AFHRL-TR-85-19

(12)

AIR FORCE HUMAN RESOURCES LABORATORY

ARMED SERVICES VOCATIONAL APTITUDE BATTERY: DEVELOPMENT OF AN ADAPTIVE ITEM POOL

J. Stephen Prestwood
C. David Vale

Assessment Systems Corporation
2233 University Avenue, Suite 310
St. Paul, Minnesota 55114

Randy H. Massey
John R. Welsh

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

September 1985

Final Report for Period October 1981 - May 1985

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

DENNIS W. JARVI, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified AD-A160 608	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-85-19
6a. NAME OF PERFORMING ORGANIZATION Assessment Systems Corporation 6b. OFFICE SYMBOL (*If applicable*)	7a. NAME OF MONITORING ORGANIZATION Manpower and Personnel Division
6c. ADDRESS (*City, State, and ZIP Code*) 2233 University Avenue, Suite 310 St. Paul, Minnesota 55114	7b. ADDRESS (*City, State, and ZIP Code*) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (*If applicable*) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F-33615-81-C-0020
8c. ADDRESS (*City, State, and ZIP Code*) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62703F; PROJECT NO. 7719; TASK NO. 18; WORK UNIT ACCESSION NO. 13

11. TITLE (*Include Security Classification*)
Armed Services Vocational Aptitude Battery: Development of an Adaptive Item Pool

12. PERSONAL AUTHOR(S)
Prestwood, J.S., Vale, C.D., Massey, R.H., and Welsh, J.R.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Final	FROM Oct 81 TO May 85	September 1985	90

16. SUPPLEMENTARY NOTATION
TS Analysis Codes 8768, 8760, 8533, 8414, 8381

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
Armed Services Vocational Aptitude Battery
calibrations
computer adaptive testing
domain specifications
item difficulty
(Continued)

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

In order to take advantage of advances in the field of mental measurement, the Armed Forces and the Department of Defense have supported the development of a computerized adaptive version of the Armed Services Vocational Aptitude Battery (ASVAB) for use in military personnel selection and classification. This report describes the development and calibration of item pools for each of nine ASVAB content areas. Domain specifications were developed for the content areas, and more than 3,600 items were written. The items were then pretested on samples of recruits. These data were used to select items for calibration in a sample of over 136,000 examinees tested in Military Entrance Processing Stations and their associated testing sites in May and June of 1983. The calibration data were then analyzed using both an equivalent-groups design and a joint-calibration design that used matched experimental and operational test data. Item response theory a, b, and c parameters based on the three-parameter logistic item response model were computed. The parameters resulting from the joint-calibration approach were recommended for operational use.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy A. Perrigo, Chief, STINFO Office	22b. TELEPHONE (*Include Area Code*) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Item 18 (Concluded):

item discrimination
item response theory
theta (ability) estimates
three-parameter logistic model
selection
classification

SUMMARY

The Armed Services Vocational Aptitude Battery (ASVAB), consisting of 10 subtests, is the primary instrument for assessing abilities of young men and women enlisting in the Armed Forces. To take advantage of advances in the field of mental measurement, the Armed Forces and the Department of Defense have supported development of a computerized adaptive version of the ASVAB. This report describes the procedures used to develop and calibrate item pools for this new test.

Content areas for the computerized ASVAB were the same as those used in the conventional ASVAB, with two exceptions--no items were written for the speeded subtest areas, and the Auto and Shop Information subtest was divided into two separate content areas. Domain specifications for the content areas were developed, and over 3,600 items were written and pretested. Pretesting took place in Recruit Training Centers (RTCs) and employed an equivalent-groups design. The item response data were analyzed using classical and item response theory (IRT) procedures. The pretesting data indicated that, while the item discrimination parameters were satisfactorily high, more easy items would be required to achieve the desired rectangular distribution of item difficulty parameters.

Additional easy items were developed, and approximately 200 items from each content area were selected for calibration in 63 Military Entrance Processing Stations (MEPS) located throughout the nation. Both an equivalent-groups design and a joint-calibration design using matched experimental and operational test data were employed in analyzing the data. The IRT *a*, *b*, and *c* parameters computed using the joint-calibration approach were recommended for operational use.

PREFACE

This technical report and the item pool development effort it describes were completed as part of the Omnibus Item Pool and Test Development Project (Contract F-33615-81-C-0020). This project was completed by Assessment Systems Corporation, St. Paul, Minnesota, for the Air Force Human Resources Laboratory, Brooks AFB, Texas.

Special appreciation is expressed to Dr. Malcolm Ree of the Air Force Human Resources Laboratory and to Dr. Jerome Lehnus of the Military Entrance Processing Command for their contributions to and support of this project.

TABLE OF CONTENTS

LIST OF TABLES

ARMED SERVICES VOCATIONAL APTITUDE BATTERY: DEVELOPMENT OF AN ADAPTIVE ITEM POOL

I. INTRODUCTION

The Armed Forces Qualification Test (AFQT), first introduced in 1950, was a mental test battery developed jointly by the Armed Forces for screening potential recruits. The AFQT was a paper-and-pencil test administered to groups of examinees. It was composed of three power subtests that contained vocabulary, arithmetic reasoning, and spatial-relations items. The AFQT was revised in 1953, 1956, and 1960. In 1972, each service began to administer its own test battery, using that battery to estimate an AFQT score. In 1975, joint development efforts among the services led to the development of a new, common test battery--the Armed Services Vocational Aptitude Battery (ASVAB). Again, the new battery was used to estimate AFQT scores for potential recruits. It was also used to make classification decisions for each of the services.

The ASVAB went through several revisions over the years. Since 1980, the operational ASVABs have consisted of 10 subtests. These subtests are General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto-Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronics Information. The scores from four of the subtests (Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Numerical Operations) are used to compute an AFQT score, which is then used to determine an examinee's eligibility for military service.

Since 1978, consideration has been given to the development of a computerized adaptive version of the ASVAB. Computerized adaptive testing (CAT) has a number of advantages over paper-and-pencil group tests. Among these advantages are improved test efficiency, more uniform measurement precision for examinees at different ability levels, and improved item-pool security. This report describes the procedures used to develop and calibrate an item pool for a computerized adaptive version of the ASVAB. It also describes the results of those efforts and suggests further development needs.

Computerized Adaptive Testing

Computerized adaptive testing, or tailored testing, is one of the newest products of psychometric research. A major development in psychometrics, CAT promises to substantially improve the quality of measurement. CAT takes advantage of modern test theories and recent developments in computer technology to improve both the efficiency and the accuracy of tests.

Computerized adaptive tests provide greater efficiency and equally precise measurement at different levels of ability because the tests

are dynamically tailored to each examinee's level of ability. A variety of testing strategies have been proposed for adaptive testing. These have been described elsewhere (e.g., Weiss, 1974) and will not be detailed here. In general, after each response by an examinee, a provisional ability estimate is calculated. This estimate is used to select from the item pool the next test item that is most appropriate for the examinee. Appropriateness is usually defined mathematically, using the tools of item response theory (IRT) described below. That item is then administered and the process is repeated until some test termination criterion is reached. Because examinees are administered different and non-randomly selected subsets of items, classical test scoring and analysis procedures are inadequate. Thus, in addition to providing a basis for item selection, IRT procedures are used to place test scores on a common metric for examinees taking different subsets of items.

A basic element of an adaptive test is the item pool. Requirements for CAT item pools differ from those used to construct conventional tests. The CAT item pool must contain sufficient numbers of items that are appropriate for each examinee's ability level. As a result, CAT item pools typically contain more items than do conventional tests, and the item difficulties span a wider range of difficulty.

Item Response Theory

IRT specifies a general mathematical relationship between an individual's status on an underlying trait and the characteristics of a test item. IRT actually refers to a general class of psychometric models. Included are models for dichotomous responses (Birnbaum, 1968; Lord & Novick, 1968), polychotomous responses (Bock, 1972; Samejima, 1969, 1972), and continuous responses (Samejima, 1974). These models have been developed for applications in which unidimensional traits are measured. Hambleton and Cook (1977) present an overview of unidimensional IRT models.

The current effort considered only one IRT model: the three-parameter logistic model. In this model, the item is characterized by the three parameters a, b, and c, and ability is characterized by a single parameter, theta. The a parameter is an index of the item's power to discriminate among different levels of ability. Theoretically, it ranges between negative and positive infinity. Practically, it ranges between 0.0 and about 2.5 when ability is expressed in a standard-score metric. A negative a parameter would mean that a low-ability examinee had a better chance of answering the item correctly than did a high-ability examinee. An a parameter of zero would mean that the item had no capacity to discriminate between different levels of ability (and would therefore be useless as an item in a power test). Items with high a parameters provide sharper discrimination among levels of ability and are generally more desirable in CAT item pools than are items with low a parameters.

The b parameter indicates the difficulty level of an item. It is scaled in the same metric as ability and indicates the value of theta at which the examinee has a 50-50 chance of knowing the correct answer to the item. However, this is not the level of theta at which the examinee has a 50-50 chance of selecting the correct answer if it is possible to answer the item correctly by guessing.

The c parameter gives the probability with which a very low-ability examinee would answer the item correctly. It is often called the guessing parameter because it is roughly the probability of answering the item correctly if the examinee does not know the answer. Intuitively, the c parameter of an item should be the reciprocal of the number of alternatives in the item. Empirically, it is usually somewhat lower than this.

All four parameters are used in the three-parameter logistic model to determine the probability of a correct response. The mathematical relationship is given by Equation 1, which shows the probability of a correct response to item g for an examinee with ability theta (θ).

$$P_g(\theta) = c_g + (1 - c_g)\ \Psi[\ 1.7a_g\ (\theta - b_g)] \qquad (1)$$

where

$$\Psi(x) = [\ 1 + \exp(-x)\ ]^{-1}.$$

discriminating power of the test items. In addition, multidimensionality may become a greater problem when the domain specifications for the content areas are broadened to increase the range of item difficulties.

Other differences between the item pools for conventional and CAT tests result from the computerized mode of administration. For example, the size of the computer terminal's screen limits the amount of text that an item may contain, and the resolution of the screen limits the complexity of the illustrations that may accompany the items. In developing a CAT version of the ASVAB, the size of the screen will have an impact on the length of the reading passages in the Paragraph Comprehension subtest. Moreover, computerized presentation may affect content areas differently. For example, the resolution of the screen may have little effect on the illustrations accompanying items in areas such as Shop Information, but may greatly limit the amount of complexity that can be used in the illustrations for Mechanical Comprehension items.

Speeded tests such as the Numerical Operations and Coding Speed subtests on the paper-and-pencil ASVAB present problems for both adaptive testing and computerized test administration. On speeded tests, the probability of correctly responding to an item is close to 1.00, assuming that the examinee reads and responds to the item before the time limit is reached. Adaptive testing is not practical in areas measured by items of essentially equal difficulty, and standard IRT procedures are not appropriate for speeded tests. They are not appropriate because the relationship between ability and a correct response to a speeded test item is not solely a function of the item discrimination, the item difficulty, and the probability of correctly responding to an item by chance. Additionally, the computer hardware used for administering speeded items will affect the scores. Consequently, speeded tests must be administered with the same hardware used to calibrate the items in order for the norming data to be useful. No items for speeded tests were developed in this project.

Procedures

Items for the initial CAT ASVAB were developed in nine content areas. These content areas were the same as those used in the conventional paper-and-pencil ASVAB with two exceptions--no items were developed for the speeded subtests, and the Auto and Shop Information subtest was divided into two separate content areas. The nine content areas were thus General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Automotive Information, Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronics Information. All of the items originally developed had five alternatives in a multiple-choice format. Appendix A contains the guidelines that were used in writing the items within each of the content areas.

All of the items were subject to a four-phase editing process. After each item was written, it was given to a technical editor who corrected grammar, spelling, and typographical errors, and who, if necessary, rewrote the item to improve its clarity. In the second phase of the editing process, the item was returned to its original author. The author then reviewed the corrections made by the technical editor to ensure that no changes had been introduced that would affect the accuracy of the item. In the third stage, the Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Mathematics Knowledge, and Mechanical Comprehension items were reviewed by staff members familiar with the content to ensure that each item was correct as written, that the subject matter was relevant to the skill being tested, and that the difficulty level was appropriate. The General Science, Automotive Information, Shop Information, and Electronics Information items were similarly reviewed by external consultants who taught classes in these areas at a large vocational-technical institute. In the fourth phase of the review, the items were returned to the technical editor who ensured that no typographical errors had been introduced during the previous editing stages.

Domain Specifications

Domain specifications were developed to outline the content of the items to be written in each area. ASVAB 8ax, an experimental version of the operational ASVAB 8a, was provided to the contractor for guidance in explicating the content domains of the nine subtests. In most cases, there were too few items in that form to comprehensively delineate the domain for an area. The content specifications, which are outlined below, were designed to adequately represent the diverse content of each of the knowledge or skill areas incorporated in the ASVAB and to assess the knowledge and skills needed by the Armed Services for selection and placement.

General Science. The General Science domain was specified using textbooks for junior and senior high school science courses (Heimler & Price, 1977; Keeton, 1968). Other science textbooks of different difficulty levels were surveyed to ensure that the domain was specified completely and that the relative representations of the areas within the domain were related to the amount of text devoted to the subjects in the textbooks. The specifications were then reviewed and modified by an instructor at the University of Minnesota who taught natural science, physical science, and biology and was responsible for developing and maintaining an item pool for the university's general biology course.

