
55. I often make people laugh.	☐	☐	☐		☐	☐	☐	
56. I like to do things on the spur of the moment.	☐	☐	☐		☐	☐	☐	
57. I like to work crossword puzzles.	☐	☐	☐		☐	☐	☐	
58. I like detailed work.	☐	☐	☐		☐	☐	☐	
59. I like to play chess.	☐	☐	☐		☐	☐	☐	
60. I usually notice the furniture arrangements in a strange house.	☐	☐	☐		☐	☐	☐	
61. I don't like having my plans changed.	☐	☐	☐		☐	☐	☐	

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
62. I like mathematics.	☐	☐	☐	☐	☐	☐
63. I usually think before I leap.	☐	☐	☐	☐	☐	☐
64. I like to solve complex problems.	☐	☐	☐	☐	☐	☐
65. I like work requiring patience and carefulness.	☐	☐	☐	☐	☐	☐
66. I don't like changes.	☐	☐	☐	☐	☐	☐
67. I consider myself always careful.	☐	☐	☐	☐	☐	☐
68. When I have to carry through some project, I prefer working on it with interested colleagues, rather than on my own.	☐	☐	☐	☐	☐	☐

	Definitely	Moderately	Somewhat	Somewhat	Moderately	Definitely
	UNCHARACTERISTIC			CHARACTERISTIC		
69. I often feel that I must discuss something I've read before I'll really understand or remember it.	☐	☐	☐	☐	☐	☐
70. I usually solve a problem better by discussing it with others than by studying it alone.	☐	☐	☐	☐	☐	☐
71. I find it helpful to discuss a problem with others before coming to a decision.	☐	☐	☐	☐	☐	☐
72. I would prefer to learn about something by class discussion, rather than by reading a book on the subject in my own time.	☐	☐	☐	☐	☐	☐
73. I would prefer a teacher who neglects me and leaves me to my own devices over one who continually watches me and makes suggestions.	☐	☐	☐	☐	☐	☐
74. When required to make a number of decisions in a comparatively short time, I prefer to make them alone rather than with the help of others.	☐	☐	☐	☐	☐	☐
75. I am bothered when someone offers me advice I didn't ask for.	☐	☐	☐	☐	☐	☐
76. I like to do my planning alone, without suggestions from or discussions with other people.	☐	☐	☐	☐	☐	☐

77. I like working alone.

Definitely	Moderately	Somewhat		Somewhat	Moderately	Definitely
UNCHARACTERISTIC				CHARACTERISTIC		
☐	☐	☐		☐	☐	☐

APPENDIX III

Types of Testing Formats

DIRECTIONS FOR MULTIPLE CHOICE

You are advised to use your time effectively and to work as rapidly as you can without losing accuracy. Do not waste your time on questions that are too difficult for you. Go on to the other questions and come back to the difficult ones later if you can.

Be sure you understand the directions before attempting to answer any questions.

YOU ARE TO INDICATE ALL YOUR ANSWERS ON THE SEPARATE ANSWER SHEET. No credit will be given for anything written on any other paper. After you have decided which of the suggested answers is correct, mark the answer in the space on the right. Give only one answer to each question. If you change an answer, be sure that all previous marks are erased completely.

EXAMPLE

1. Chicago is a

(A) state
(B) city
(C) country
(D) continent

SAMPLE ANSWER

1. B

Answer Sheet for Multiple Choice

1. _____

2. _____

3. _____

4. _____

5. _____

6. _____

7. _____

8. _____

9. _____

10. _____

11. _____

12. _____

13. _____

14. _____

15. _____

DIRECTIONS FOR MULTIPLE CHOICE: PICK-ONE

The test that you are about to take is a little different than most other tests you have taken while in the Air Force. You are advised to use your time effectively and to work as rapidly as you can without losing accuracy. Do not waste your time on questions that are too difficult for you. Go on to the other questions and come back to the difficult ones later if you can.

BE SURE YOU UNDERSTAND THE DIRECTIONS BEFORE ATTEMPTING TO ANSWER ANY QUESTIONS.

YOU ARE TO INDICATE ALL YOUR ANSWERS ON THE SEPARATE ANSWER SHEET. No credit will be given for anything written on any other paper. Read the question carefully and read each alternative. After you have decided which of the suggested answers is correct, mark that answer on the line on the left hand side of the answer sheet in the space beside the question number. BE SURE THAT YOU MARK THE ANSWER CLEARLY. Give only one answer to each question. If you change your answer, be sure that all previous marks are erased completely.

Now, you are asked to indicate how sure you are that the answer you just marked was correct. This is to be indicated in the column to the right of your first mark. If you are very sure your answer was the correct one, check the top line. If you cannot make such a strong statement about your answer but are sure your answer was correct, mark the second line from the top. If you are fairly sure your answer was correct, mark the middle line. If you are not very sure of your answer but are not making a complete guess, mark the fourth line from the top. If each alternative seems equally possible and you guess one from these, mark the bottom line.

EXAMPLE 1

Chicago is a

(A) state
(B) city
(C) country
(D) village

EXAMPLE 1. ___B___

__X__ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

In the above example the subject was 100% sure that the answer he marked was the correct answer.

EXAMPLE 2

Rantoul is in what Illinois county?

(A) Ford
(B) Champaign
(C) Mercer
(D) Greene

EXAMPLE 2. ____B____

_____ very sure
_____ sure
__X__ fairly sure
_____ not very sure
_____ not sure

In the case above the subject quickly eliminated (C) and (D). He was not sure of his choice between the other two possible answers. He knew that it was either Ford or Champaign but did not know which of the two was correct. He decided to pick B and indicated his lack of sureness by marking the middle line.

