AD 655263

Technical Research Note 188

AD

AN EXPLORATORY STUDY OF BRANCHING TESTS

by A. G. Bayroff and Leonard C. Seeley

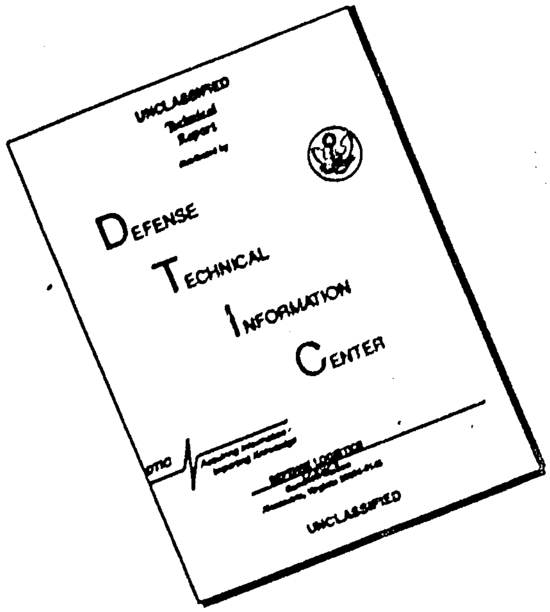
MILITARY SELECTION RESEARCH DIVISION

JUNE 1967

Distribution of this document is unlimited.

U. S. Army
Behavioral Science Research Laboratory

DISCLAIMER NOTICE



AD

Technical Research Note 188

AN EXPLORATORY STUDY OF BRANCHING TESTS

by A. G. Bayroff and Leonard C. Seeley

MILITARY SELECTION RESEARCH DIVISION
Edmund F. Fuchs, Chief

U. S. ARMY BEHAVIORAL SCIENCE RESEARCH LABORATORY

Office, Chief of Research and Development
Department of the Army

Washington, D. C. 20315

June 1967

Army Project Number
2J024701A722

Input Quality i-00

BESRL Technical Research Reports and Technical Research Notes are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The INPUT QUALITY Task conducts a continuing research program on screening and induction techniques. Objectives are (1) to improve the system for screening potential enlisted input so as to identify and reject more effectively those who are not readily trainable and usable in the service; (2) to aid in manpower planning by developing methods for estimating the mental abilities of the civilian pool available for service under various conditi ons; and (3) to develop technical information for use in consultative assistance to staff agencies responsible for procurement and standards policies.

As one avenue to the development of technical information and methods for improving input screening, the potential contribution of programmed testing has been explored. In this connection, experimental branching tests were developed. The present publication reports on the trial administration of two branching tests by means of a computerized system with teletypewriter input/output developed by the National Bureau of Standards. Research was a part of Subtask i, "An exploratory study of branching tests." The entire INPUT QUALITY Task is responsive to special requirements of the Deputy Chief of Staff for Personnel, as well as to requirements to contribute to achievement of the objectives of RDT&E Project 2J024701A722, "Selection and Behavioral Evaluation," FY 1967 Work Program.

J. E. UHLANER, Director
Behavioral Science
Research Laboratory

AN EXPLORATORY STUDY OF BRANCHING TESTS

BRIEF

REQUIREMENT:

To explore the comparability of computerized branching tests and conventional paper-and-pencil tests with respect to reliability, information conveyed by the test score, and rationale of test construction.

PROCEDURE:

Two specially constructed 8-9-item branching tests (verbal and arithmetic reasoning) and the corresponding conventional 40- and 50-item tests of the Army Classification Battery were administered to a sample of 102 enlisted men. Scores were analyzed to estimate reliability and to study relationships between corresponding branching and conventional tests.

FINDINGS:

The short branching tests were substantially correlated with counterpart longer conventional tests ($r = .83$ and $.79$, higher than would be expected with equally short conventional tests). Classical test theory developed for the construction of linear tests is not entirely appropriate in developing branching tests. For example, item difficulty indexes based on population performance are not fully appropriate when each examinee is tested with questions geared to his own ability level.