The General Science domain was divided into three main content areas: life science, physical science, and earth science. The representation of each of these areas in the domain was approximately 40%, 40%, and 20%, respectively.

Life science items dealt with the animal and plant kingdoms and with ecology and the environment. Included were questions concerning cell structures and functions, human nutrition, health, genetics, and the classification of animal systems and groups. Also included were questions about plant structures and photosynthesis.

Sixty of the physical science items (15% of the entire domain) were devoted to chemistry and the classification of matter. Other items dealt with the concepts of force, work, energy, and simple machines. The remaining 100 items of the physical science domain included items about heat, light, sound, electricity, and magnetism.

The earth science items dealt with astronomy (specifically Earth and the solar system), weather, and the atmosphere. Also included were items dealing with the formation and classification of rocks and soil.

Table 1 lists the areas of the domain and shows the numbers of items written and pretested in each area. The area called "Important Names in Science" consists of items dealing with the names of important figures in the history of science, whether they are associated with life, physical, or earth science.

Arithmetic Reasoning. The Arithmetic Reasoning domain consisted of items requiring the recognition and application of basic mathematical concepts and operations in problems encountered in everyday life. The items were designed to emphasize the concepts or operations required for solution rather than computational complexity. Six basic concept/operation areas were included. The items required skills represented by one or more of these areas.

The first area involved the recognition and application of the four basic arithmetic operations: addition, subtraction, multiplication, and division. The algebraic forms and illustrative examples are shown below (a, b, and c represent integers, decimals, or fractions, while x represents the unknown value).

a+b=x	If a 10-foot and 15-foot extension cord are connected together, how far will they reach?
a-b=x	If 5 feet are cut from a 10-foot board, how many feet will be left?
ab=x	If four 6-foot hoses are connected, how many feet will they reach?
a/b=x	If 12 apples are split evenly among 4 children, how many will each get?

The second area involved a rearrangement of the basic operations and thus required some algebraic manipulation to find the answer. The algebraic forms and several examples are shown below.

a+x=b	If you connect a 15-foot extension cord to another cord and find that both will reach 25 feet, how long is the other cord?
a-x=b	If you cut 5 feet from a board and find that 5 feet are left, how long was the original board?
ax=b	If four hoses of equal length connected together reach 24 feet, how long is each hose?
a/x=b	If 12 apples are split evenly among a group of children and each child gets 3 apples, how many children are in the group?

The third area assessed skill in dealing with percentages. Three basic forms for these items and examples are shown below.

a% of b=x	If 20% of Bill's $150.00 check is taken away for taxes, how much tax does he pay?
x% of b=a	If $30.00 of Bill's $150.00 check goes to taxes, what percent of his check is this?
a% of x=b	If Bill pays 20% of his check to taxes and he pays $30.00 in taxes, how much was his original check?

The fourth area assessed skill in solving rate problems and other problems involving equivalent-fractions operations. One of these problems, for instance, might have asked, "If 3 workers can produce 6 widgets in 4 hours, how many widgets can they produce in 8 hours?" The fifth area assessed skill in converting simple units of weight, time, and distance. The sixth area required the determination of perimeters, areas, and volumes of circles, squares, rectangles, triangles, and cubes.

Table 2 shows the number of items that were written and pretested in each of the areas. Items employing two areas or three or more areas are tallied separately.

Word Knowledge. The Word Knowledge subtest of the ASVAB 8ax contained two types of item stems. Most of the item stems were of the form "_____ most nearly means...." Approximately 37% of the stems, however, were complete sentences containing the key word in context. Only the first form shown above was used for the CAT pool in order to

ensure that the resulting subtest would be as unidimensional as possible. The key words for the item stems were selected from the Thorndike and Lorge (1944) word frequency lists by frequency category in an attempt to achieve the difficulty levels appropriate for the population to be tested. The frequency categories used were determined by analyzing the relationship between frequency and item difficulty in the ASVAB 8ax data and in pretest data collected as part of the ASVAB 11, 12, and 13 development effort. The key words were also chosen so that they would not duplicate any used in ASVAB 8ax or in the pre-operational CAT item pool developed by the Navy.

Paragraph Comprehension. The Paragraph Comprehension items were designed to assess an examinee's ability to understand what he or she reads. Six facets of the comprehension domain were measured by the items: (a) the ability to recall literal detail, (b) the ability to paraphrase or summarize a passage, (c) the ability to recognize main ideas, (d) the ability to make inferences regarding material in the passage, (e) the ability to apply the material in the passage to other material, and (f) the ability to recognize and understand sequential, cause/effect, and comparative relationships. Some items tapped only one of these abilities, although most assessed more than one.

The paragraphs varied in length from about 50 to 130 words, excluding the item stem and alternatives. For each paragraph, a single question was provided. This was done to meet the functional independance required for adaptive testing and the local independence required by IRT. Longer passages with multiple paragraphs like those found in the conventional ASVAB Paragraph Comprehension subtests were not written for this project because of the limitations imposed by the size of CRT screens.

Factual paragraphs, fictional paragraphs, and paragraphs stating opinions were written for each facet of the domain. The specific content of the paragraphs was selected to minimize the effects of examinees' prior experiences on performance and thus to require the examinees to read and understand the information presented in the paragraph in order to choose the correct answer.

Table 3 shows the number and percentage of items measuring each facet of the domain in pretesting. Approximately 77% of the paragraphs were factual, 15% were fictional, and 8% stated opinions.

Automotive Information. The Automotive Information domain was developed with the aid of three basic texts in automobile mechanics used by local vocational-technical schools (Ellinger, 1977; Stockel, 1978; Toboldt & Johnson, 1981). A preliminary domain was established by examining the tables of contents of these texts and determining target representations of the areas as percentages of pages devoted to each in the texts. The items were written and then subjected to a

review by three content experts who were automotive repair instructors at a vocational-technical school. The reviewers identified items that were ambiguous, mis-keyed, or obsolete. They also rated the expected difficulty of the items. As a result of their review, some items were replaced or edited. The new and edited items were then subjected to the same review process. The areas within the domain and their representations in pretesting are shown in Table 4.

Shop Information. The Shop Information domain was developed by surveying texts used in shop instruction in high schools and in vocational-technical schools (Feirer, 1977; Jackson & Day, 1978; Walker, 1973). Popular home maintenance and do-it-yourself manuals (Better Homes and Gardens, 1980; Reader's Digest Association, 1973) were also reviewed. The domain was divided into three main categories--tools, materials, and miscellaneous. The item difficulties ranged from very easy (identification of household tools such as saws and hammers) to very difficult (identification of symbols used to specify types of welds). All Shop Information items were reviewed by two shop instructors from a vocational-technical school. The review identified items that were ambiguous, too difficult, or poorly written. Approximately 20 of the 400 items originally written in this area were replaced as a result of the review. Another 20 items were modified. Table 5 outlines the Shop Information domain and shows the number of items written and pretested in each area of the domain.

Mathematics Knowledge. The specification of the Mathematics Knowledge domain was based on recommendations from the General College mathematics coordinator at the University of Minnesota. The first area of the domain covered the conversion of fractions, decimals, percentages, and mixed numbers to other forms (e.g., fractions to decimals, mixed numbers to improper fractions, etc.). Other items required the examinee to compare the sizes of different fractions, to obtain reciprocals, and to reduce fractions to lowest terms. This area also included the computation of least common denominators, greatest common factors, and smallest common multiples.

Approximately 15% of the Mathematics Knowledge domain covered a variety of arithmetic and algebra topics including prime numbers, factorials, absolute values, and logarithms. Also included were items requiring knowledge of the correct order of operations, rounding and place values, and the rectangular coordinate system. Ten items required an examinee to transform a verbal statement into symbolic (algebraic) form.

Geometry items made up approximately 18% of the Mathematics Knowledge domain. Half of these items required knowledge of analytic geometry, and the other half required knowledge of plane and solid geometry. Most of the analytic geometry items dealt with linear equations; the other analytic geometry items dealt with the equations

for a sphere and the conic sections (circle, parabola, hyperbola, ellipse). The plane geometry items required an examinee to compute the perimeters, areas, and volumes of circles, triangles, rectangles, trapezoids, cubes, spheres, and cylinders.

Seventy items were written in the next content area. These items required knowledge of square roots and cube roots and how to do computations with variables raised to different powers. This area also covered simple operations with polynomials. Knowledge of the quadratic equation for finding roots of polynomials was required in several items.

The remaining items involved solving equations and inequalities. These equations ranged from very simple forms (e.g., $x + 3 = 4$) to more complex forms that required several operations before the solution could be obtained. The areas and the number of items written in each area are shown in Table 6.

Mechanical Comprehension. Mechanical Comprehension items were written to assess the ability of the examinee to apply mechanical principles to simple devices in order to determine some aspect of their operation. The Mechanical Comprehension area included simple devices such as gears, pulleys, wheels, and levers. Items were written which involved one or more actions of these simple devices. Items were also written involving complex machines which employed two or more different simple devices. Other items required analyses of static systems under stress or assessed knowledge of basic topics in physics such as gravity, inertia, magnetism, centrifugal forces, and diffraction. The remaining items involved hydraulics or pneumatic systems. Table 7 outlines the content areas and shows the number of items written and pretested for each.

Electronics Information. The Electronics Information domain specifications were initially developed using three texts in elementary electronics used by local vocational-technical schools (Gerrish & Dugger, 1980; Grob, 1977; Matt, 1980). A preliminary domain was established by examining the contents of these texts and establishing target representations of the areas using the proportions of pages devoted to each in the texts. Several sample items were written to tap knowledge in each of the areas established in this domain. A review of the items developed in this manner, however, suggested that these items would be much too difficult for the ASVAB item pool.

The second attempt at domain specification was accomplished by surveying the items in the three unique Electronics Information subtests in ASVAB 8, 9, and 10. This survey resulted in the identification of nine areas. The first area, Devices, included items that tapped very basic knowledge and understanding of certain electrical devices. Examples of such devices are batteries, wire,

motors, and transformers. Term Recognition items consisted of items listing one technical term and four additional terms drawn from the areas of automotive, shop, mathematics, and general science; the general stem for these items was "Which of the following is an electrical term?" The third area, Term Definition, either provided a term in the stem and asked for a definition, or provided a definition in the stem and asked for the term. Terms included basic units of electricity such as units of conductance, capacitance, or resistance, and also some slightly more advanced terminology such as types of transformers and types of wire. The fourth area, Advanced Electronics, included most of the areas that had originally been developed from the three books on electronics. These items assessed knowledge of more advanced electronic concepts such as AC and DC circuits in theory and practice, and basic design of power supplies, amplifiers, oscillators, transmitters, and receivers. The fifth area, Physics, assessed basic knowledge in areas such as magnetism, electrostatics, and electrodynamics. The sixth area, Important Names, assessed the examinee's ability to recognize the name of a famous person associated with electricity or electronics and to identify what that person was famous for. The seventh area, Instruments, assessed the examinee's knowledge of the purpose of different electronic instruments and how to use them. The eighth area, Household Electrics, assessed the examinee's knowledge of basic wiring of household appliances such as stoves, refrigerators, toasters, etc. Finally, the ninth area, Schematic Diagrams, assessed the examinee's ability to read, understand, and trace signals in a schematic circuit.

The areas and the number of items pretested in each are shown in Table 8. The items were reviewed by two instructors at a local vocational-technical school. The reviewers were asked to indicate the correct alternative for each item and to rate the difficulty of each item on a five-point scale. A set of 10 benchmark items was used to define the rating levels. Ten items were deleted as a result of the reviewers' suggestions, and new items were written to replace them.

III. ITEM POOL PRETESTING

Method

The design of this item development effort called for administering all items in Recruit Training Centers (RTCs) during May and June of 1982. The item statistics and IRT item parameters would then be used to select items for calibration in Military Entrance Processing Stations (MEPS). Pretesting and calibration each included four steps--booklet construction, data collection, data editing, and data analysis.

Booklet Construction. A total of 3,654 items were assembled into 71 test booklets. Each booklet contained items from a single content area. The items within a content area were randomly assigned to the booklets representing that area and were randomly ordered within booklets. If random assignment and ordering resulted in two similar items being presented on the same page of a booklet, one of the items was moved to another location. This was particularly important in the Mechanical Comprehension area where the items and their accompanying illustrations often required very similar analytical skills. Examinees were given 50 minutes to complete all of the items in a booklet. The number of booklets representing each content area varied as a function of the estimated number of items that could be answered within that time frame. Table 9 shows the number of booklets within each content area, the number of items per booklet, and the total number of items within each content area.

Because the items were eventually to be administered in an adaptive setting where each examinee would move through the various subtests at a different pace, the instructional procedures varied from those usually associated with the ASVAB subtests. When the ASVAB is administered, the examiner reads both general and content-specific instructions to the examinees. For the pretest data collection, general instructions were included in each test booklet and were read by the examiner. The content-specific instructions were read silently by the examinees prior to testing. In the self-paced CAT test, examinees will read the content-specific instructions on their own. In addition to making pretest conditions more like CAT conditions, this variation from the ASVAB procedure allowed any combination of test booklets to be administered simultaneously.