EXAMPLE 3

Who is the present Postmaster General

(A) Volpe
(B) Richardson
(C) Blount
(D) Hickle

EXAMPLE 3. ____C____

_____ very sure
_____ sure
_____ fairly sure
__X__ not very sure
_____ not sure

In this case the subject knew that Hickle was the Secretary of the Interior and so that could be eliminated. He had heard the other three names on the news recently but couldn't remember their cabinet offices. For some reason Blount's name seemed to stick in his mind and so he chose this as the correct alternative. Since he couldn't really tell the difference between the first three alternatives, he indicated that he was not very sure of his answer.

EXAMPLE 4

The capitol of Illinois is

(A) Chicago
(B) Springfield
(C) Rantoul
(D) Decatur

EXAMPLE 3. _____B_____

_____ very sure
__X__ sure
_____ fairly sure
_____ not very sure
_____ not sure

In this case the subject knew that neither Rantoul nor Decatur was the capitol city. He thought that Chicago might be the capitol, but there was only a slim chance of this being correct. He therefore chose (B) as being correct and indicated that he was not completely sure of his answer by checking the second line.

EXAMPLE 5

Webster's NEW COLLEGIATE DICTIONARY has how many pages?

(A) 1175
(B) 1352
(C) 1189
(D) 1174

EXAMPLE 5. _____D_____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
__X__ not sure

In this case the subject didn't know which one was the answer. Each alternative looked like it was possible. The subject guessed at D being the correct answer.

Answer Sheet for Pick-One

1. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

2. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

3. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

4. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

5. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

6. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

7. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

8. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

9. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

10. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

11. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

12. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

13. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

14. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

15. _____

_____ very sure
_____ sure
_____ fairly sure
_____ not very sure
_____ not sure

DIRECTIONS FOR MULTIPLE CHOICE: DISTRIBUTE 100 POINTS

The test that you are about to take is a little different than most other tests you have taken while in the Air Force. You are advised to use your time effectively and to work as rapidly as you can without losing accuracy. Do not waste your time on questions that are too difficult for you. Go on to the other questions and come back to the difficult ones later if you can.

Be sure you understand the directions before attempting to answer any questions.

YOU ARE TO INDICATE ALL YOUR ANSWERS ON THE SEPARATE ANSWER SHEET. No credit will be given for anything written on any other paper. Read the question carefully and read each alternative. After you have decided which of the suggested answers is correct, mark that answer on the line on the left hand side of the answer sheet in the space beside the question number. BE SURE THAT YOU MARK THE ANSWER CLEARLY. Give only one answer to each question. If you change your answer, be sure that all previous marks are erased completely.

Now you are asked to indicate how sure you are that the answer you gave was correct. You are given 100 points to distribute over the possible alternatives. You are to distribute the 100 points over the alternatives as they appear on the right hand column of your answer sheet. If you are completely sure of your answer, place your 100 points all on that alternative. If you are making a complete guess, place 25 points on each alternative.

The first step is to decide what alternatives are completely wrong and place zero points on those alternatives. You should show your confidence in an alternative by the number of points you assign to it. The alternative which you marked on the left should always have the largest number of points placed beside it on the right. You should keep in mind that the points you indicate to the right should add up to 100, no more and no less.

Example 1

Chicago is a

(A) state
(B) city
(C) country
(D) village

EXAMPLE 1 B

0	(A)
100	(B)
0	(C)
0	(D)

In the example above the subject was 100% sure that the answer he marked was the correct answer.

EXAMPLE 2.

Webster's NEW COLLEGIATE DICTIONARY has how many pages?

(A) 1175
(B) 1352
(C) 1189
(D) 1174

EXAMPLE 2 D

25	(A)
25	(B)
25	(C)
25	(D)

In this case the subject didn't know the answer. As far as he was concerned each of the alternatives were equally possible. He chose D as being his response, but this was only a guess.

EXAMPLE 3.

The capitol of Illinois is

(A) Chicago
(B) Springfield
(C) Rantoul
(D) Decatur

EXAMPLE 3 B

25	(A)
75	(B)
0	(C)
0	(D)

In this case the subject knew that neither Rantoul nor Decatur was the capitol for sure. He thought that Chicago might be the capitol but there was only a slim chance of that being true. He was pretty sure that Springfield was the capitol but did not want to place all of his confidence in that answer.

EXAMPLE 4.

Who is the present Postmaster General?

(A) Volpe
(B) Richardson
(C) Blount
(D) Hickle

EXAMPLE 4 C

20	(A)
20	(B)
60	(C)
0	(D)

In the case above the subject knew that Hickle was the Secretary of the Interior and therefore not the Postmaster General. He therefore places no weight on alternative D. He has heard the other three names recently

on the radio but does not know for sure which one is the Postmaster General. He decides that Blount is probably the correct answer and decides to give that alternative 60 points. The other two alternatives appear to be equally likely so he gives them each 20 points.

EXAMPLE 5.

Rantoul is in what Illinois county?

(A) Ford
(B) Champaign
(C) Mercer
(D) Greene

EXAMPLE 5 B

50 (A)
50 (B)
0 (C)
0 (D)

In this case the subject quickly eliminated C and D. He therefore assigned them zero points. The choice between A and B was a toss-up. He decided to choose B but this was only a guess. He felt that alternative A had the same chance of being correct.