UTILIZATION OF FINDINGS:

The exploratory study reinforced the promise of branching tests and pointed to the need for reexamination of test theory.

AN EXPLORATORY STUDY OF BRANCHING TESTS

CONTENTS

AN EXPLORATORY STUDY OF BRANCHING TESTS

In the past several years, the U. S. Army Behavioral Science Research Laboratory (BESRL) has been interested in new approaches to testing which might prove to be improvements over conventional methods. One line of interest has been the branching technique. Branching is provided by programming a test so that an examinee who answers a test item correctly is presented next with a more difficult item, and an examinee who answers incorrectly is presented with an easier item. By contrast, the conventional test is linearly programmed so that all examinees answer the same items regardless of the correctness of their responses. The branching program has the potential of reducing error in test scores or of providing scores of validity equal to that of the linear test--but with fewer items. A preliminary study[1] had indicated the theoretical promise of branching tests. However, the promise could not be followed up in the format of the conventional printed test[2], and automated methods were required.

BESRL contracted with the National Bureau of Standards (NBS) to conduct a preliminary design study of a programmed testing machine which would meet a number of specified requirements, including the requirements for branching[3]. Following completion of the design study, NBS, out of its own interest in the technique, developed a computer system with teletypewriter input/output which was programmed to provide branching but not to meet the other requirements covered in the design study. NBS invited BESRL to use the system for exploratory research. The impending move of NBS to a new location made it necessary to act quickly.

Accordingly, two branching tests, one of verbal ability and the other of arithmetic reasoning, were assembled from item data readily available. These tests and counterpart conventional tests were administered to a group of enlisted men and the results were compared. The objective was to obtain indications of the research promise of the branching technique and to uncover some of the major problems likely to be encountered in a systematic study of the branching technique, as well as to provide NBS with a use test of its computer system.

---

[1] Waters, Carrie Jean. Preliminary evaluation of simulated branching test. BESRL Technical Research Note 140. June 1964.

[2] Seeley, L. C., Morton, Mary A., and Anderson, Alan A. Exploratory study of a sequential item test. BESRL Technical Research Note 129. December 1962.

[3] Bayroff, A. G. Feasibility of a programmed testing machine. BESRL Research Study 64-3. November 1964.

The Branching Tests

Items for the branching tests were selected from the experimental forms of the Armed Forces Qualification Test, AFQT 7-8 and AFQT 5-6. Selection was mainly of items not included in the operational forms of AFQT. Each test plan (Figure 1) called for a pool of items with a difficulty range of p = .95 to .25, beginning with an item of p = .60. All examinees were to answer 8 items with difficulty differences of p = .05. Examinees who reached the most difficult item (p = .25) and answered it correctly were to be presented with an additional item of greater difficulty (p = .20) as a means of increasing the ceiling. The items were selected to meet this plan as closely as possible. The four-choice items were modified by adding two incorrect choices as a means of reducing chance success.

The score was determined by the relative difficulty of the item reached in the final stage. This stage had a difficulty range of p = .95 to .20 and provided a scale with a raw score range of 1 to 17. Each of these final items had two score values--a score for answering the item incorrectly, and the score increased by 1 for answering the item correctly.

The Conventional Linear Tests

The two conventional linear tests administered were the Verbal Test (VE-2B) and the Arithmetic Reasoning Test (AR-4B) of the Army Classification Battery. These tests are power tests of 50 items and 40 items, respectively, each item having four alternatives. Total scores were corrected for chance success.

PROCEDURE

The two branching tests and the two linear tests were administered to all examinees in counterbalanced order, half taking the two branching tests first and half taking the linear tests first. The examinees were 102 enlisted men from Fort Belvoir, Virginia, with a wide range of scores on the General Technical Aptitude Area, a composite of VE and AR. No particular sampling design was attempted. Examinees were told they were taking part in an experiment.