The test booklets were assigned six-digit form numbers. The last two digits of the form numbers were 01 through 71. The middle two digits were equal to the last two digits plus seven. The first two digits were equal to the last two digits plus 14. Thus, the first form number was 150801 and the last was 857871. The redundant coding was used so that the correct booklet number could be recovered even if one of the digits was encoded incorrectly, if two digits were transposed,

or if the entire booklet number was shifted to the right or left when encoded in the 20-character grid on the answer sheet. Booklet numbers were assigned to the different test forms in a quasi-random fashion so that booklets with items from the same content area were somewhat evenly spaced throughout the series of 71 forms.

Data Collection. Eleven RTCs participated in the study. Each of the RTCs was assigned a quota for the number of examinees to be tested. The quota for the RTCs within a single service was divided evenly among the participating RTCs. The quotas for the services corresponded roughly to the proportion of enlistees accepted into that service. The participating RTCs and their assigned examinee quotas are shown in Table 10.

At each RTC, test proctors were instructed to stack the booklets in order of their booklet numbers and to distribute them in a sequential fashion within each testing session. After the session was finished, the booklets were to be collected and returned to the bottom of the stack. This distribution plan ensured that the booklets would be distributed in an approximately random order to the examinees and that each test booklet would be administered an equal number of times. Testing took place during May and June of 1982. Examinees recorded their responses on optically scannable answer sheets. The answer sheets were then returned to the Air Force Human Resources Laboratory (AFHRL) for scanning. Because the booklets were distributed in a quasi-random manner to the examinees and were administered simultaneously, an equivalent-groups design was appropriate for the data analyses.

Data Editing. The answer sheets were scanned by AFHRL and the data were sent to Assessment Systems for editing and analysis. Several data editing procedures were employed. The form numbers encoded on the answer sheets were checked for errors, and the data for examinees who responded to fewer than five items on the test were excluded from further analysis. An algorithm for detecting response strings and response patterning (Prestwood, Vale, Massey, & Welsh, 1985) was then used to examine the response records of examinees with proportion-correct scores near chance level.

Data Analysis. The item response data were analyzed using both classical and IRT procedures. For each item, the proportion correct, biserial item-total correlation coefficient, and point-biserial item-total correlation coefficient were computed. These statistics were also computed for each of the alternatives as if they had been scored correctly. The statistics for the alternatives were used to detect items that were incorrectly keyed. Items with more than one correct answer were deleted, items that were mis-keyed were corrected, and the analyses were repeated.

An equivalent-groups design was used for the IRT data analyses. This design was appropriate because (a) the assignment of items to booklets was essentially random, (b) all booklets were administered in each RTC, (c) the distribution of booklets within a testing session ensured that each subject had an equal chance of taking each booklet, and (d) the booklets used within a testing session were simultaneously administered.

IRT parameters were computed using Version 1.0 of the program ASCAL. ASCAL is a joint maximum-likelihood/modal-Bayesian item calibration program for the three-parameter logistic item response model (cf., Prestwood, Vale, Massey, & Welsh, 1985). The parameter estimates were then transformed from the RTC-based ability metric to a MEPS-based metric using data in a draft AFHRL report entitled "ASVAB Form 8b and AFQT-7A Summary Distributional Statistics for MEPS, Air Force Qualified, and Army Qualified Samples," which was provided for this purpose. The paper contained mean number-correct scores for the traditional ASVAB subtests administered to MEPS (then referred to as AFEES) examinees and to Air Force, Army, and combined Air Force and Army samples of recruits. The Air Force and Army combined data were used for estimating the restricted ability distribution mean and standard deviation for the RTC sample. The MEPS data were used for estimating the unrestricted ability distribution parameters. These data are shown in Table 11. The data for the Auto-Shop subtest on the conventional ASVAB were used for both the Auto Information and Shop Information experimental subtests. The transformations used for the $\underline{a}$ and $\underline{b}$ parameters are shown below in Equations 2 and 3, respectively. The $\underline{c}$ parameters did not change. The development of Equations 2 and 3 is described in Appendix B.

$$\underline{a}_{MEPS} = \underline{a}_{RTCs} \left(\sigma_{MEPS} / \sigma_{RTCs} \right) \quad (2)$$

$$\underline{b}_{MEPS} = \left(\mu_{RTCs} - \mu_{MEPS} + \underline{b}_{RTCs} * \sigma_{RTCs} \right) / \sigma_{MEPS} \quad (3)$$

where

μ_{MEPS} = mean subtest score for MEPS samples,

μ_{RTCs} = mean subtest score for combined samples,

σ_{MEPS} = standard deviation of scores for MEPS samples, and

σ_{RTCs} = standard deviation of scores for combined samples.

Results and Discussion

Data Collection. The response records of 21,093 examinees were collected in the course of pretesting. Table 12 shows the number of examinees tested in each of the RTCs and the percentage of each RTC's

quota which that number represents. The Air Force, Army, and Navy tested slightly more examinees than required (107.6%, 101.5%, 101.2%, respectively). The Marine Corps tested 637 additional examinees (126.5% of quota). Each of the experimental tests was administered to approximately 300 examinees.

Data Editing. Of the 41,093 examinees tested, approximately 99% had correctly encoded or recoverable form numbers. Table 13 shows the number of examinees with correctly encoded or recoverable form numbers for each of the 71 experimental test booklets. Also shown are the numbers of examinees deleted from the analyses during the data editing process (less than 0.6% overall) and the number of examinees with usable data for each booklet. An average of 292 usable response records per booklet were available for analysis after editing. More data were available for the booklets which fell earlier in the sequence, suggesting that the test administrators did not always follow the booklet distribution instructions. However, there is no evidence to suggest that the differences in numbers of examinees per booklet should affect the analyses.

Data Analysis. Table 14 shows descriptive statistics for three conventional item statistics: proportion correct, biserial item-total correlation, and point-biserial item-total correlation. The last row in this table shows the numbers of items included in the analyses. The numbers of items do not match those shown in Table 9 because some items were deleted from the analyses. The items deleted were those for which IRT parameters could not be estimated and those for which some ambiguity in the item caused the statistics for the response alternatives to differ from the expected values.

Mean proportions correct ranged from 0.392 for Electronics Information to 0.626 for Word Knowledge. The minimum proportion correct for the nine areas ranged from 0.012 for Electronics Information to 0.125 for Paragraph Comprehension. Maximum proportions correct ranged from 0.916 for Mathematics Knowledge to 0.984 for Arithmetic Reasoning. These statistics suggested that, with the possible exception of Paragraph Comprehension, all content areas contained sufficient numbers of difficult items; but some content areas had too few easy items.

Mean biserial correlations for the nine content areas ranged from 0.362 in Electronics Information to 0.613 in Word Knowledge. The point-biserial item-total correlation coefficients showed, with a few minor transpositions, the same general pattern as the biserial correlation coefficients. This was expected because the two coefficients are closely related.

Table 15 shows descriptive statistics for the IRT parameters computed in pretesting. Mean a parameters ranged from 1.040 for

Mechanical Comprehension to 1.273 for Mathematics Knowledge. Minimum a parameters were uniformly 0.400 across all the content areas. This minimum value was a lower bound for the item calibration program used. Maximum a parameters ranged from 2.030 for General Science up to 2.400 for Automotive Information. The pattern of mean a parameters across the nine content areas did not closely match the pattern observed in the biserial correlation coefficients.

Mean b parameters ranged from a low of -0.404 for Word Knowledge to a high of 1.328 for Electronics Information. With the exception of Mathematics Knowledge, which had a minimum of -2.718, all content areas uniformly had minima of -3.000 and maxima of 3.000. As was the case with the a parameter minima, these values were bounds imposed by the calibration program. The data show that, with the exception of Paragraph Comprehension and Word Knowledge (which respectively had mean b parameters of -0.131 and -0.404), all content areas appeared to have items that were more difficult than desired for the RTC population. In the case of Electronics Information, the items appeared to be considerably more difficult than desired.

Mean c parameters for items included in pretesting ranged from a low of 0.182 for Mathematics Knowledge to a high of 0.199 for Paragraph Comprehension. Minima were uniformly 0.100 and maxima were uniformly 0.300. Program bounds were set to these values for this calibration. In general, mean c parameters appeared very near 0.200, the expected value for the five-alternative multiple-choice items.

Table 16 shows descriptive statistics for IRT item parameters transformed to estimated values in a MEPS population using the procedures described in the Method section. The MEPS population was expected to be of lower and more variable ability than the RTC population used in pretesting. As a result of the transformations, the a parameters were higher, and the b parameters were more positive. Mean transformed a parameters ranged from 1.188 for Mechanical Comprehension to 1.825 for Word Knowledge. Mean transformed b parameters ranged from 0.358 for Word Knowledge to 1.597 for Electronics Information. The minimum and maximum a and b parameters shown in Table 16 are simple transformations of corresponding values shown in the previous table. Since no transformations were applied to the c parameter, the c parameter descriptive statistics shown in Table 16 are identical to those shown in Table 15.

In general, the results shown in Table 16 suggest that the items that were pretested were more difficult than those that would ultimately be required for the CAT ASVAB. They also suggested, however, that the a parameters would be sufficiently high for adaptive testing to work very efficiently.

Table 17 provides a distribution of transformed difficulty parameters. This table, like the preceding tables, suggests that the

item pool for most areas was deficient at the easy end of the difficulty range, but that all areas, with the possible exceptions of Word Knowledge and Paragraph Comprehension, had more than enough items at the difficult end of the range.

Thus, the general results of the pretesting of the CAT items suggest that additional easy items would be required to provide the desired rectangular distribution of difficulty parameters but that the discrimination parameters of the items written thus far were adequately high.

IV. ITEM SELECTION

The item difficulty range between $\underline{b} = -2.2$ and $\underline{b} = 2.2$ was divided into 20 categories of equal width. The bounds for the extreme categories were then expanded to ± 2.5 in order to include additional items with difficulty parameters that could be expected to regress toward the mean in calibration. The items within each content area were assigned to their appropriate categories on the basis of their estimated MEPS-metric difficulty parameters. Table 17 summarizes the distribution of items by category. Each of the five difficulty categories in the table includes four of those used in item selection. Table 17 shows that the item difficulty distribution was not rectangular over the entire range--there were fewer items at the easy end of the range in all content areas.

A decision was made to select items from those pretested with as rectangular a distribution of item difficulties as possible and then to produce new items that could be expected to be easier than any in the pretesting pool. Approximately 200 items were selected for calibration from each of the content areas.

The items initially selected had $\underline{a} \geq 0.65$, and $-2.5 \leq \underline{b} \leq 2.5$ on the MEPS metric. The $\underline{c}$ parameter was bounded at 0.30 for the pretesting analyses and thus was not used as a selection criterion. Some items were included even though they did not meet these specifications if they appeared to have been too easy for the parameters to have been appropriately estimated from the small (and presumably relatively high-ability) RTC sample. In selecting the specific items to be calibrated, an attempt was made to draw an equal number of items from each of the difficulty categories. When too few items were available in a particular category, the items were chosen from other categories. For instance, if a total of 13 items were available in the first three categories, then those items were selected, and the remaining 187 items were selected by choosing the 11 most discriminating items from each of the remaining categories.

The items selected were sent to AFHRL for review. In some cases, AFHRL substituted items of approximately equal statistical characteristics for the items which had been initially selected. The changes resulted primarily from concerns regarding item content. Table 18 shows distributions of the item difficulty parameter estimates for the pretested items that were selected for calibration in the MEPS.

In addition, a number of new items were written for MEPS calibration in each of the content areas. These new items were designed to be easier than any of the items in the pretesting pool. For each content area, Table 19 shows the number of pretested items in each content area which were selected for calibration, the number of

new items which were written and selected for calibration, and the total number of items selected.

Tables 20 through 27 show the distribution of items selected for calibration by area within item domains for General Science, Arithmetic Reasoning, Paragraph Comprehension, Automotive Information, Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronics Information, respectively. Some of the new areas included in calibration were not included in pretesting; an example is the Device Recognition area of the Electronics Information domain (Table 27). This area contained pictorial representations of common household devices which the examinee was required to identify. Examples of the devices were an electric iron, a light bulb, and an extension cord. Other new items fell into areas such as Term Recognition, where the new items required the examinee to identify which of five simple household objects (e.g., radio, basketball, chair) uses electricity. The new items each had five response alternatives except those written in Mechanical Comprehension. The new Mechanical Comprehension items had three alternatives to allow simpler mechanical processes and singular mechanical outcomes (e.g., speed or direction) to be used in the items.

V. ITEM POOL CALIBRATION

Method

Booklet Construction. The items selected for calibration were assembled into 43 test booklets. As in pretesting, each booklet contained items from a single content area. The items within a content area were randomly assigned to the booklets for that area and were randomly ordered within the booklets. If random assignment and ordering resulted in two similar items being presented on the same page of a booklet, one of the items was moved. The number of booklets for each content area varied as a function of the estimated number of items that examinees could answer within 50 minutes. Table 28 shows the number of booklets within each content area and the number of items per booklet. Each booklet contained general instructions which were read by the examiner and instructions specific to the individual content area which were read silently by the examinees.

The test booklets were once again assigned six-digit form numbers. The last two digits of the form numbers were 01 through 43. The middle two digits were equal to the last two digits plus 11. The first two digits were equal to the last two digits plus 18. Thus, the first form number was 191201 and the last was 615443. As in pretesting, the redundant coding was used to allow recovery of the correct form number when one of the digits was encoded incorrectly, two digits were transposed, or the entire number was shifted to the left or right in the response grid on the answer sheet. The booklet numbers were assigned to the different test forms in a quasi-random fashion. The procedure used ensured that the booklets with items from a single content area were somewhat evenly spaced throughout the series of 43 forms.