Many people wonder how to mark their confidence in terms distributing the 100 points so that they will make the highest possible score. This test is scored in such a way that you will do your best in the long run if you distribute the 100 points as honestly as you possibly can. The more points you place on the right answer the higher will be your score. Only the number of points placed on the correct answer will be scored. If you place all of your points on only one alternative and that alternative is wrong, you will not get any score. On the other hand, if you put 25 points on each alternative, you will not receive as high a score as you would if you placed more points on the correct answer.

REMEMBER THIS: THE MORE ACCURATELY YOU DISTRIBUTE THE 100 POINTS, THE HIGHER YOU CAN EXPECT YOUR SCORE TO BE.

ALSO: BE SURE YOUR POINTS ADD TO 100.

Many people wonder how to mark their sureness for questions about which they are not completely certain. As it turns out, the test is scored so that you will do your best in the long run if you indicate how sure you were as honestly as you possibly can. If you mark your answer as being very sure and in fact your answer is wrong, you will get a lower score than you would if you had marked not sure. On the other hand, it doesn't pay to be too conservative either. If you mark an answer as being not sure when in fact it is correct, you will get a lower score than if you would had you marked very sure.

REMEMBER THIS: THE MORE ACCURATELY YOU INDICATE YOUR SURENESS, THE HIGHER YOU CAN EXPECT YOUR SCORE TO BE.

Answer Sheet for Distribute 100 Points

1. _____

A. _____

B. _____

C. _____

D. _____

2. _____

A. _____

B. _____

C. _____

D. _____

3. _____

A. _____

B. _____

C. _____

D. _____

4. _____

A. _____

B. _____

C. _____

D. _____

5. _____

A. _____

B. _____

C. _____

D. _____

6. _____

A. _____

B. _____

C. _____

D. _____

7. _____

A. _____

B. _____

C. _____

D. _____

8. _____

A. _____

B. _____

C. _____

D. _____

9. _____

A. _____

B. _____

C. _____

D. _____

10. _____

A. _____

B. _____

C. _____

D. _____

11. _____

A. _____

B. _____

C. _____

D. _____

12. _____

A. _____

B. _____

C. _____

D. _____

13. _____

A. _____

B. _____

C. _____

D. _____

14. _____

A. _____

B. _____

C. _____

D. _____

15. _____

A. _____

B. _____

C. _____

D. _____

APPENDIX IV

Student Attitudes

ATTITUDES TOWARDS TESTING FORM CON1

1. How well do you think the results of the test showed what you really know?

   A. _____ very satisfactorily
   B. _____ satisfactorily
   C. _____ so-so
   D. _____ not very satisfactorily
   E. _____ not satisfactorily

2. To what extent are you satisfied with classroom testing as it helps <u>you</u> find out what you know and what you don't know?

   A. _____ to a large extent
   B. _____ to an extent
   C. _____ to some extent
   D. _____ to a little extent
   E. _____ not at all

3. How satisfied, in general, are you with the testing that takes place in this class?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ I can take it or leave it
   D. _____ not very satisfied
   E. _____ not satisfied

4. How advantageous do you find testing in your course?

   A. _____ very advantageous
   B. _____ advantageous
   C. _____ doubtful
   D. _____ disadvantageous
   E. _____ very disadvantageous

IF YOU HAVE EVER BEEN IN A REMEDIAL SESSION, ANSWER NUMBERS 5 AND 6, OTHERWISE DO NOT ANSWER THEM.

5. How satisfied are you with classroom testing as it helps you with remediation?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ so-so
   D. _____ not very satisfied
   E. _____ not satisfied

6. How well do your test results help you during remediation?

   A. _____ very well
   B. _____ well
   C. _____ I can't really say
   D. _____ not very well
   E. _____ not well

ATTITUDES TOWARDS TESTING FORM CON2

1. How well do you think the results of the test showed what you really know?

   A. _____ very satisfactorily
   B. _____ satisfactorily
   C. _____ so-so
   D. _____ not very satisfactorily
   E. _____ not satisfactorily

2. To what extent are you satisfied with classroom testing as it helps you find out what you know and what you don't know?

   A. _____ to a large extent
   B. _____ to an extent
   C. _____ to some extent
   D. _____ to a little extent
   E. _____ not at all

3. How satisfied, in general, are you with the testing that takes place in this class?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ I can take it or leave it
   D. _____ not very satisfied
   E. _____ not satisfied

4. How advantageous do you find testing in your course?

   A. _____ very advantageous
   B. _____ advantageous
   C. _____ doubtful
   D. _____ disadvantageous
   E. _____ very disadvantageous

IF YOU HAVE BEEN TO A REMEDIAL SESSION, ANSWER NUMBERS 5 AND 6, OTHERWISE, GO TO NUMBER 7.