The linear tests were administered to groups of about 25 men. The branching tests were administered to one examinee at a time. The teletypewriter typed out the branching test item, and the examinee responded by pressing a typewriter key appropriate to the alternative selected. So long as the item was on display, the examinee could change his answer by pressing the key for another alternative. A "Record" key entered the last alternative selected as the answer. The computer scored omitted items as wrong; hence, examinees were instructed to guess if necessary. Examinees were also instructed to guess on the linear tests. Examinees were not informed of the nature of the branching tests.

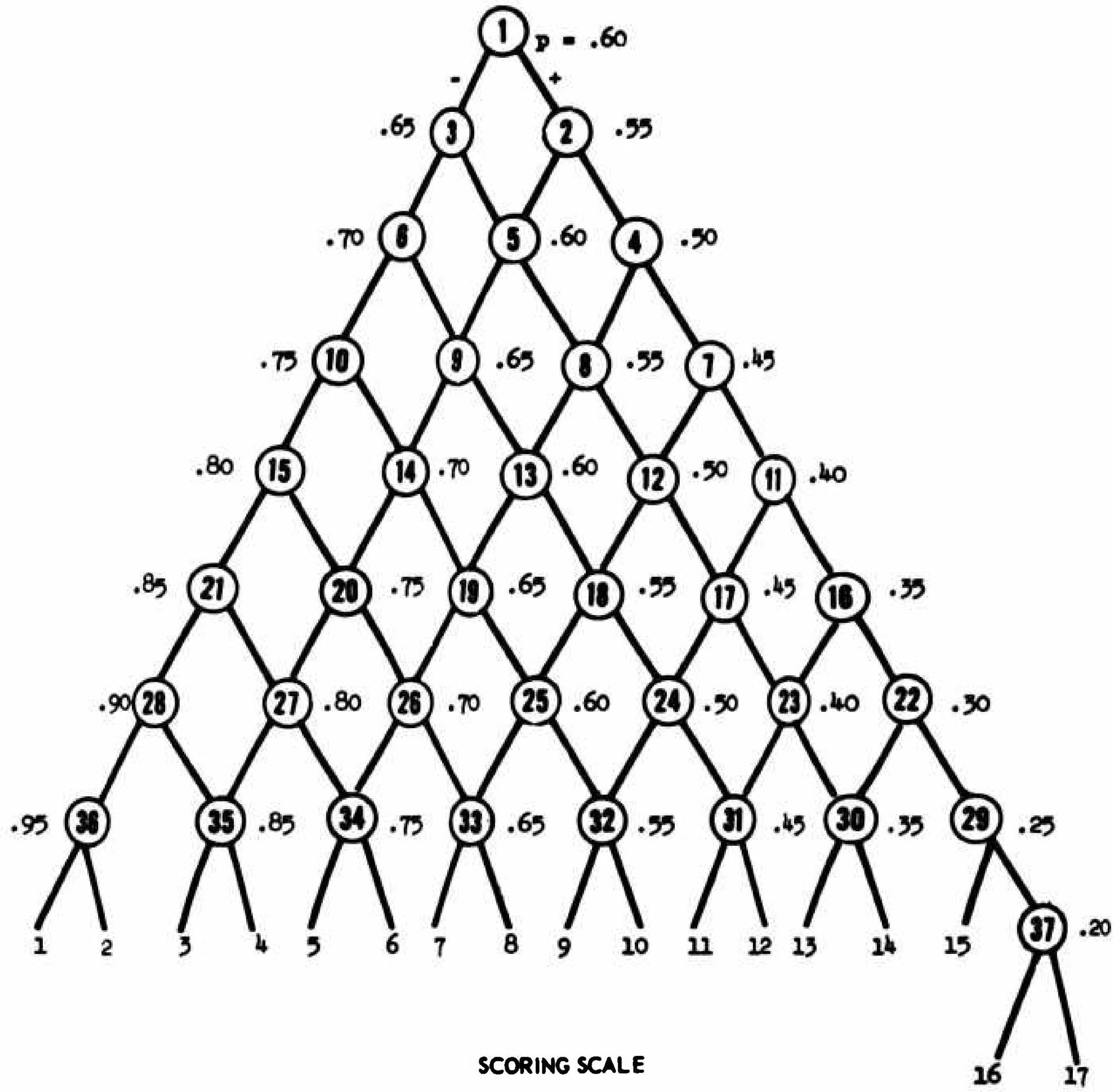


Figure 1. Branching test plan

RESULTS AND DISCUSSION

A correlation coefficient was computed between each of the branching tests and its counterpart linear test. The two orders of testing were combined. Correlation between the two verbal ability tests was r = .78; between the two arithmetic reasoning tests, r = .74. When corrected to the mobilization population, using the GT aptitude area as the selector, correlation was increased to r = .83 for VE and .79 for AR (Table 1).

Table 1

CORRELATION OF BRANCHING TESTS WITH CONVENTIONAL LINEAR TESTS

Tests	No. of Items	M	S.D.	Correlation Coefficient[a]			
				Branching		Linear	
Branching		Raw Scores		VE	AR	VE	AR
VE	8-9	10.6	2.8		.57	.78	.64
AR	8-9	9.9	3.9			.50	.74
Linear		Army Standard Scores					
VE	50	105.9	19.1	.83		.91[b]	.65
AR	40	105.6	17.7		.79		.85[b]

[a] Coefficients above diagonal uncorrected for selection; coefficients below diagonal corrected to mobilization population.

[b] Test-retest reliability in mobilization population.

The test-retest reliability estimates of the ACB tests, as recently determined in carefully constructed samples of enlisted men and corrected to the mobilization population, are r = .91 for VE and r = .35 for AR. These coefficients represent the maximum correlation that could practically be expected between the 8- or 9-item branching tests and the 40- and 50-item linear tests of the ACB.

To provide a frame of reference for these data, two estimates were made: (1) the length a linear test would have to be in order to be as reliable as the 8-item branching test, and (2) the correlation between 8-item linear tests and their counterpart long linear tests.[1]

[1] The following analyses were contributed by Dr. John J. Mellinger, Chief, Statistical Research and Consultation Branch, Statistical Research and Analysis Division, BESRL.