Data Collection. The Military Entrance Processing Command (MEPCOM) coordinated the administration of the experimental forms in 63 MEPS and their associated Mobile Examining Team (MET) and Office of Personnel Management (OPM) testing sites. MEPCOM assigned each MEPS and its associated sites an examinee quota which reflected the anticipated number of examinees to be processed through the MEPS during May and June of 1983. MEPS commanders were responsible for distributing the appropriate experimental forms to their associated MET and OPM sites according to instructions provided by MEPCOM. Test administrators in the MEPS, MET sites, and OPM sites were instructed to use the forms on a rotating basis such that each form available at a site would be used once before any form was used twice. Examinees recorded their responses on optically scannable answer sheets. Each examinee was given one experimental form and an operational ASVAB. The experimental form was always administered first. The operational forms in use during May and June of 1983 were 9a, 9b, 10a, 10b, 10x, and 10y.

Data Editing. The answer sheets for the experimental tests were sent to AFHRL. The answer sheets were then scanned, and the data were sent to Assessment Systems for editing and analysis. Several data editing procedures were employed with the experimental forms. The redundant form numbers were checked for accuracy. Inaccurate form numbers were recovered where possible. Examinees responding to fewer than five items were deleted from the analyses. The response records of examinees with proportion-correct scores of 0.35 or less were rescored using the keys for the other forms to ensure that the correct form number had been coded. The algorithm used in pretesting for detecting response strings and response patterning was also employed.

The answer sheets for the operational tests were also sent to AFHRL. After they were scanned, the data were sent to Assessment Systems for analysis. The operational response records were matched, where possible, to experimental response records using the examinees' recorded social security numbers. Only exact matches were considered valid. Operational response records with incorrectly encoded form numbers were excluded from the joint calibrations of experimental and operational items.

Data Analysis. The item response data from the experimental tests were analyzed using both classical and IRT procedures. For each item, the proportion correct, biserial item-total correlation coefficient, and point-biserial item-total correlation coefficient were computed. The proportion of examinees endorsing each response alternative and the biserial and point-biserial correlation coefficients for each alternative were also computed. As in pretesting, these alternative statistics were used to verify the keys assigned to the items.

Two basic data analysis designs were used for estimating IRT item parameters--an equivalent-groups design and a simultaneous-calibration design. The equivalent-groups design was considered appropriate because (a) the assignment of items to booklets was essentially random, (b) all booklets were administered in each MEPS, (c) the distribution of booklets within a testing session ensured that each examinee had an equal chance of taking each booklet, and (d) the booklets used within each testing session were simultaneously administered. In the equivalent-groups design, the IRT a, b, and c parameters were computed for the items in each of the booklets separately.

If the groups of examinees taking each of the experimental tests within a content area were equivalent in ability, the parameters estimated for each content area using the equivalent-groups design should be on a common metric. To test the equivalent-groups assumption within each content area, analyses of variance were used to contrast the scores on the like-named operational ASVAB subtests for examinees taking different experimental forms. Arc-sine transformations of the proportion-correct scores were used in the analyses (Winer, 1971). For

the Automotive Information and Shop Information content areas, only the items which were auto-specific or shop-specific from the operational Auto-Shop subtests were used for computing the ability distributions. If the groups taking different experimental forms within a content area were equivalent in terms of their ability distributions on the operational subtests, then the assumption of equivalent groups in the initial IRT analyses would be demonstrated. If the groups taking different experimental tests within a content area had different ability distributions on the like-named operational subtests, then the equivalent-groups IRT parameters could be adjusted to take these differences into account.

In the second data analysis design, IRT item parameters were simultaneously estimated within each content area for all of the experimental items in that area and all of the items from like-named operational subtests. In this simultaneous or "joint" calibration of the experimental and operational items, inclusion of the operational ASVAB items ensured that the IRT parameters for the various experimental forms and for the operational subtests would be estimated on a single metric.

In both designs, chi-square fit statistics were computed for all of the items. These statistics, like the conventional response-alternative statistics, were used to detect possible problems in the item keys. Items with high chi-squares were individually inspected, and no mis-keyed items were found.

Results and Discussion

Data Editing. A total of 138,424 examinees were tested. More examinees were tested per booklet in the MEPS in order to increase the stability and accuracy of the IRT parameter estimates. Of the total number tested, 136,327 had properly coded form numbers. The data for an additional 1,292 examinees were recovered through analysis of redundantly coded form numbers. Thus 137,619 response records had identifiable form numbers. The response data for 241 examinees were then removed during subsequent editing. Of these, 48 responded to too few items, 92 scored near chance on the form ostensibly administered and much higher on another form, and 101 were eliminated because of response patterning. This left a total of 137,378 response records for the classical and equivalent-groups IRT analyses. This figure was just over 99% of the total experimental tests administered. Table 29 shows the results of editing for each of the 43 experimental test forms.

The sampling plan implemented should have resulted in an approximately equal number of examinees for each test booklet. Table 29 shows, as Table 13 did for the pretest data, that more low-numbered than high-numbered booklets were administered. The administrators apparently did not distribute all of the booklets in rotation as

instructed. Although greater numbers of examinees took the low-numbered booklets, there is no evidence to suggest that the administration process introduced any systematic bias related to ability level into the assignment of booklets.

The experimental response records were then matched to the operational response records using the recorded social security numbers. As Table 30 shows, 84.6% of the examinees with valid experimental data could be matched to operational ASVAB data. This percentage was relatively constant across content areas, ranging from 84.2% for Shop Information to 85.4% for Automotive Information.

The conventional and equivalent-groups IRT analyses used all of the experimental data remaining after editing. The analyses of ASVAB score distributions and the joint calibration of experimental and operational items used only the matched data.

Data Analysis. Table 31 shows descriptive statistics for the conventional item parameters computed on the total calibration sample used in the equivalent-groups analyses. Mean proportions correct ranged from 0.573 for Mechanical Comprehension to 0.717 for Word Knowledge. Comparing these results to those shown in Table 14, it can be seen that the items administered in calibration had substantially higher proportions correct than did those administered in pretesting. This was probably because of differences in the sets of items administered in pretesting and calibration. Minimum proportions correct ranged from a low of 0.072 in Electronics Information to a high of 0.119 in General Science. Maximum proportions correct ranged from 0.975 in Mathematics Knowledge to 0.995 in both Word Knowledge and Automotive Information.

Mean biserial correlations ranged from a low of 0.494 in Electronics Information to a high of 0.661 in Word Knowledge. In general, the mean biserials in the calibration sample were higher than corresponding values in the pretesting sample. Minimum biserial correlations ranged from a low of -0.051 for Arithmetic Reasoning to a high of 0.250 for Mathematics Knowledge. Maximum biserial correlations ranged from 0.771 for Shop Information to 1.000 for Word Knowledge.

Table 32 shows these same statistics computed on a sample containing only male examinees. In general, the same patterns and levels observed in Table 31 are again observed in Table 32. The only notable differences are a rather slight increase in mean proportion correct for the males-only sample in Automotive Information, Shop Information, Mechanical Comprehension, and Electronics Information. These differences range from an increase of 0.016 for Electronics Information to an increase of 0.028 for Automotive Information. The largest difference in mean biserial correlation was an increase of 0.015 for General Science in the males-only group.

Table 33 shows descriptive statistics for the IRT parameters computed on the total calibration sample using the equivalent-groups design. Mean a parameters ranged from 1.004 for Electronics Information to 1.485 for Mathematics Knowledge. Minimum a parameters ranged from a low of 0.400 (the program's lower bound) to 0.423 for Mathematics Knowledge. Maximum a parameters were uniformly 2.500 (the program's upper bound) except in Shop Information, where the maximum a parameter was 2.004. In general, the a parameters were somewhat lower than those expected based on the pretesting data used to select items for calibration.

Mean b parameters ranged from -0.866 for Word Knowledge to -0.019 for Mechanical Comprehension. Minimum b values were uniformly -3.000 (the program's lower bound). Maximum b parameters ranged from 2.001 for Mathematics Knowledge to 3.000 (the program's upper bound). Mean b parameters for Shop Information, Mathematics Knowledge, and Mechanical Comprehension were all near the desired value of 0.0. Mean b parameters for the remaining areas were substantially lower than the desired value, suggesting that the items selected for calibration were too easy. This is in sharp contrast to the pretest analyses which suggested that the pool contained an insufficient number of easy items and an adequate number of difficult ones. It is also, in part, a reflection of the steps that were taken to remedy the anticipated problem of too few easy items.

Mean c parameters, shown in Table 33, ranged from a low of 0.149 for Mathematics Knowledge to a high of 0.213 for Word Knowledge. Minimum c parameters ranged from 0.000 to 0.040. Maximum c parameters ranged from a low of 0.390 for Paragraph Comprehension to a high of 0.630, which was the program's upper bound for the Mechanical Comprehension items with three alternatives.

Table 34 shows comparable statistics for the males-only calibration sample. The largest change for the a parameters was in Automotive Information; the mean a parameter for Automotive Information was 0.068 lower for the males-only group than for the total sample. Mean b parameters for the content areas reflected the changes in proportions correct between the two samples. Mean c parameters for the males-only sample were quite similar to those for the total sample.

Analyses of variance using the matched data to investigate differences in proportion-correct scores on the operational tests for samples that took different experimental tests showed few significant differences. Only five of the 54 comparisons were statistically significant ($p < .05$), and four of those were in one content area, Shop Information. For all other areas the results clearly showed that only chance deviations occurred among the ability distributions of samples given different experimental tests within a content area. Adjustments to place the parameters on a common metric were therefore not required for these areas.

Table 35 shows that samples taking operational forms 10a, 10b, 10x, and 10y performed quite differently across the four experimental Shop Information forms. Although this may be a chance significance, the fact that four of six comparisons were significant suggests it was not. This particular content area presents special problems for making corrections, however. Because the number of shop-specific items in each operational subtest was small (i.e., nine), a correction using conventional score means and standard deviations as substitutes for theta moments was unacceptable. Similarly, nine items are insufficient for estimating IRT parameters. Thus, there appeared to be two acceptable options: the parameters could be left unchanged, or the operational item parameters could be estimated by pooling the test forms and then making the appropriate transformations. Because the latter option is similar, but inferior, to the joint-calibration approach, no parameter adjustments were made for the experimental Shop Information items.

Table 36 shows descriptive statistics for the IRT parameters computed using the joint-calibration procedure for the total matched sample. Mean $\underline{a}$ parameters ranged from a low of 0.965 in Mechanical Comprehension to a high of 1.436 in Mathematics Knowledge. Minimum $\underline{a}$ parameters ranged from 0.400 to 0.467. Maximum $\underline{a}$ parameters ranged from 2.089 to 2.500. The $\underline{a}$ parameters for all content areas were somewhat lower in joint-calibration analyses than in the equivalent-groups analyses.

Mean $\underline{b}$ parameters ranged from a low of -0.903 for Word Knowledge to a high of -0.025 for Mechanical Comprehension. Minimum $\underline{b}$ parameters were uniformly -3.000. Mean $\underline{b}$ parameters were all slightly more negative in the joint-calibration analyses than in the equivalent-groups analyses. The difference ranged from a decrease of 0.024 for Shop Information to a decrease of 0.070 for Arithmetic Reasoning.

Mean $\underline{c}$ parameters ranged from 0.153 for Mathematics Knowledge to 0.205 for Word Knowledge. Minimum $\underline{c}$ parameters ranged from 0.000 for both Mathematics Knowledge and Mechanical Comprehension to 0.040 for Paragraph Comprehension. Maximum $\underline{c}$ parameters ranged from 0.370 for Paragraph Comprehension to 0.620 for Mechanical Comprehension. The mean $\underline{c}$ parameters were generally similar to those observed in the equivalent-groups analyses. The largest difference observed was in Mechanical Comprehension, in which the mean $\underline{c}$ parameter rose from 0.171 in the equivalent-groups design to 0.181 in the joint-calibration design.

Table 37 shows the distribution of IRT difficulty parameters for all items based on the equivalent-groups analyses. As discussed earlier, an ideal item pool for an adaptive test would include items with difficulties that were rectangularly distributed over the range of

ability in the target population. Ideally, the items calibrated in this study would be evenly distributed in the five categories ranging from $b \geq -2.5$ to $b < 2.5$. Slight to substantial deficiencies are noted in all content areas in the difficulty category with b between 1.5 and 2.5. Deficiencies are noted in the easy range from -2.5 to -1.5 in all areas except Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension. Adequate numbers of items appeared in the three central categories for all content areas.

Good items for an adaptive item pool should also have reasonably high discriminations and appropriately low guessing parameters. Table 38 presents data similar to the data presented in Table 37 except that items included all had an a parameter greater than or equal to 0.65 and a c parameter less than 0.301 (0.501 for the three-alternative Mechanical Comprehension items). Ideally, 200 items in each content area would be uniformly distributed across the five categories. Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension contained more items than required in the easy category while the remaining areas showed deficiencies. In the next category, all areas except Paragraph Comprehension, Mathematics Knowledge, and Mechanical Comprehension showed slight deficiencies. In the middle-difficulty category, General Science, Word Knowledge, Paragraph Comprehension, and Electronics Information showed deficiencies, the largest being a deficiency of 13 items in Word Knowledge. In the category for $0.5 \leq b < 1.5$, Word Knowledge, Paragraph Comprehension, Automotive Information, and Electronics Information showed slight deficiencies. In the most difficult category, all content areas showed moderate deficiencies. Arithmetic Reasoning and Paragraph Comprehension each had only nine items in the most difficult category. The content area most adequately represented in the most difficult category was Shop Information, with 33 items. Total numbers of items with $a \geq 0.65$, $-2.5 < b \leq 2.5$, and $c < 0.301$ (0.501 for three-alternative items) ranged from 125 in Electronics Information to 208 in Mathematics Knowledge. Overall, 1,574 items met these criteria in the equivalent-groups analyses.