5. How satisfied are you with classroom testing as it helps you with remediation?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ so-so
   D. _____ not very satisfied
   E. _____ not satisfied

6. How well do your test results help you during remediation?

   A. _____ very well
   B. _____ well
   C. _____ I can't really say
   D. _____ not very well
   E. _____ not well

7. How important was it for you to score high on your tests?

A. _____ very important
B. _____ important
C. _____ so-so
D. _____ not very important.
E. _____ not important

8. How well did you understand how the tests were graded?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

9. How well did you understand the ways of marking your answers?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

10. When you came to a question that you did not know, how accurately do you think you marked your confidence?

A. _____ very accurately
B. _____ accurately
C. _____ fairly accurately
D. _____ not very accurately
E. _____ not accurately

11. How easy was it for you to decide how to mark your answers?

A. _____ very easy
B. _____ easy
C. _____ so-so
D. _____ fairly difficult
E. _____ difficult

12. How well do you think your confidence marks showed your real confidence?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

13. How honest were you in making your confidence marks?

A. _____ very honest
B. _____ honest
C. _____ fairly honest
D. _____ not very honest
E. _____ not honest

14. How comfortable did you feel when you took tests where you marked your confidence?

A. _____ very comfortable
B. _____ comfortable
C. _____ a little uneasy
D. _____ uneasy
E. _____ uncomfortable

15. How often did you gamble in answering?

A. _____ very often
B. _____ often
C. _____ a few times
D. _____ once or twice
E. _____ never

16. To what extent do you think it was in your best interest to mark your confidence accurately?

A. _____ to a large extent
B. _____ to an extent
C. _____ to a slight extent
D. _____ to a small extent
E. _____ not at all

17. How well did you like taking multiple choice tests where you picked the answer you thought was right and then told how sure you were the answer was right?

A. _____ very well
B. _____ well
C. _____ so-so
D. _____ not very well
E. _____ not well

18. To what degree has your prior experience in taking tests by more conventional methods affected your use of the testing format that you have been using in this class?

A. _____ to a great degree
B. _____ to a degree
C. _____ to a slight degree
D. _____ to a very small degree
E. _____ not at all

19. Compared to the usual method of taking a test, how would you rate the testing you have been using as a method of relearning?

A. _____ very well
B. _____ well
C. _____ I can't really say
D. _____ not very well
E. _____ not well

20. To what extent do you agree or disagree with these strengths and weaknesses of the confidence testing system you have been using? Check the appropriate blank for each statement.

	strongly agree	agree	doubtful	don't agree that much	do not agree
It better identifies my strengths and weaknesses.	____	____	____	____	____
It allows instructor to better reteach material.	____	____	____	____	____
It reduces guessing.	____	____	____	____	____
It identifies level of knowledge helpful to me.	____	____	____	____	____
It is fairer than most systems.	____	____	____	____	____
It requires more thought before making a response.	____	____	____	____	____
It is a useful device for relearning material.	____	____	____	____	____
It is difficult to overcome old testing habits.	____	____	____	____	____
It tends to make me lose confidence in selecting one answer.	____	____	____	____	____
It tends to make me hedge in answers - play safe.	____	____	____	____	____
Uninformed students try to beat the system.	____	____	____	____	____
The method encourages guessing.	____	____	____	____	____

ATTITUDES TOWARDS TESTING FORM CON3

1. How well do you think the results of the test showed what you really know?

   A. _____ very satisfactorily
   B. _____ satisfactorily
   C. _____ so-so
   D. _____ not very satisfactorily
   E. _____ not satisfactorily

2. To what extent are you satisfied with classroom testing as it helps <u>you</u> find out what you know and what you don't know?

   A. _____ to a large extent
   B. _____ to an extent
   C. _____ to some extent
   D. _____ to a little extent
   E. _____ not at all

3. How satisfied, in general, are you with the testing that takes place in this class?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ I can take it or leave it
   D. _____ not very satisfied
   E. _____ not satisfied

4. How advantageous do you find testing in your course?

   A. _____ very advantageous
   B. _____ advantageous
   C. _____ doubtful
   D. _____ disadvantageous
   E. _____ very disadvantageous

IF YOU HAVE BEEN TO A REMEDIAL SESSION, ANSWER NUMBERS 5 AND 6, OTHERWISE, GO TO NUMBER 7.

5. How satisfied are you with classroom testing as it helps you with remediation?

   A. _____ very satisfied
   B. _____ satisfied
   C. _____ so-so
   D. _____ not very satisfied
   E. _____ not satisfied

6. How well do your test results help you during remediation?

   A. _____ very well
   B. _____ well
   C. _____ I can't really say
   D. _____ not very well
   E. _____ not well

7. How important was it for you to score high on your tests?

A. _____ very important
B. _____ important
C. _____ so-so
D. _____ not very important
E. _____ not important

8. How well did you understand how the tests were graded?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

9. How well did you understand the ways of marking your answers?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

10. When you came to a question that you did not know, how accurately do you think you marked your confidence?

A. _____ very accurately
B. _____ accurately
C. _____ fairly accurately
D. _____ not very accurately
E. _____ not accurately

11. How easy was it for you to decide how to mark your answers?

A. _____ very easy
B. _____ easy
C. _____ so-so
D. _____ fairly difficult
E. _____ difficult

12. How well do you think your confidence marks showed your real confidence?

A. _____ very well
B. _____ well
C. _____ fairly well
D. _____ not very well
E. _____ not at all

13. How honest were you in making your confidence marks?

A. _____ very honest
B. _____ honest
C. _____ fairly honest
D. _____ not very honest
E. _____ not honest

14. How comfortable did you feel when you took tests where you marked your confidence?

A. _____ very comfortable
B. _____ comfortable
C. _____ a little uneasy
D. _____ uneasy
E. _____ uncomfortable

15. How often did you gamble in answering?

A. _____ very often
B. _____ often
C. _____ a few times
D. _____ once or twice
E. _____ never

16. To what extent do you think it was in your best interest to mark your confidence accurately?

A. _____ to a large extent
B. _____ to an extent
C. _____ to a slight extent
D. _____ to a small extent
E. _____ not at all

17. How difficult was it for you to distribute the 100 points and make sure they added to 100?

A. _____ very difficult
B. _____ difficult
C. _____ so-so
D. _____ not very difficult
E. _____ not difficult

18. How well did you like taking multiple choice tests where you distributed 100 points over the possible answers?

A. _____ very well
B. _____ well
C. _____ I can't really say
D. _____ not very well
E. _____ not well

19. To what degree has your prior experience in taking tests by more conventional methods affected your use of the testing format that you have been using in this class?