The reliability coefficients of the branching tests were based on the reliability indexes of the two long linear tests (.95 for VE, .92 for AR) obtained from the test-retest reliability coefficients. To obtain the reliability indexes of the 8-item branching tests, the coefficients of correlation between the branching tests and their long linear counterparts (.83 for VE, .79 for AR) were divided by the reliability indexes of the long linear tests. From these indexes (.87 for VE, .86 for AR) the reliability coefficients of the branching tests were computed (.76 for VE, .73 for AR). The Spearman-Brown formula indicated that a linear VE test with a reliability coefficient of .76 would require 16 items; a linear AR test with a reliability coefficient of .73 would require 19 items in contrast to the branching tests of 8-9 items.

Estimates of the correlation between 8-item linear tests and the longer linear tests were derived from their respective reliability indexes. On this basis, it was estimated that the correlation between the 8-item linear tests and the 40- and 50-item tests was $r = .75$ for VE and $r = .67$ for AR. Comparable correlation coefficients for the branching tests were .83 for VE and .79 for AR.

The correlation coefficient (uncorrected for selection) between the branching VE test and the linear AR test ($r = .64$) was the same as the uncorrected correlation coefficient between the linear VE test and the linear AR test ($r = .65$). The comparable coefficient between the branching AR test and the linear VE test was lower ($r = .50$), presumably a function of the marked skew in the branching AR test distribution, as described below.

The distributions of the linear tests and of the branching VE test were approximately normal. However, the distribution of the branching AR test departed markedly from normality, with 12 of the 102 examinees obtaining the maximum score. These 12 examinees had Army standard scores of 107 to 141 on the linear AR test, indicating that a higher ceiling for the branching test might have resulted in higher correlation with the conventional test. The behavior of the AR items in the only sequence possible for those obtaining the maximum score (correct answers for items 1, 2, 4, 7, 11, 16, 22, 29, 37) was examined. These items differed by successive decrements of $p = .05$, approximately. Beginning with the fourth item in this sequence (item 7), all the succeeding items were answered correctly by most of the examinees who attempted them--in spite of the range of difficulty ($p = .45$ to $.20$). For these examinees, the items were apparently equal in difficulty.