Table 39 shows the distribution of IRT difficulty parameters resulting from the joint-calibration analyses. Generally, the same results that were observed in Table 37 for the equivalent-groups design can be seen in Table 39. The only notable difference is the slight tendency for items to appear easier in the joint-calibration data. Table 40 presents the distribution of IRT difficulty parameters obtained in the joint-calibration design for items with a parameters greater than 0.65 and c parameters less than 0.301 (0.501 for the three-alternative Mechanical Comprehension items). The results shown in Table 40 are similar to those shown for the equivalent-groups design except that fewer of the items met the criteria in the joint-calibration design. Math Knowledge again had 208 items which met the criteria. Electronics Information had three more items which met the criteria than in the equivalent-groups analyses. All of the remaining areas had slightly

fewer acceptable items as a result of the joint-calibration analyses. Appendix C shows item pool information functions calculated using the joint-calibration parameters for all of the newly developed items in each content area. These information functions are calculated for a normal distribution of ability.

Further Analyses of RTC and MEPS Population Differences. Previous data and conventional wisdom suggested that examinees in the RTC population on which the items were pretested should have had higher ability in most content areas than MEPS examinees. The pretest item parameters were therefore transformed to estimate the parameters that would have been obtained had the items been pretested in a MEPS population. As shown in the calibration analyses, the final item pools contained items that were, in general, easier than desired. This was apparently due to the parameter transformations made in pretesting and the use of the transformed parameters in selecting items to achieve a rectangular distribution of difficulties with $-2.5 < b < 2.5$.

Table 41 presents mean item statistics for only those items that were administered in both pretesting and calibration. Table 41 shows that the proportions correct in the pretesting and calibration samples ranged from virtually identical--as in the cases of General Science, Automotive Information, Shop Information, and Mathematics Knowledge--to substantial increases in proportions correct in the calibration group. In Arithmetic Reasoning, for example, the mean proportion correct was 0.075 higher in the calibration group than in the pretesting group. Although the calibration group consisting of MEPS examinees was expected to have generally lower proportions correct than the pretest group consisting of RTC examinees, the average proportions correct were actually lower in the RTC sample for all areas except Shop Information. No consistent differences appeared between the biserial correlation coefficients for the pretesting and calibration samples.

Similarly, no consistent differences appeared in the level of the IRT a parameters between the two groups. The assumptions made in the pretesting phase that the MEPS calibration sample should be more heterogeneous than the pretesting sample suggested that the a parameters on equivalent items should be higher in the MEPS calibration. The mean IRT difficulty parameters ranged from being virtually identical for the two samples to being substantially lower in the calibration group, as in Arithmetic Reasoning. In Arithmetic Reasoning, the b parameters were 0.512 lower in calibration than in pretesting. This suggests that the calibration group was of somewhat higher ability than the pretesting group rather than somewhat lower ability as had been expected. No noteworthy trends were apparent in the IRT c parameters (also shown in Table 41) when the two groups were compared.

Table 42 also presents the mean IRT b parameters for the items that were administered in both pretesting and calibration and the

estimated MEPS parameters based on the pretesting data and the transformations. In all content areas, the actual b parameters obtained in the MEPS calibration sample were substantially lower than those predicted from the RTC data.

Five possible explanations for the discrepancy between predicted and actual MEPS parameters are suggested. First, it is possible that some peculiarity in the IRT calibration procedures resulted from the use of the two different populations. Second, the transformation used to estimate MEPS parameters from the RTC data could have been incorrect. Third, some error in form assignment or data editing might have confounded one of the data sets. A fourth possibility is that the MEPS population increased in ability between the time of the pretest and the time of the calibration. The final possibility is that the conventional test scores used for the RTC to MEPS transformation were incorrect or inappropriate.

Data presented in Tables 41 and 42 suggest that the difference was not due to a peculiarity in IRT calibration procedures or to inappropriate transformations of the RTC data. Both the conventional proportion-correct statistics and the IRT difficulty statistics suggest that the items were at least as difficult for the pretest group as they were for the calibration group. If the problem were limited to a peculiarity in calibration or to the transformations used, the proportions correct in the pretest sample should have been uniformly higher than those observed in the calibration sample.

It is virtually impossible to prove that an error did not occur in the assignment and editing process that preceded the data analyses. All of the procedures used, however, were carefully scrutinized in light of these findings, and no errors were detected that could have resulted in the reversed pattern of difficulties.

The fourth possibility, that the level of ability of the MEPS examinees (and consequently the RTCs examinees) increased in the year between the pretest and calibration, is a reasonable hypothesis. Military service became a more attractive occupational option for many individuals as the state of the economy worsened during this period of time. Table 43, which contains data from Gialluca, Crichton, Vale, and Ree (1984), shows a distribution of AFQT percentile scores for the periods October to December 1981 and October to December 1982. Column three of Table 43 shows normal curve z-score equivalents corresponding to the percentile midpoints of each of the intervals. Using these z-scores and the raw category proportions, mean AFQT z-scores can be computed for the two groups. The mean z-score for the 1981 group was -0.123. The mean for the 1982 group was 0.037. The difference between these two values is 0.160. This difference, while confirming that the 1982 group had a higher ability than the 1981 group, is still considerably smaller than the differences observed

between the predicted and actual MEPS difficulty parameters shown in Table 42.

The remaining explanation for the difference between predicted and actual MEPS difficulties is that the means and standard deviations used in making the transformations were inappropriate. There were three possible problems in the data provided for these transformations. The first problem was that the scores were conventional number-correct scores, or their linear transformations, rather than IRT theta estimates. No correction for this problem was obvious, and it was not viewed as a serious problem at the time the transformations were made. Second, the data in the paper were computed using samples explicitly selected on a composite formed by some of the subtest scores.

Although a correction could have been made for this second problem, it would probably have reduced the differences between MEPS and RTC means by only about 10%, the amount of measurement error in the observed test-score variance. The final possible problem was that the combined Air Force and Army sample was assumed to be representative of the overall RTC population.

Data presented by Prestwood, Vale, Massey, and Welsh (1985) suggest that theta estimates of RTC groups exceed those of contemporary MEPS groups by about 0.14 theta units (see Tables 48 through 55 in Prestwood et al.). This is roughly the difference observed between the 1981 and 1982 mean AFQT scores in the data cited above and suggests that the 1981 RTC examinees should have ability levels nearly equal to 1982 MEPS examinees. This is essentially what was found in this study but is substantially different from the conclusions drawn from the draft AFHRL report titled "ASVAB Form 8b and AFQT-7A Summary Distributional Statistics for MEPS, Air Force Qualified, and Army Qualified Samples." This suggests that the data in that report were, for whatever reason, inappropriate for making the transformations.

In summary, the data in the two studies previously cited (Gialluca, Crichton, Vale, & Ree, 1984; Prestwood, Vale, Massey, & Welsh, 1985) support the observations made in this study that the MEPS and RTC examinees are much more similar in ability than was initially believed. They also suggest that many of the items that were excluded from further evaluation after pretesting because they appeared to be too difficult may be very good items for calibration and use in the MEPS population.

VI. CONCLUSIONS AND RECOMMENDATIONS

A total of 3,973 items were written and administered to examinees in RTCs and MEPS. Of these, 2,118 items were calibrated in the MEPS using both equivalent-groups and joint-calibration procedures. The joint-calibration approach probably provided better estimates of the parameters since more items were used to estimate examinee ability. The joint-calibration method also implicitly linked the parameters in the CAT and operational item pools. For these reasons, the joint-calibration parameters are suggested for operational use.

The joint-calibration analyses resulted in 1,551 items with $a \geq 0.65$, $-2.5 < b \leq 2.5$, and $c < 0.301$ (0.501 for the three-alternative items in Mechanical Comprehension). Although there is an adequate number of items in each of the nine content areas for operational implementation of the CAT ASVAB, each area is somewhat lacking in difficult items. These deficiencies are due in large part to the RTC to MEPS transformation that was applied to the pretesting data. Many of the pretested items that were discarded because they appeared to be too difficult had high discriminations. They would probably be very useful for operational testing in the MEPS if they were calibrated using sufficiently large samples from the MEPS population.

All of the items calibrated were administered in printed form and the examinees responded using optically scannable answer sheets. Computerized administration and response encoding may affect the manner in which the items function. This may be especially true for items requiring illustrations. Items with illustrations are found in the Automotive Information, Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronics Information content areas. When the hardware for administering the CAT ASVAB is selected, the adequacy of the parameters for items in these areas, estimated under traditional paper-and-pencil conditions, should be carefully investigated.

Future calibrations can be accomplished on-line while the operational data are collected. Many of the pretested items not selected for calibration in this study are excellent candidates for early on-line calibration.

VII. REFERENCES

Better Homes and Gardens. (1980). Complete guide to home repair, maintenance, and improvement. Des Moines, IA: Meredith Corporation.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Ellinger, H. E. (1977). Automechanics (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Feirer, J. L. (1977). Cabinetmaking and millwork (rev. ed.). Peoria, IL: Chas. A. Bennett.

Gerrish, H. H., & Dugger, W. E. (1980). Electricity and electronics. South Holland, IL: Goodheart-Willcox.

Gialluca, K. A., Crichton, L. I., Vale, C. D., & Ree, M. J. (1984). Methods for equating mental tests (AFHRL-TR-84-35, AD-A149544). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Grob, B. (1977). Basic electronics (4th ed.). New York: McGraw-Hill.

Hambleton, R. K., & Cook, L. (1977). Latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Heimler, C. H., & Price, J. (1977). Focus on physical science. Columbus, OH: Merrill.

Jackson, A., & Day, D. (1978). Tools and how to use them. New York: Alfred A. Knopf.

Keeton, W. T. (1968). Elements of biological science. New York: Norton.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Matt, S. R. (1980). Electricity and basic electronics. South Holland, IL: Goodheart-Willcox.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). Armed Services Vocational Aptitude Battery: Development of forms 11, 12, and 13 (AFHRL-TR-85-16). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Reader's Digest Association. (1973). Complete do-it-yourself manual. Pleasantville, NY: Author.

Reckase, M. D. (1978). Unifactor latent trait models applied to multifactor tests: Results and implications. In D. J. Weiss (ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 354-372). Minneapolis: Psychometric Methods Program, University of Minnesota.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, No. 17.

Samejima, F. (1972). A general model for free-response data. Psychometrika Monograph Supplement, No. 18.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. Psychometrika, 39, 111-121.

Stockel, M. W. (1978). Auto mechanics fundamentals. South Holland, IL: Goodheart-Willcox.

Thorndike, E. L., & Lorge, I. (1944). The teacher's word book of 30,000 words. New York: Columbia University Bureau of Publications.

Toboldt, W. K., & Johnson, L. (1981). Automotive encyclopedia. South Holland, IL: Goodheart-Willcox.

Walker, J. R. (1973). Modern metalworking. South Holland, IL: Goodheart-Willcox.

Weiss, D. J. (1974). Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: Psychometric Methods Program, University of Minnesota.

Weiss, D. J., & Suhadolnik, D. (1982). Robustness of adaptive testing strategies to multidimensionality. Paper presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill, MN.

Winer, J. B. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Table 1

Domain Coverage in Pretesting for General Science

Area	N	%
Life Science		
Human and Animal	92	20.9
Plant	39	8.8
Ecology	16	4.3
Cellular	34	7.7
Physical Science		
Chemistry	65	14.7
Work and Energy	31	7.0
Electricity and Magnetism	10	2.3
Sound	9	2.0
Measurement	15	3.4
Light	10	2.3
Heat	10	2.3
Miscellaneous	10	2.3
Earth Science		
Astronomy	36	8.2
Weather	25	5.7
Geology	34	7.7
Important Names in Science	5	1.1
Total	441	100.7

Note. Total percentage does not equal 100.0 because of rounding.

Table 2

Domain Coverage in Pretesting for Arithmetic Reasoning

Area	N	%
1. Basic Operations	40	10.0
2. Algebraic Manipulations	32	8.0
3. Percentages	26	6.5
4. Rate-type Problems	40	10.0
5. Unit Conversion Problems	9	2.2
6. Simple Geometry	12	3.0
7. Two Operations	160	40.0
8. Three or More Operations	81	20.2
Total	400	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 3

Domain Coverage in Pretesting for Paragraph Comprehension

Area	N	%
Recall of literal detail	189	46.9
Paraphrase or summary of passage	34	8.4
Recognition of main idea	42	10.4
Inference regarding material in passage	56	13.9
Application of passage material	21	5.2
Recognition of sequential or cause-and-effect relationships	25	6.2
Recognition of comparison relationships	36	8.9
Total	403	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 4

Domain Coverage in Pretesting for Automotive Information

Area	N	%
Basic Engine	53	13.0
Lubrication	15	3.7
Cooling	16	3.9
Fuel	37	9.1
Battery	15	3.7
Starter	11	2.7
Charging System	12	2.9
Ignition System	16	3.9
Intake/Exhaust System	21	5.1
Engine Testing and Service	39	9.6
Clutch	15	3.7
Transmission--Standard	15	3.7
Transmission--Automatic	24	5.9
Differential and Rear Axle	15	3.7
Brakes	29	7.1
Tires	16	3.9
Suspension	14	3.4
Accessories	20	4.9
Body and Body Repair	4	0.9
Steering	16	3.9
Basic Operating Procedures	5	1.2
Total	408	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 5

Domain Coverage in Pretesting for Shop Information

Area	N	%
Tools		
Marking/Measuring	21	5.2
Cutting	45	11.2
Shaping	32	8.0
Fastening/Welding	29	7.2
Sanding/Grinding	14	3.5
Drilling	20	5.0
Construction	23	5.7
Miscellaneous Tools	14	3.5
Materials		
Wood	31	7.7
Metal	37	9.2
Fastening	45	11.2
Miscellaneous Materials	29	7.2
Miscellaneous		
Design/Blueprints	33	8.2
Shop Safety	7	1.7
Other	21	5.2
Total	401	99.7

Note. Total percentage does not equal 100.0 because of rounding.