A. _____ to a great degree
B. _____ to a degree
C. _____ to a slight degree
D. _____ to a very small degree
E. _____ not at all

20. Compared to the usual method of taking a test, how would you rate the testing you have been using as a method of relearning?

A. _____ very well
B. _____ well
C. _____ I can't really say
D. _____ not very well
E. _____ not well

21. To what extent do you agree or disagree with these strengths and weaknesses of the confidence testing system you have been using? Check appropriate blank for each statement.

	strongly agree	agree	doubtful	don't agree that much	do not agree
It better identifies my strengths and weaknesses.	____	____	____	____	____
It allows instructor to better reteach material.	____	____	____	____	____
It reduces guessing.	____	____	____	____	____
It identifies level of knowledge helpful to me.	____	____	____	____	____
It is fairer than most systems.	____	____	____	____	____
It requires more thought before making a response.	____	____	____	____	____
It is a useful device for relearning material.	____	____	____	____	____
It is difficult to overcome old testing habits.	____	____	____	____	____
It tends to make me lose confidence in selecting one answer.	____	____	____	____	____
It tends to make me hedge in answers - play safe.	____	____	____	____	____
Uninformed students try to beat the system.	____	____	____	____	____
The method encourages guessing.	____	____	____	____	____

APPENDIX V

Detailed Analysis of Block Scores

Table 1

Means and Standard Deviations of Block Scores by
Type of Testing, Type of Remediation, and Shift in AGE

Type of Testing	Type of Remediation	Shift	N		Block 6	Block 7	Block 8
M. C.	Cont.	A	12	Mean	90.3	87.8	79.1
				S D	10.3	13.8	17.0
M. C.	Cont.	C	4	"	79.8	88.5	86.5
					4.8	3.0	1.7
M. C.	Cont.	D	10	"	93.9	88.4	83.3
					3.9	8.7	4.8
M. C.	Spec.	A	10	"	80.4	85.1	86.6
					8.2	14.1	7.9
M. C.	Spec.	B	8	"	91.5	80.1	88.3
					3.3	6.6	4.8
M. C.	Spec.	D	9	"	88.7	87.7	86.6
					8.0	8.9	8.5
P. O.	Cont.	A	10	"	85.4	85.5	86.9
					6.6	7.4	5.7
P. O.	Cont.	C	8	"	89.0	82.8	81.6
					8.2	10.9	9.3
P. O.	Cont.	D	15	"	88.3	86.9	86.7
					6.6	5.7	7.8
P. O.	Spec.	A	10	"	83.8	80.9	81.1
					4.9	8.7	6.0
P. O.	Spec.	B	8	"	88.0	88.8	86.9
					6.4	5.8	3.4
P. O.	Spec.	C	12	"	82.3	84.4	84.9
					7.1	9.4	6.6
100 P	Cont.	A	9	"	88.1	83.2	76.8
					5.3	5.9	6.9
100 P	Cont.	B	16	"	89.1	91.5	88.9
					7.4	6.6	4.8
100 P	Cont.	C	8	"	89.9	87.1	82.4
					6.6	6.4	6.7
100 P	Cont.	D	10	"	91.4	87.5	91.4
					4.9	11.3	4.6
100 P	Spec.	A	10	"	86.2	87.4	86.6
					5.2	7.4	8.5
100 P	Spec.	C	12	"	90.0	77.9	80.8
					5.0	8.4	6.7
100 P	Spec.	D	11	"	87.8	81.0	84.5
					7.4	10.8	6.6

Table 2

Multivariate Test of the Type of Testing x Shift Interaction in AGE Using Wilks' Lambda Criterion

Test of Roots	F	Degrees of Freedom For Hypothesis	Degrees of Freedom For Error	p less than
1 Through 3	2.584	18	484.146	0.001
2 Through 3	1.891	10	474.957	0.044
3 Through 3	0.062	4	456.715	0.993

Table 3

Multivariate Test of the Type of Testing Effects Within Shift D in AGE

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	1.769	6	342.0	0.105
2 Through 2	0.848	2	171.5	0.430

Univariate F tests

Variable	F(2,173)	Mean Square	p less than	Standardized Discriminant Coefficients
Block 6 Score	0.832	37.318	0.437	0.623
Block 7 Score	0.934	75.493	0.395	0.789
Block 8 Score	0.767	45.048	0.466	-1.164

Table 4

Multivariate Test of the Type of Testing Effects Within Shift C in AGE

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	3.562	6	342.0	0.002
2 Through 2	0.103	2	171.5	0.902

Univariate F tests

Variable	F(2,173)	Mean Square	p less than	Standardized Discriminant Coefficients
Block 6 Score	4.595	206.099	0.011	1.150
Block 7 Score	0.732	59.178	0.482	-0.326
Block 8 Score	0.736	43.229	0.481	-0.700

Table 5

Multivariate Test of the Type of Testing
Effects Within Shift B in AGE

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	2.424	6	342.0	0.026
2 Through 2	0.201	2	171.5	0.818

Univariate F tests

Variable	F(2,173)	Mean Square	p less than	Standardized Discriminant Coefficients
Block 6 Score	0.581	26.062	0.560	0.542
Block 7 Score	4.307	348.293	0.015	-1.140
Block 8 Score	0.182	10.671	0.834	0.304