The apparently lesser difficulty of the items in the AR branching test raises a question concerning the index of difficulty. The conventional p-value is a population value--the proportion of a population that answers an item correctly. It does not indicate the proportion of a particular level of ability that answers correctly. An item that is considered difficult because of its low p-value is not necessarily equal in difficulty for all levels of ability. Items which differ in p-value for the entire population may, in fact, be equal in the proportion of higher levels of the population which answer correctly; conversely, items which are equal in p-value for the entire population may differ in the proportion of the lower levels of the population which answer correctly. The logic of the branching program does not appear compatible with population indexes of difficulty. After the initial item, the difficulty indexes must be related to ability level, uncontaminated by the contributions of the rest of the population.

Another problem concerns the homogeneity of items. The problem is not unique to branching tests, of course, but is emphasized. The small number of items that are answered and the variation in the particular items that are responded to by different examinees make it more difficult to sample only the common content than in the linear tests with their larger numbers of items, all of which are responded to by all the examinees. Moreover, as with the computation of difficulty indexes, homogeneity indexes such as item-test correlation coefficients need to differentiate the effect of examinees who are presented with particular items in the branching test and those who are not.

Since it is possible to arrive at a terminal item by a variety of pathways or item sequences--except for the easiest and most difficult terminal items--it was of interest to determine if, in fact, such variety did occur. Accordingly, the pathways taken by each examinee were tabulated and grouped according to terminal item. In most instances, as Table 2 indicates, a variety of pathways, differing in the average p-value of the items, were taken to arrive at the same terminal item. This finding has several implications, in addition, of course, to the possibility that these are chance variations: (1) Item p-values as indexes of difficulty are relatively imprecise--not a new finding. (2) The branching program permits the individual to respond according to item difficulty for himself, which is different from the difficulty represented by population p-values. (3) Variation in pathway may be a significant parameter of branching tests, and if incorporated in the scoring may contribute to more effective discrimination.

General Observations

The machine mode of administering the branching tests apparently aroused great interest in the examinees. The occasional difficulties with the equipment were promptly dealt with and did not appear to introduce error into the scores nor adversely affect motivation. All the examinees had taken the same or similar linear tests within the past few months. However, it is not possible to tell how much change in scores occurred, since the original scores were not available.

Table 2

NUMBER OF DIFFERENT PATHWAYS TO EACH TERMINAL ITEM ON BRANCHING TESTS

Score	Terminal Item	f VE	f AR	Pathways Possible	Pathways Taken VE	Pathways Taken AR
1	36	0	0	1	0	0
2	36	0	2	1	0	1
3	35	1	2	7	1	2
4	35	1	3	7	1	2
5	34	0	6	21	0	5
6	34	4	6	21	4	6
7	33	5	12	35	4	12
8	33	18	7	35	13	6
9	32	7	11	35	6	8
10	32	20	13	35	12	10
11	31	5	9	21	4	7
12	31	6	5	21	5	4
13	30	16	4	7	5	3
14	30	10	9	7	5	6
15	29	8	0	1	1	0
16	37	0	1	1	0	1
17	37	1	12	1	1	1
Totals		102	102	256	62	74

Print-outs of responses to the branching tests indicated that practically no examinee had changed an answer, although the computer system permitted changes to be made while the item was displayed. It is not clear whether machine presentation increased the confidence of the examinees or whether they were more interested in the machine operation than in their test scores, especially since they knew they were taking part in an experiment.

The computer system used with the branching program displayed one item at a time. The examinee could see the print-out of the preceding items and his responses but could not change the responses. The printed format, of course, makes many items available at a time and does permit the examinee to vary his order of responding and to change his responses to preceding items. Furthermore, the computer system all but precluded omissions, whereas omissions did occur on the conventional tests. The extent to which these incidental differences affected the correlation is not known.