Table 6

Domain Coverage in Pretesting for Mathematics Knowledge

Area	N	%
Converting Fractions	32	8.0
Reducing and Building Fractions	5	1.2
Reciprocals	15	3.7
Comparing Fractions	3	0.8
Least Common Denominators	25	6.2
Prime Numbers	12	3.0
Factorials	8	2.0
Writing Equations	10	2.5
Absolute Values	4	1.0
Rounding and Place Values	6	1.5
Order of Operations	5	1.2
Cartesian Coordinates	10	2.5
Logarithms	3	0.8
Geometry: Lines	25	6.2
Geometry: Equations	10	2.5
Geometry: Basic Forms	12	3.0
Geometry: Angles	6	1.5
Geometry: Areas	7	1.8
Geometry: Volumes	6	1.5
Geometry: Perimeters	6	1.5
Exponents, Roots, and Powers	40	10.0
Polynomials	30	7.5
Solving Equations	90	22.5
Solving Inequalities	30	7.5
Total	400	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 7

Domain Coverage in Pretesting for Mechanical Comprehension

Area	N	%
Gears	63	15.8
Ratchet Mechanisms	5	1.2
Pulleys -- Lifting	21	5.2
Pulleys -- Driving	21	5.2
Wheels	4	1.0
Cams	13	3.2
Levers -- Machines	26	6.5
Levers -- Pivoting Arms	13	3.2
Screws	6	1.2
Cranks	3	0.8
Crankshafts	8	2.0
Pendulums	8	2.0
Mobiles	7	1.8
Springs	7	1.8
Scales	9	2.2
Miscellaneous Devices	28	7.0
Complex Machines	62	15.5
Stress and Supports	17	4.2
Hydraulic Systems	25	6.2
Pneumatic Systems	15	3.8
Physics	39	9.8
Total	400	99.6

Note. Total percentage does not equal 100.0 because of rounding.

Table 8

Domain Coverage in Pretesting for Electronics Information

Area	N	%
Devices	102	25.2
Term Recognition	30	7.5
Term Definition	73	18.2
Advanced Electronics	90	22.4
Physics	38	9.5
Important Names	15	3.7
Instruments	18	4.5
Household Electrics	20	5.0
Schematic Diagrams	15	3.7
Total	401	99.7

Note. Total percentage does not equal 100.0 because of rounding.

Table 9

Number and Length of Booklets per Content Area in Pretesting

Content Area	Number of Booklets	Number of Items per Booklet	Total Number of Items
General Science (GS)	7	63	441
Arithmetic Reasoning (AR)	10	39-41	400
Word Knowledge (WK)	4	100	400
Paragraph Comprehension (PC)	12	33-35	403
Automotive Information (AI)	7	58-59	408
Shop Information (SI)	7	57-58	401
Mathematics Knowledge (MK)	9	44-45	400
Mechanical Comprehension (MC)	8	50	400
Electronics Information (EI)	7	57-58	401
Total	71		3654

Table 10

Participating RTCs and Assigned Examinee Quotas

RTC	Quota
Air Force	
Lackland AFB	4150
Army	
Ft. Bliss	1350
Ft. Dix	1350
Ft. Jackson	1350
Ft. Knox	1350
Ft. Leonard Wood	1350
Ft. McClellan	1350
Ft. Sill	1350
Marine Corps	
Parris Island	1200
San Diego	1200
Navy	
San Diego	3950

Table 11

Data Used to Transform RTC Pretesting Data to a MEPS-based Metric

Content Area	MEPS		Army & Air Force	
	Mean	SD	Mean	SD
General Science	15.087	4.920	17.853	3.802
Arithmetic Reasoning	17.087	7.122	21.421	5.794
Word Knowledge	23.409	7.558	28.085	4.887
Paragraph Comprehension	9.824	3.340	11.747	2.299
Automotive Information	15.307	5.682	17.865	4.686
Shop Information	15.307	5.682	17.865	4.686
Mathematics Knowledge	11.170	5.413	13.788	5.487
Mechanical Comprehension	14.117	5.385	16.709	4.716
Electronics Information	11.503	4.255	13.725	3.445

Table 12

Participating RTCs and Examinees Pretested

RTC	Number of Examinees	Percent of Quota
Air Force - Lackland AFB	4466	107.6
Army	9591	101.5
(Ft. Bliss)	(1119)	(82.9)
(Ft. Dix)	(1358)	(100.6)
(Ft. Jackson)	(1942)	(143.9)
(Ft. Knox)	(1404)	(83.4)
(Ft. Leonard Wood)	(1015)	(75.2)
(Ft. McClellan)	(1427)	(105.5)
(Ft. Sill)	(1326)	(98.2)
Marine Corps	3037	126.5
(Parris Island)	(499)	(41.6)
(San Diego)	(2538)	(211.5)
Navy - San Diego	3999	101.2

Table 13

Number of Examinees per Booklet in Pretesting

Booklet	Correct or Recoded Form Numbers	Removed in Editing	Final Number	Booklet	Correct or Recoded Form Numbers	Removed in Editing	Final Number
1	340	5	335	38	277	0	277
2	339	0	339	39	284	1	283
3	333	2	331	40	288	5	283
4	335	1	334	41	284	6	278
5	330	0	330	42	281	2	279
6	332	0	332	43	282	0	282
7	330	0	330	44	284	4	280
8	341	1	340	45	273	1	272
9	321	0	321	46	275	2	273
10	322	4	318	47	272	2	270
11	317	2	315	48	280	1	279
12	304	3	301	49	276	0	276
13	311	0	311	50	277	0	277
14	312	4	308	51	285	3	282
15	320	1	319	52	250	2	248
16	336	1	335	53	271	0	271
17	328	2	326	54	261	0	261
18	328	2	326	55	271	0	271
19	320	5	315	56	262	0	262
20	317	3	314	57	267	1	266
21	311	0	311	58	275	1	274
22	314	0	314	59	275	0	275
23	302	1	301	60	266	3	263
24	327	5	322	61	264	0	264
25	317	1	316	62	269	1	268
26	318	2	316	63	241	1	240
27	320	2	318	64	256	2	254
28	298	3	295	65	245	1	244
29	293	2	291	66	273	1	272
30	303	4	299	67	264	1	263
31	319	0	319	68	255	2	253
32	294	0	294	69	267	3	264
33	297	0	297	70	259	0	259
34	307	1	306	71	253	1	252
35	285	1	284				
36	283	0	283	Total	20,843	107	20,736
37	277	2	275				

Table 14

Descriptive Statistics for Conventional Item Statistics Computed in Pretesting

Statistic	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
Proportion Correct									
Mean	0.534	0.435	0.626	0.617	0.475	0.453	0.459	0.450	0.392
Standard Deviation	0.239	0.219	0.242	0.173	0.242	0.208	0.185	0.186	0.230
Variance	0.057	0.048	0.058	0.030	0.059	0.043	0.034	0.035	0.053
Skew	-0.064	0.525	-0.589	-0.412	0.253	0.351	0.085	0.146	0.516
Kurtosis	-1.177	-0.667	-0.867	-0.472	-1.175	-0.801	-0.590	-0.637	-0.822
Minimum	0.069	0.053	0.066	0.125	0.060	0.063	0.060	0.018	0.012
Maximum	0.946	0.984	0.973	0.951	0.945	0.934	0.916	0.929	0.932
Biserial Correlation									
Mean	0.485	0.479	0.613	0.603	0.463	0.443	0.561	0.435	0.362
Standard Deviation	0.197	0.174	0.273	0.168	0.219	0.198	0.160	0.188	0.197
Variance	0.039	0.030	0.074	0.028	0.048	0.039	0.026	0.036	0.039
Skew	-0.294	-0.983	-0.787	-0.546	-0.617	-0.468	-1.008	-0.725	-0.525
Kurtosis	-0.172	0.844	0.459	0.445	0.012	-0.365	1.093	0.821	0.100
Minimum	-0.101	-0.174	-0.365	0.015	-0.158	-0.137	-0.015	-0.287	-0.260
Maximum	1.000	0.767	1.000	1.000	0.879	0.852	0.865	0.827	0.771
Point-Biserial Correlation									
Mean	0.351	0.357	0.428	0.448	0.340	0.334	0.430	0.334	0.268
Standard Deviation	0.135	0.136	0.173	0.116	0.157	0.150	0.132	0.145	0.147
Variance	0.018	0.019	0.030	0.013	0.025	0.022	0.018	0.021	0.022
Skew	-0.642	-0.796	-1.140	-0.876	-0.690	-0.480	-0.871	-0.678	-0.376
Kurtosis	-0.140	0.258	1.219	0.914	0.002	-0.406	0.590	0.379	-0.403
Minimum	-0.060	-0.095	-0.206	0.009	-0.099	-0.097	-0.009	-0.162	-0.134
Maximum	0.615	0.600	0.695	0.678	0.637	0.619	0.689	0.626	0.608
Number of Items	435	400	400	401	405	397	399	395	400

Table 15

Descriptive Statistics for IRT Parameters Computed in Pretesting

Parameter	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
a									
Mean	1.085	1.269	1.180	1.062	1.197	1.097	1.273	1.040	1.182
Standard Deviation	0.385	0.399	0.525	0.412	0.430	0.389	0.410	0.407	0.451
Variance	0.148	0.159	0.276	0.170	0.185	0.152	0.168	0.165	0.204
Skew	0.263	-0.028	-0.041	0.476	0.135	0.217	0.072	0.253	0.036
Kurtosis	-0.661	-0.708	-1.045	-0.324	-0.564	-0.526	-0.633	-1.028	-1.008
Minimum	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400
Maximum	2.030	2.227	2.250	2.279	2.400	2.253	2.248	2.118	2.319
b									
Mean	0.455	0.864	-0.404	-0.131	0.701	0.902	0.727	0.925	1.328
Standard Deviation	1.393	1.266	1.636	1.110	1.405	1.227	0.991	1.123	1.380
Variance	1.941	1.602	2.677	1.233	1.974	1.505	0.982	1.260	1.905
Skew	0.015	-0.432	0.001	-0.363	-0.279	-0.229	0.049	-0.059	-0.497
Kurtosis	-0.837	-0.379	-0.727	0.488	-0.650	-0.546	-0.203	-0.448	-0.862
Minimum	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-2.718	-3.000	-3.000
Maximum	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000
c									
Mean	0.194	0.189	0.195	0.199	0.189	0.191	0.182	0.195	0.185
Standard Deviation	0.036	0.043	0.034	0.024	0.039	0.040	0.045	0.034	0.043
Variance	0.001	0.002	0.001	0.001	0.001	0.002	0.002	0.001	0.002
Skew	-0.404	-0.361	-0.505	-1.161	-0.466	-0.187	-0.290	-0.484	-0.280
Kurtosis	1.819	0.178	2.447	8.187	0.609	0.801	-0.433	2.166	0.087
Minimum	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
Maximum	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300
Number of Items	435	400	400	401	405	397	399	395	400

Table 16

Descriptive Statistics for Transformed IRT Parameters Computed in Pretesting

Parameter	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
a									
Mean	1.404	1.560	1.825	1.543	1.451	1.330	1.256	1.188	1.460
Standard Deviation	0.498	0.491	0.812	0.598	0.521	0.472	0.404	0.464	0.557
Variance	0.248	0.241	0.659	0.358	0.272	0.223	0.164	0.216	0.311
Skew	0.263	-0.028	-0.041	0.476	0.135	0.217	0.072	0.253	0.036
Kurtosis	-0.661	-0.706	-1.045	-0.324	-0.564	-0.526	-0.633	-1.028	-1.008
Minimum	0.518	0.492	0.619	0.581	0.485	0.485	0.395	0.457	0.494
Maximum	2.627	2.737	3.480	3.312	2.910	2.732	2.218	2.418	2.864
b									
Mean	0.914	1.311	0.358	0.486	1.029	1.194	1.220	1.218	1.597
Standard Deviation	1.077	1.030	1.058	0.764	1.159	1.012	1.005	0.983	1.118
Variance	1.159	1.060	1.119	0.584	1.342	1.024	1.009	0.967	1.249
Skew	0.015	-0.432	0.001	-0.363	-0.279	-0.229	0.049	-0.059	-0.497
Kurtosis	-0.837	-0.379	-0.727	0.488	-0.650	-0.546	-0.203	-0.448	-0.862
Minimum	-1.756	-1.832	-1.321	-1.489	-2.024	-2.024	-2.272	-2.157	-1.907
Maximum	2.880	3.049	2.558	2.641	2.924	2.924	3.525	3.097	2.951
c									
Mean	0.194	0.189	0.195	0.199	0.189	0.191	0.182	0.195	0.185
Standard Deviation	0.036	0.043	0.034	0.024	0.039	0.040	0.045	0.034	0.043
Variance	0.001	0.002	0.001	0.001	0.001	0.002	0.002	0.001	0.002
Skew	-0.404	-0.361	-0.505	-1.161	-0.466	-0.187	-0.290	-0.484	-0.280
Kurtosis	1.819	0.178	2.447	8.187	0.609	0.801	-0.433	2.166	0.087
Minimum	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
Maximum	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300
Number of Items	435	400	400	401	405	397	399	395	400