Table 6

Multivariate Test of the Type of Testing
Effects Within Shift A in AGE

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	1.259	6	342.0	0.275
2 Through 2	0.651	2	171.5	0.523

Univariate F tests

Variable	F(2,173)	Mean Square	p less than	Standardized Discriminant Coefficients
Block 6 Score	0.744	33.361	0.477	0.813
Block 7 Score	0.686	55.448	0.505	0.625
Block 8 Score	0.442	25.992	0.643	-1.139

Table 7

Means and Standard Deviations of Block Scores by Type of Testing, Type of Remediation, and Shift in JEM

Type of Testing	Type of Remediation	Shift	N		Block 2	Block 3
M.C.	Cont.	A	18	Mean	84.5	83.0
				S.D.	8.1	7.5
M.C.	Cont.	B	24	"	88.5	89.0
					6.2	5.9
M.C.	Spec.	A	6	"	77.0	81.3
					5.5	3.6
M.C.	Spec.	B	8	"	78.9	77.8
					7.3	5.3
P.O.	Cont.	A	26	"	91.3	84.2
					5.6	6.7
P.O.	Cont.	B	29	"	85.7	88.0
					7.6	6.5
P.O.	Spec.	A	25	"	89.6	87.1
					5.9	4.9
P.O.	Spec.	B	27	"	85.6	88.8
					7.7	4.6
100P	Cont.	A	27	"	89.8	88.5
					6.7	5.8
100P	Cont.	B	15	"	87.1	85.1
					8.2	5.0
100P	Spec.	A	17	"	81.3	81.8
					5.9	6.3
100P	Spec.	B	32	"	85.8	83.4
					7.0	7.4

Table 8

Multivariate Test of the Type of Testing x Type of Remediation x Shift Interaction in JEM Using Wilks' Lambda Criterion

Test of Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 through 2	2.585	4	482.000	0.036
2 through 2	0.110	1	241.500	0.740

Table 9

Multivariate Test of the Type of Testing
Effects Within Shift B Special Remediation in JEM

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	7.363	4	482.000	0.001
2 Through 2	4.878	1	241.500	0.028

Univariate F tests

Variable	F(2,242)	Mean Square	p less than	Standardized Discriminant Coefficients 1	2
Block 2 Score	3.496	165.707	0.032	-0.295	1.157
Block 3 Score	11.524	443.185	0.001	1.130	-0.386

Table 10

Multivariate Test of the Type of Testing
Effects Within Shift A Special Remediation in JEM

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	6.028	4	482.000	0.001
2 Through 2	0.610	1	241.500	0.436

Univariate F tests

Variable	F(2,242)	Mean Square	p less than	Standardized Discriminant Coefficients
Block 2 Score	12.001	568.557	0.001	0.946
Block 3 Score	4.686	174.174	0.010	0.093

Table 11

Multivariate Test of the Type of Testing
Effects Within Shift B Control Remediation in JEM

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	1.902	4	482.000	0.109
2 Through 2	2.105	1	241.500	0.148

Univariate F tests

Variable	F(2,242)	Mean Square	p less than	Standardized Discriminant Coefficients 1	2
Block 2 Score	1.092	51.712	0.337	-0.792	0.894
Block 3 Score	2.015	74.899	0.136	1.182	0.175

Table 12

Multivariate Test of the Type of Testing
Effects Within Shift A Control Remediation in JEM

Tests of significance using Wilks' lambda criterion

Tests of the Roots	F	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	p less than
1 Through 2	5.894	4	482.000	0.001
2 Through 2	9.892	1	241.500	0.002

Univariate F tests

Variable	F(2,242)	Mean Square	p less than	Standardized Discriminant Coefficients 1	2
Block 2 Score	5.450	258.176	0.005	-1.069	0.533
Block 3 Score	5.340	198.477	0.005	1.031	0.604

APPENDIX VI

Frequency Distributions of Responses to Student Attitude Questionnaires

Appendix VI

Frequency Distributions of Responses to Student Attitude Questionnaires

Question	AGE MULT CH	AGE PICK ONE	AGE DIST 100	JEM MULT CH	JEM PICK ONE	JEM DIST 100
1. How well do you think the results of the test showed what you really know?**						
A. very satisfactorily	3	3	3	8	14	8
B. satisfactorily	16	12	13	31	38	43
C. so-so	5	7	13	8	21	25
D. not very satisfactorily	0	1	7	6	14	8
E. not satisfactorily	1	3	4	3	2	0
2. To what extent are you satisfied with classroom testing as it helps you find out what you know and what you don't know.**						
A. to a large extent	10	12	9	24	28	30
B. to an extent	10	7	7	17	25	31
C. to some extent	4	6	15	10	23	21
D. to a little extent	0	0	7	5	8	4
E. not at all	1	1	2	0	6	3
3. How satisfied, in general, are you with the testing that takes place in this class?*, **						
A. very satisfied	2	2	2	15	14	17
B. satisfied	17	15	14	30	49	58
C. I can take it or leave it	3	3	13	3	10	11
D. not very satisfied	2	3	9	8	14	1
E. not satisfied	1	3	2	0	3	2
4. How advantageous do you find testing in your course?*						
A. very advantageous	6	10	5	22	28	17
B. advantageous	15	10	20	28	40	64
C. doubtful	2	5	13	6	18	6
D. disadvantageous	2	1	2	0	3	2
E. very disadvantageous	0	0	0	0	0	0

*Indicates significant chi-square at .05 level in JEM.