SUMMARY OF FINDINGS AND CONCLUSIONS

Correlation between the short branching tests and their counterpart long linear tests was substantial ($r = .83$ and $.79$, corrected for restriction in range). Coefficients were considerably higher than would be expected if equally short linear tests were substituted for the branching tests ($r = .75$ and $.67$), and approached the test-retest reliability of the long linear tests ($r = .91$ and $.85$). Linear tests to be as reliable as the branching tests would have to be ~~3.2~~ 2.4 and ~~2.1~~ 2.0 times as long.

Classical test theory from which the linear model is derived does not appear completely helpful in understanding the branching model. The two models, while they appear to treat test items as independent samples of ability, differ in other respects. The linear model requires all examinees to respond to the same set of items; the branching model presents different items to examinees of different ability levels. In the linear model, the items presented are unrelated to the preceding responses; in the branching model, the items presented are determined by the preceding responses. In the linear model, item statistics are based on the performance of the population; in the branching model, performance by ability level must be considered. In the linear model, the method of limiting the score to reflect correct knowledge is a statistical correction for chance success; in the branching model, the method reduces the opportunity for chance success. The linear score is based on the number of items answered correctly; the branching score, on the relative difficulty of the last item. In sum, the branching model resembles the psychophysical concept of the limen to which the classical additive theory seems only partially applicable. If the results of the present study are substantiated, new developments in test theory seem necessary.

The net results of this exploratory study indicated the definite research promise of the branching program for tests. The problems uncovered do not seem insuperable. Research directed toward such problems as effectiveness of branching variants, determination of optimum test length, size of difficulty interval, contribution of other item and score parameters, and generalizability to other content areas should prove profitable and may lead to the eventual development of branching tests for operational use. The immediate need is for equipment designed specifically for experimentation with the branching technique.

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) U.S. Army Behavioral Science Research Laboratory, OCRD	2a. REPORT SECURITY CLASSIFICATION Unclassified
	2b. GROUP

3. REPORT TITLE

An Exploratory Study of Branching Tests

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (Last name, first name, initial)

Bayroff, A. G. and Seeley, Leonard C.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1967	20	4

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Research Note 188
b. PROJECT NO. DA R&D No. 2J024701A722	
c. Input Quality	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. 1-00	

10. AVAILABILITY/LIMITATION NOTICES

Qualified requestors may obtain copies of this report directly from DDC. Available, for sale to the public, from the Clearinghouse for Federal Scientific and Technical Information, Department of Commerce, Springfield, Virginia 22151.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Dep Chief of Staff for Pers., Wash., D.C.

13. ABSTRACT

In the conduct of continuing research to improve screening and induction techniques, the INPUT QUALITY Task has explored the potential contribution of programmed testing particularly with respect to use of the branching technique which tests the examinee with items appropriate to his own ability level. The present publication reports on the trial administration of two branching tests by means of a computerized system with teletypewriter input/output developed by the National Bureau of Standards. A specific objective was to determine the comparability of computerized branching tests and conventional paper-and-pencil tests with reference to reliability, test score information, and test construction rationale. The two specially constructed 8-9-item branching tests (verbal and arithmetic reasoning) and corresponding conventional 40- and 50-item tests of the Army Classification Battery (ACB) were administered, in counter-balanced order, to a sample of 102 enlisted men. Statistical analyses of score results and observations of response patterns are presented. Findings indicated substantial correlation of the short branching tests with their longer conventional counterpart tests (r = .83 and .79, greater than would be expected with equally short conventional tests). Findings also reinforced the research promise of branching tests and indicated the need for reexamination of classical test theory.

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
*Automated testing						
*Branching technique						
*Linear tests						
*Test construction theory						
Statistics						
Armed Forces Qualification Test (AFQT)						
Psychometrics						
Army Classification Battery (ACB)						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
_______________ ."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
_______________ ."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
_______________ ."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.