Table 17

Distribution of Items by Estimated Difficulty Parameters on the MEPS Metric

Difficulty Range	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
$b < -2.50$	0	0	0	0	0	0	0	0	0
$-2.50 \leq b < -1.32$	5	4	49	15	9	2	1	2	1
$1.32 \leq b < -0.44$	38	16	49	29	43	24	19	12	15
$-0.44 \leq b < 0.44$	115	69	100	145	78	71	71	62	64
$0.44 \leq b < 1.32$	119	100	137	165	91	114	133	142	78
$1.32 \leq b < 2.50$	119	166	53	46	138	143	130	126	128
$2.50 \leq b$	39	45	12	1	46	43	45	51	114
Total Analyzed	435	400	400	401	405	397	399	395	400

Table 18

Distribution of Items Selected for Calibration by Estimated Difficulty Parameter on the MEPS Metric

Difficulty Range	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
$-2.50 \leq b < -1.32$	4	2	20	15	8	1	0	1	1
$1.32 \leq b < -0.44$	35	15	38	25	37	23	18	10	13
$-0.44 \leq b < 0.44$	54	63	52	56	51	55	54	55	56
$0.44 \leq b < 1.32$	54	58	52	60	52	58	64	69	63
$1.32 \leq b < 2.50$	53	59	48	43	52	60	64	65	63
Total Selected	200	197	210	199	200	197	200	200	196

Table 19

Number of Items Selected for and Written for Calibration

Content Area	Selected	Written	Total
General Science	200	28	228
Arithmetic Reasoning	197	48	245
Word Knowledge	210	48	258
Paragraph Comprehension	199	32	231
Automotive Information	200	40	240
Shop Information	197	31	228
Mathematics Knowledge	200	30	230
Mechanical Comprehension	200	30	230
Electronics Information	196	32	228
Total	1799	319	2118

Table 20

Domain Coverage in Calibration for General Science

Area	N	%
Life Science		
Human and Animal	44	19.3
Plant	17	7.5
Ecology	12	5.3
Cellular	12	5.3
Physical Science		
Chemistry	44	19.3
Work and Energy	10	4.4
Electricity and Magnetism	6	2.6
Sound	5	2.2
Measurement	8	3.5
Light	5	2.2
Heat	8	3.5
Miscellaneous	5	2.2
Earth Science		
Astronomy	19	8.3
Weather	13	5.7
Geology	15	6.6
Important Names in Science	5	2.2
Total	228	100.1

Note. Total percentage does not equal 100.0 because of rounding.

Table 21

Domain Coverage in Calibration for Arithmetic Reasoning

Area	N	%
1. Basic Operations	75	30.6
2. Algebraic Manipulations	26	10.6
3. Percentages	14	5.7
4. Rate-type Problems	29	11.8
5. Unit Conversion Problems	6	2.4
6. Simple Geometry	3	1.2
7. Two Operations	70	28.6
8. Three or More Operations	22	9.0
Total	245	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 22

Domain Coverage in Calibration for Paragraph Comprehension

Area	N	%
Recall of literal detail	122	52.8
Paraphrase or summary of passage	15	6.5
Recognition of main idea	20	8.7
Inference regarding material in passage	35	15.2
Application of passage material	8	3.5
Recognition of sequential or cause-and-effect relationships	12	5.2
Recognition of comparison relationships	19	8.2
Total	231	100.1

Note. Total percentage does not equal 100.0 because of rounding.

Table 23

Domain Coverage in Calibration for Automotive Information

Area	N	%
Basic Engine	27	11.2
Lubrication	9	3.8
Cooling	12	5.0
Fuel	18	7.5
Battery	4	1.7
Starter	7	2.9
Charging System	5	2.1
Ignition System	12	5.0
Intake/Exhaust System	14	5.8
Engine Testing and Service	23	9.6
Clutch	5	2.1
Transmission--Standerd	8	3.3
Transmission--Automatic	8	3.3
Differential and Rear Axle	7	2.9
Brakes	17	7.1
Tires	9	3.8
Suspension	9	3.8
Accessories	5	2.1
Body and Body Repair	2	0.8
Steering	6	2.5
Basic Parts Recognition	15	6.2
Basic Operating Procedures	12	5.0
Basic Driving Skills	6	2.5
Total	240	100.0

Table 24

Domain Coverage in Calibration for Shop Information

Area	N	%
Tools		
Marking/Measuring	19	8.3
Cutting	28	12.3
Shaping	14	6.1
Fastening/Welding	17	7.5
Sanding/Grinding	6	2.6
Drilling	12	5.3
Construction	19	8.3
Miscellaneous Tools	6	2.6
Materials		
Wood	13	5.7
Metal	11	4.8
Fastening	33	14.5
Miscellaneous Materials	10	4.4
Miscellaneous		
Design/Blueprints	16	7.0
Shop Safety	5	2.2
Other	19	8.3
Total	228	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 25

Domain Coverage in Calibration for Mathematics Knowledge

Area	N	%
Identifying Fractions	5	2.2
Converting Fractions	21	9.1
Reducing and Building Fractions	4	1.7
Reciprocals	13	5.7
Comparing Fractions	0	0.0
Least Common Denominators	14	6.1
Prime Numbers	2	0.9
Factorials	2	0.9
Writing Equations	8	3.5
Absolute Values	2	0.9
Rounding and Place Values	1	0.4
Order of Operations	6	2.6
Cartesian Coordinates	5	2.2
Logarithms	0	0.0
Geometry: Lines	5	2.2
Geometry: Equations	1	0.4
Geometry: Basic Forms	4	1.7
Geometry: Angles	4	1.7
Geometry: Areas	3	1.3
Geometry: Volumes	1	0.4
Geometry: Perimeters	3	1.3
Exponents, Roots, and Powers	25	10.9
Polynomials	22	9.6
Solving Equations	67	29.1
Solving Inequalities	12	5.2
Total	230	100.0

Table 26

Domain Coverage in Calibration for Mechanical Comprehension

Area	N	%
Gears	31	13.5
Ratchet Mechanisms	5	2.2
Pulleys -- Lifting	15	6.5
Pulleys -- Driving	18	7.8
Wheels	5	2.2
Cams	7	3.0
Levers -- Machines	25	10.9
Levers -- Pivoting Arms	13	5.7
Screws	3	1.3
Cranks	4	1.7
Crankshafts	5	2.2
Pendulums	1	0.4
Mobiles	3	1.3
Springs	6	2.6
Scales	3	1.3
Miscellaneous Devices	17	7.4
Complex Machines	23	10.0
Stress and Supports	10	4.3
Hydraulic Systems	8	3.5
Pneumatic Systems	9	3.9
Physics	19	8.3
Total	230	100.0

Table 27

Domain Coverage in Calibration for Electronics Information

Area	N	%
Devices	62	27.2
Term Recognition	27	11.8
Term Definition	32	14.0
Advanced Electronics	12	5.3
Physics	28	12.3
Important Names	11	4.8
Instruments	9	3.9
Household Electrics	18	7.9
Schematic Diagrams	9	3.9
Device Recognition	20	8.8
Total	228	99.9

Note. Total percentage does not equal 100.0 because of rounding.

Table 28

Number of Booklets and Items per Booklet in Calibration

Content Area	Number of Booklets	Items per Booklet	Total Items
General Science	4	57	228
Arithmetic Reasoning	7	35	245
Word Knowledge	3	86	258
Paragraph Comprehension	7	33	231
Automotive Information	4	60	240
Shop Information	4	57	228
Mathematics Knowledge	5	46	230
Mechanical Comprehension	5	46	230
Electronics Information	4	57	228

Table 29

Number of Examinees per Booklet in Calibration

Booklet	Correct Form Numbers	Recoded Form Numbers	Removed in Editing	Final Number
01	3475	41	6	3510
02	3465	36	5	3496
03	3445	35	5	3475
04	3427	28	5	3450
05	3397	32	3	3426
06	3401	35	10	3426
07	3376	32	8	3400
08	3307	55	3	3359
09	3322	29	3	3348
10	3285	32	12	3305
11	3217	38	5	3250
12	3269	29	7	3291
13	3265	23	5	3283
14	3226	26	5	3247
15	3246	28	4	3270
16	3220	30	5	3245
17	3211	29	4	3236
18	3180	39	8	3211
19	3178	24	7	3195
20	3116	38	8	3146
21	3184	24	1	3207
22	3154	21	7	3168
23	3161	37	6	3192
24	3159	36	8	3187
25	3175	17	5	3187
26	3164	20	3	3181
27	3111	33	4	3140
28	3086	43	7	3122
29	3101	32	8	3125
30	3093	39	6	3126
31	3092	31	10	3113
32	3086	25	6	3105
33	3074	18	6	3086
34	3038	23	7	3054
35	3013	21	10	3024
36	2988	28	4	3012
37	2999	25	2	3022
38	2972	38	7	3003
39	2963	22	1	2984
40	2989	25	1	3013
41	2931	22	4	2949
42	2908	21	8	2921
43	2858	32	2	2888
Total	136,327	1,292	241	137,378

Table 30

Examinees with Valid Data for Joint Calibration

Content Area	Experimental Data Cases	Matched Experimental and Operational Data Cases	Percentage of Cases Matched
General Science	12,801	10,843	84.7
Arithmetic Reasoning	22,252	18,776	84.4
Word Knowledge	9,659	8,171	84.6
Paragraph Comprehension	22,523	19,097	84.8
Automotive Information	13,050	11,140	85.4
Shop Information	12,973	10,923	84.2
Mathematics Knowledge	15,581	13,138	84.3
Mechanical Comprehension	15,608	13,154	84.3
Electronics Information	12,931	10,931	84.5
Total	137,378	116,173	84.6

Table 31

Descriptive Statistics for Conventional Item Statistics Computed on the Total Calibration Sample

Statistic	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
Proportion Correct									
Mean	0.648	0.688	0.717	0.699	0.636	0.580	0.575	0.573	0.624
Standard Deviation	0.267	0.265	0.287	0.241	0.286	0.272	0.220	0.223	0.257
Variance	0.071	0.070	0.082	0.058	0.082	0.074	0.048	0.050	0.066
Skew	-0.279	-0.514	-0.804	-0.597	-0.347	0.001	0.049	-0.034	-0.175
Kurtosis	-1.340	-1.077	-0.903	-1.050	-1.287	-1.308	-1.062	-1.003	-1.101
Minimum	0.119	0.093	0.075	0.106	0.080	0.086	0.075	0.093	0.072
Maximum	0.992	0.992	0.995	0.990	0.995	0.993	0.975	0.976	0.993
Biserial Correlation									
Mean	0.562	0.630	0.661	0.650	0.568	0.531	0.638	0.527	0.494
Standard Deviation	0.150	0.115	0.216	0.144	0.156	0.145	0.133	0.126	0.130
Variance	0.023	0.013	0.047	0.021	0.024	0.021	0.018	0.016	0.017
Skew	-0.801	-1.300	-0.717	-0.672	-0.829	-0.801	-0.480	-1.065	-0.444
Kurtosis	0.548	4.462	-0.121	0.531	0.401	0.410	0.179	1.182	0.211
Minimum	0.071	-0.051	0.074	0.194	0.090	0.079	0.250	0.059	0.107
Maximum	0.821	0.857	1.000	0.985	0.820	0.771	0.926	0.789	0.799
Point-Biserial Correlation									
Mean	0.366	0.401	0.374	0.421	0.370	0.364	0.472	0.387	0.330
Standard Deviation	0.118	0.123	0.122	0.095	0.142	0.136	0.128	0.106	0.111
Variance	0.014	0.015	0.015	0.009	0.020	0.018	0.016	0.011	0.012
Skew	-0.389	-0.406	-0.444	-0.452	-0.147	-0.254	-0.380	-0.759	-0.060
Kurtosis	-0.540	-0.318	-0.530	-0.464	-0.851	-0.907	-0.439	0.525	-0.396
Minimum	0.048	-0.030	0.046	0.144	0.050	0.061	0.117	0.034	0.069
Maximum	0.619	0.640	0.593	0.596	0.647	0.592	0.725	0.626	0.638
Number of Items	228	245	258	230	240	228	230	230	226

Table 32

Descriptive Statistics for Conventional Item Statistics Computed on a Males-Only Calibration Sample