**Indicates significant chi-square at .05 level in AGE.

Appendix VI (Continued)

Question	AGE MULT CH	AGE PICK ONE	AGE DIST 100	JEM MULT CH	JEM PICK ONE	JEM DIST 100
5. How satisfied are you with classroom testing as it helps you with remediation?						
A. very satisfied	2	4	1	9	11	8
B. satisfied	11	10	13	10	3	1
C. so-so	6	6	9	3	4	7
D. not very satisfied	2	1	3	2	1	2
E. not satisfied	1	1	2	0	1	0
6. How well do your test results help you during remediation?						
A. very well	3	7	3	9	6	8
B. well	9	5	6	8	7	11
C. I can't really say	7	7	15	7	3	7
D. not very well	1	2	1	0	3	0
E. not well	2	1	3	0	1	1

Appendix VI (Continued)

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
7. How important was it for you to score high on your tests?				
A. very important	7	14	52	55
B. important	12	14	26	31
C. so-so	6	7	7	7
D. not very important	1	1	3	2
E. not important	0	3	1	1
8. How well did you understand how the tests were graded?*				
A. very well	11	9	28	29
B. well	4	8	22	33
C. fairly well	3	11	18	21
D. not very well	3	6	15	4
E. not at all	5	6	6	2
9. How well did you understand the ways of marking your answers?				
A. very well	13	13	47	46
B. well	4	17	24	27
C. fairly well	8	6	11	14
D. not very well	0	4	6	0
E. not at all	1	0	1	2
10. When you came to a question that you did not know, how accurately do you think you marked your confidence?				
A. very accurately	0	3	10	13
B. accurately	10	11	25	29
C. fairly accurately	13	20	42	37
D. not very accurately	1	5	9	9
E. not accurately	2	1	3	1

*Indicates significant chi-square at .05 level in JEM.

Appendix VI (Continued)

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
11. How easy was it for you to decide how to mark your answers?				
A. very easy	7	5	24	15
B. easy	13	14	36	48
C. so-so	5	19	19	19
D. fairly difficult	0	2	8	4
E. difficult	1	0	2	3
12. How well do you think your confidence marks showed your real confidence?				
A. very well	5	3	20	16
B. well	7	7	30	31
C. fairly well	7	17	26	28
D. not very well	5	6	9	9
E. not at all	2	7	4	5
13. How honest were you in making your confidence marks?				
A. very honest	9	7	30	27
B. honest	6	11	39	39
C. fairly honest	9	17	16	20
D. not very honest	1	3	3	0
E. not honest	1	2	1	3
14. How comfortable did you feel when you took tests where you marked your confidence?*				
A. very comfortable	8	5	20	12
B. comfortable	11	15	49	43
C. a little uneasy	5	15	14	29
D. uneasy	0	4	1	4
E. uncomfortable	2	1	5	1

*Indicates significant chi-square at .05 level in JEM.

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
15. How often did you gamble in answering?				
A. very often	3	5	1	5
B. often	3	7	11	11
C. a few times	35	19	52	41
D. once or twice	5	9	24	27
E. never	0	0	1	2
16. To what extent do you think it was in your best interest to mark your confidence accurately?				
A. to a large extent	8	5	19	20
B. to an extent	11	14	42	39
C. to a slight extent	5	10	15	16
D. to a small extent	0	7	7	8
E. not at all	2	4	6	6
17. How well did you like taking multiple choice tests where you picked the answer you thought was right and then told how sure you were the answer was right?				
A. very well	6	7	25	20
B. well	9	9	22	35
C. so-so	6	14	25	19
D. not very well	1	5	10	4
E. not well	4	5	7	11
18. To what degree has your prior experience in taking tests by more conventional methods affected your use of the testing format that you have been using in this class?				
A. to a great degree	4	5	8	11
B. to a degree	7	11	38	31
C. to a slight degree	7	17	28	24
D. to a very small degree	3	3	4	10
E. not at all	5	4	11	13

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
19. Compared to the usual method of taking a test, how would you rate the testing you have been using as a method of relearning?				
A. very well	6	4	22	14
B. well	8	9	22	39
C. I can't really say	9	15	35	26
D. not very well	1	6	10	8
E. not well	2	5	0	1
To what extent do you agree or disagree with these strengths and weaknesses of the confidence testing system you have been using?				
20. It better identifies my strengths and weaknesses.				
A. strongly agree	3	5	13	12
B. agree	11	13	47	45
C. doubtful	6	13	17	20
D. don't agree that much	0	6	1	5
E. do not agree	5	3	10	5
21. It allows instructor to better reteach material.				
A. strongly agree	6	3	25	12
B. agree	10	15	36	58
C. doubtful	5	11	15	10
D. don't agree that much	1	6	2	3
E. do not agree	3	4	10	4
22. It reduces guessing.				
A. strongly agree	1	3	9	4
B. agree	5	11	23	33
C. doubtful	11	12	20	20
D. don't agree that much	2	7	14	14
E. do not agree	6	7	22	16

Appendix VI (Continued)

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
23. It identifies level of knowledge helpful to me.*				
A. strongly agree	3	2	12	4
B. agree	13	15	43	60
C. doubtful	5	18	18	11
D. don't agree that much	0	5	8	6
E. do not agree	4	0	7	4
24. It is fairer than most systems.				
A. strongly agree	1	6	12	7
B. agree	12	17	37	47
C. doubtful	6	12	18	20
D. don't agree that much	2	1	10	7
E. do not agree	4	4	11	6
25. It requires more thought before making a response.*				
A. strongly agree	6	4	17	10
B. agree	13	19	28	50
C. doubtful	1	8	18	14
D. don't agree that much	3	2	9	7
E. do not agree	2	7	16	6
26. It is a useful device for relearning material.*				
A. strongly agree	3	10	21	10
B. agree	13	8	33	51
C. doubtful	5	14	15	13
D. don't agree that much	1	4	11	10
E. do not agree	3	4	8	3
27. It is difficult to overcome old testing habits.				
A. strongly agree	1	5	10	6
B. agree	5	11	14	18
C. doubtful	3	10	18	21
D. don't agree that much	7	8	20	18
E. do not agree	9	6	26	24

*Indicates significant chi-square at .05 level in JEM.

Appendix VI (Continued)

QUESTION	AGE PICK ONE	AGE DISTRIBUTE 100	JEM PICK ONE	JEM DISTRIBUTE 100
28. It tends to make me lose confidence in selecting one answer.				
A. strongly agree	0	1	6	5
B. agree	5	14	13	15
C. doubtful	6	12	22	26
D. don't agree that much	6	7	18	13
E. do not agree	8	6	29	27
29. It tends to make me hedge in answers - play safe.				
A. strongly agree	1	7	3	1
B. agree	6	10	22	24
C. doubtful	8	15	26	35
D. don't agree that much	4	5	17	14
E. do not agree	6	3	20	12
30. Uninformed students try to beat the system.				
A. strongly agree	2	2	6	5
B. agree	10	17	21	26
C. doubtful	6	13	31	27
D. don't agree that much	4	3	10	8
E. do not agree	3	4	20	21
31. The method encourages guessing.				
A. strongly agree	1	6	6	4
B. agree	3	6	10	16
C. doubtful	8	7	20	27
D. don't agree that much	6	4	12	13
E. do not agree	7	5	40	27

APPENDIX VII

Instructor Questionnaires

INSTRUCTOR INTERVIEW

1. Instructor Name:

2. Instructor class:

3. Do you look at your students' test results? If so, what do you look for?

No	Yes
1	36

Items Missed	Total Scores
34	3

4. Do you look at the questions that each student misses?

No	Yes
4	33

5. Do you correct the tests yourself?

No	Yes
3	34

6. Do you use a total score in assigning students to remediation or can you use responses to particular questions?

Total score 37

7. How do you determine the instruction in the remedial classes?

By questions missed	Miscellaneous
30	7

8. How far ahead are students scheduled for remediation?

Daily	Other
33	4

9. Are any people required to attend remediation for reasons other than poor test results? If so, for what other reasons are people assigned to remediation?

No	Yes
10	27

10. Does the attendance vary according to individuals problems with particular subjects or are people assigned to remedial sequences regularly until they have improved?

Assigned regularly	Both	Attendance varies
7	7	23

11. Have you had any instruction in how to use the test results in making remedial sessions?

No	Yes
18	19

FOR INSTRUCTORS OF CLASSES USING CONFIDENCE TESTING

12. How many people were you able to find who placed a large amount of confidence in wrong answers?

Few	About half	Many
30	4	3

13. How often, approximately, did students place all their confidence in their responses?

Seldom	Sometimes	Very often
2	2	33

14. Did you treat students who placed large amounts of confidence in wrong answers any differently?

No	Yes
30	7

15. Were separate remediation sessions set up for both students who placed large amounts of confidence in wrong answers and students who did not know the answer as evidenced by placing small amounts of confidence in their preferred answer?

No	Yes
37	0

16. Did the confidence test scores influence your decisions about the remediation?

No	Yes
31	6

17. What difficulties did you have in interpreting the confidence scores?

Some	Grading	Couldn't interpret
26	9	2

18. Were students pressed for time when taking the confidence test?

No	Yes
36	1

19. Did students have much difficulty in assigning their confidence?

No	Yes
36	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Educational Testing Service
Princeton, New Jersey 08540

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE
AN EVALUATION OF THE FEASIBILITY OF CONFIDENCE TESTING AS A DIAGNOSTIC AID IN TECHNICAL TRAINING

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Report (July 1970 to July 1971)

5. AUTHOR(S) *(First name, middle initial, last name)*
Gary J. Echternacht
Wayne S. Sellman
Robert F. Boldt
Joseph D. Young

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July 1971	127	22

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F41609-70-C-0044	AFHRL-TR-71-33
b. PROJECT NO. 1121	
c. Task No. 112103	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d. Work Unit No. 112103003	

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Technical Training Division Lowry Air Force Base, Colorado 80230

13. ABSTRACT

This report describes a study to determine the feasibility and the cost-effectiveness of using confidence testing as a diagnostic aid in technical training programs. Two types of confidence testing, Pick-One and Distribute 100 Points, were developed for comparison to conventional multiple-choice testing. The study was carried out in two technical training courses, Aerospace Ground Equipment Repairman (AGE) and Jet Engine Mechanic (JEM), currently being taught at Chanute Air Force Base, Illinois. The criteria for feasibility included end of block examination scores, number of student remediational sessions, and both student and instructor attitudes. In addition, the relationship of various personality variables to confidence test scores was examined for both types of confidence testing. The major finding was that while scoring was somewhat more time consuming, end of block examination scores improved slightly and the number of remediations required declined slightly when either confidence testing method was employed. Other areas of investigation produced essentially null results.

DD FORM 1 NOV 65 1473

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Psychological testing						
psychometrics						
confidence testing						
subjective probability						
technical training						