Statistic	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
Proportion Correct									
Mean	0.654	0.691	0.714	0.695	0.664	0.607	0.573	0.595	0.640
Standard Deviation	0.266	0.261	0.287	0.240	0.284	0.271	0.218	0.223	0.253
Variance	0.071	0.068	0.082	0.057	0.081	0.073	0.048	0.050	0.064
Skew	-0.307	-0.530	-0.806	-0.585	-0.479	-0.134	0.072	-0.123	-0.259
Kurtosis	-1.335	-1.040	-0.904	-1.066	-1.184	-1.311	-1.064	-0.977	-1.038
Minimum	0.116	0.099	0.072	0.108	0.085	0.099	0.081	0.089	0.078
Maximum	0.992	0.992	0.995	0.989	0.994	0.992	0.977	0.976	0.992
Biserial Correlation									
Mean	0.577	0.636	0.662	0.652	0.581	0.536	0.644	0.532	0.506
Standard Deviation	0.155	0.116	0.220	0.147	0.154	0.140	0.131	0.122	0.132
Variance	0.024	0.013	0.048	0.022	0.024	0.020	0.017	0.015	0.017
Skew	-0.707	-1.204	-0.750	-0.672	-0.798	-0.911	-0.449	-0.982	-0.420
Kurtosis	0.416	3.645	-0.126	0.604	0.725	0.687	0.208	0.833	0.265
Minimum	0.101	-0.012	0.082	0.169	0.063	0.093	0.250	0.138	0.093
Maximum	0.850	0.870	1.000	0.999	0.941	0.756	0.936	0.778	0.811
Point-Biserial Correlation									
Mean	0.372	0.406	0.377	0.425	0.367	0.360	0.477	0.386	0.334
Standard Deviation	0.118	0.123	0.124	0.096	0.134	0.125	0.126	0.103	0.109
Variance	0.014	0.015	0.015	0.009	0.018	0.016	0.016	0.011	0.012
Skew	-0.404	-0.424	-0.475	-0.505	-0.126	-0.255	-0.384	-0.685	-0.109
Kurtosis	-0.513	-0.324	-0.502	-0.291	-0.705	-0.801	-0.405	0.324	-0.314
Minimum	0.069	-0.007	0.058	0.124	0.035	0.068	0.124	0.083	0.073
Maximum	0.626	0.649	0.604	0.600	0.658	0.591	0.732	0.620	0.624
Number of Items	228	245	258	230	240	228	230	230	226

Table 33

Descriptive Statistics for Equivalent Groups IRT Parameters Computed on the Total Calibration Sample

Parameter	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
a									
Mean	1.189	1.325	1.320	1.240	1.306	1.082	1.485	1.019	1.004
Standard Deviation	0.406	0.498	0.462	0.474	0.498	0.327	0.597	0.414	0.465
Variance	0.165	0.248	0.213	0.224	0.248	0.107	0.357	0.172	0.216
Skew	0.816	0.767	0.482	0.950	0.447	0.232	0.214	1.381	1.186
Kurtosis	0.526	-0.189	-0.282	0.471	-0.481	-0.386	-1.229	2.312	1.184
Minimum	0.417	0.400	0.400	0.423	0.400	0.410	0.411	0.400	0.400
Maximum	2.500	2.500	2.500	2.500	2.500	2.004	2.500	2.500	2.500
b									
Mean	-0.420	-0.786	-0.866	-0.689	-0.442	-0.100	-0.116	-0.019	-0.296
Standard Deviation	1.571	1.515	1.703	1.362	1.635	1.590	1.085	1.251	1.614
Variance	2.469	2.295	2.899	1.855	2.674	2.528	1.178	1.564	2.606
Skew	-0.057	0.044	0.444	0.195	-0.014	-0.219	-0.400	-0.115	-0.110
Kurtosis	-1.153	-1.273	-1.040	-1.135	-1.115	-0.910	-0.444	-0.690	-1.050
Minimum	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000
Maximum	2.853	2.609	2.934	2.251	3.000	2.980	2.001	3.000	3.000
c									
Mean	0.198	0.182	0.213	0.187	0.186	0.170	0.149	0.171	0.196
Standard Deviation	0.081	0.075	0.078	0.062	0.090	0.082	0.085	0.103	0.192
Variance	0.007	0.006	0.006	0.004	0.008	0.007	0.007	0.011	0.009
Skew	0.458	0.450	0.242	0.527	0.538	0.527	0.639	1.228	0.380
Kurtosis	-0.157	0.128	-0.262	0.589	-0.302	-0.045	0.159	2.424	-0.634
Minimum	0.030	0.020	0.010	0.040	0.020	0.020	0.000	0.000	0.010
Maximum	0.400	0.400	0.400	0.390	0.400	0.400	0.400	0.630	0.400
Number of Items	228	245	258	230	240	228	230	230	226

Table 34

Descriptive Statistics for Equivalent-Groups IRT Parameters Computed on a Males-Only Calibration Sample

Parameter	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
a									
Mean	1.218	1.333	1.321	1.252	1.238	1.049	1.506	0.990	0.995
Standard Deviation	0.410	0.497	0.463	0.461	0.466	0.324	0.585	0.387	0.451
Variance	0.168	0.247	0.214	0.213	0.217	0.105	0.342	0.150	0.203
Skew	0.695	0.701	0.474	0.939	0.503	0.419	0.175	1.330	1.137
Kurtosis	0.132	-0.319	-0.378	0.501	-0.290	-0.157	-1.236	2.327	1.162
Minimum	0.442	0.400	0.405	0.464	0.400	0.403	0.448	0.400	0.400
Maximum	2.463	2.500	2.500	2.500	2.463	2.037	2.500	2.500	2.486
b									
Mean	-0.456	-0.791	-0.841	-0.658	-0.618	-0.237	-0.098	-0.141	-0.393
Standard Deviation	1.548	1.491	1.702	1.343	1.663	1.602	1.063	1.270	1.598
Variance	2.396	2.222	2.897	1.804	2.764	2.567	1.130	1.613	2.552
Skew	-0.053	0.032	0.447	0.197	0.097	-0.122	-0.419	-0.089	-0.050
Kurtosis	-1.170	-1.264	1.022	-1.131	-1.155	-0.980	-0.402	-0.612	-1.042
Minimum	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000
Maximum	2.567	2.589	2.998	2.113	3.000	2.990	1.962	3.000	3.000
c									
Mean	0.198	0.183	0.211	0.187	0.192	0.178	0.150	0.177	0.199
Standard Deviation	0.077	0.072	0.075	0.060	0.085	0.076	0.081	0.097	0.086
Variance	0.006	0.005	0.006	0.004	0.007	0.006	0.006	0.009	0.007
Skew	0.463	0.361	0.135	0.521	0.396	0.462	0.564	1.317	0.382
Kurtosis	-0.059	0.166	-0.154	0.595	-0.384	-0.204	0.009	3.107	-0.515
Minimum	0.040	0.020	0.010	0.040	0.020	0.030	0.000	0.010	0.030
Maximum	0.400	0.400	0.400	0.370	0.400	0.400	0.400	0.630	0.400
Number of Items	228	245	258	230	240	228	230	230	226

Table 35

Proportion-Correct Scores on Operational Tests for Examinees Taking Different Experimental Tests in Shop Information

Statistics	Operational Test 9a	9b	10a	10b	10x	10y
			Experimental Test 1			
Mean	0.709	0.695	0.732	0.721	0.723	0.726
Std. Dev.	0.216	0.219	0.225	0.219	0.218	0.228
			Experimental Test 2			
Mean	0.706	0.680	0.702	0.693	0.704	0.714
Std. Dev.	0.222	0.237	0.235	0.233	0.222	0.231
			Experimental Test 3			
Mean	0.703	0.696	0.686	0.667	0.672	0.670
Std. Dev.	0.215	0.230	0.228	0.230	0.229	0.237
			Experimental Test 4			
Mean	0.697	0.696	0.693	0.714	0.706	0.690
Std. Dev.	0.221	0.228	0.236	0.227	0.217	0.239
			Analyses of Variance of Arc-Sine Transformed Values			
F	0.237	0.526	3.864*	5.340*	3.986*	4.360*
df	3, 2066	3, 1968	3, 1818	3, 1843	3, 1646	3, 1557

*$p < .05$

Table 36

Descriptive Statistics for Joint-Calibration IRT Parameters Computed on the Matched Experimental/Operational Sample

Parameter	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
a									
Mean	1.137	1.175	1.267	1.150	1.259	1.074	1.436	0.965	0.975
Standard Deviation	0.341	0.387	0.409	0.399	0.459	0.326	0.547	0.342	0.425
Variance	0.117	0.150	0.168	0.159	0.211	0.106	0.299	0.117	0.181
Skew	0.410	0.737	0.318	0.941	0.329	0.285	0.242	1.122	1.077
Kurtosis	-0.303	0.279	-0.366	0.781	-0.500	-0.225	-1.093	2.019	0.882
Minimum	0.400	0.400	0.400	0.467	0.400	0.431	0.421	0.400	0.400
Maximum	2.102	2.326	2.406	2.457	2.441	2.089	2.500	2.500	2.305
b									
Mean	-0.470	-0.856	-0.903	-0.744	-0.498	-0.124	-0.141	-0.025	-0.326
Standard Deviation	1.585	1.534	1.705	1.385	1.650	1.597	1.077	1.279	1.633
Variance	2.513	2.353	2.908	1.918	2.724	2.550	1.160	1.636	2.667
Skew	-0.064	0.078	0.478	0.220	-0.030	-0.233	-0.406	-0.056	-0.129
Kurtosis	-1.194	-1.323	-1.027	-1.131	-1.167	-0.911	-0.436	-0.524	-1.066
Minimum	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000	-3.000
Maximum	2.668	2.351	3.000	2.269	3.000	3.000	1.939	3.000	3.000
c									
Mean	0.200	0.182	0.205	0.187	0.187	0.173	0.153	0.181	0.201
Standard Deviation	0.080	0.077	0.075	0.064	0.091	0.081	0.082	0.102	0.091
Variance	0.006	0.006	0.006	0.004	0.008	0.007	0.007	0.010	0.008
Skew	0.391	0.386	0.302	0.396	0.377	0.479	0.656	1.123	0.314
Kurtosis	-0.033	0.009	-0.185	0.457	-0.465	-0.057	0.190	1.868	-0.718
Minimum	0.030	0.020	0.010	0.040	0.020	0.020	0.000	0.000	0.030
Maximum	0.400	0.400	0.390	0.370	0.400	0.400	0.400	0.620	0.400
Number of Items	228	245	258	230	240	228	230	230	226

Table 41

Mean Item Statistics for Items Administered Both in Pretesting and Calibration

Mean Statistics	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
					Pretesting (RTC)				
Proportion Correct	0.603	0.548	0.622	0.613	0.570	0.527	0.526	0.512	0.537
Biserial Correlation	0.590	0.574	0.649	0.635	0.593	0.557	0.632	0.547	0.503
IRT a	1.280	1.421	1.344	1.261	1.388	1.285	1.448	1.195	1.123
IRT b	-0.012	0.154	-0.404	-0.210	0.040	0.341	0.329	0.455	0.368
IRT c	0.196	0.194	0.193	0.199	0.186	0.186	0.187	0.195	0.194
					Calibration (MEPS)				
Proportion Correct	0.607	0.623	0.662	0.661	0.571	0.522	0.534	0.530	0.572
Biserial Correlation	0.561	0.628	0.636	0.643	0.569	0.533	0.648	0.532	0.486
IRT a	1.218	1.411	1.386	1.290	1.371	1.093	1.546	1.045	1.008
IRT b	-0.138	-0.358	-0.480	-0.411	-0.010	0.282	0.101	0.222	0.075
IRT c	0.191	0.173	0.200	0.186	0.176	0.159	0.146	0.152	0.188

Note. The IRT parameters in this table are from equivalent-groups analyses.

Table 42

Mean IRT Difficulty Parameters for Items Administered in Both Pretesting and Calibration and Mean Estimated Difficulty Parameters on the Estimated MEPS Metric

Data Source/Data Metric	Content Area								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
RTCs/Pretesting Actual	-0.012	0.154	-0.404	-0.210	0.040	0.341	0.329	0.455	0.368
RTCs/MEPS Estimated	0.553	0.734	0.357	0.431	0.483	0.732	0.817	0.869	0.820
MEPS/Calibration Actual	-0.138	-0.358	-0.480	-0.411	-0.010	0.282	0.101	0.222	0.075

Table 43

Distribution of Applicants Across AFQT Categories

AFQT category	Score interval (percentile)	z-score	Oct-Dec 1981 Proportion Raw	Oct-Dec 1981 Proportion Cum.	Oct-Dec 1982 Proportion Raw	Oct-Dec 1982 Proportion Cum.
I	93-99	1.75	0.026	0.999	0.034	1.000
II	65-92	0.79	0.260	0.973	0.311	0.966
IIIa	50-64	0.18	0.154	0.713	0.172	0.655
IIIb	31-49	-0.25	0.202	0.559	0.213	0.483
IVa	21-30	-0.66	0.139	0.357	0.128	0.270
IVb	16-20	-0.92	0.081	0.218	0.064	0.142
IVc	10-15	-1.15	0.082	0.137	0.052	0.078
V	01-09	-1.64	0.055	0.055	0.026	0.026
Number of examinees			127,188		92,817	

Note. These data are for non-prior-service male applicants (first ASVAB administration) only. The data are from Gialluca, Crichton, Vale, and Ree (1984).

APPENDIX A

GUIDELINES FOR ITEM WRITERS

The Item Stem

1. The item stem should contain only one central concept or idea.

2. The item stem should be complete enough that an examinee need not read the response alternatives in order to understand what is being asked.

3. The item stem should be stated as concisely as possible to reduce reading time and avoid unnecessary complexity.

4. The item stem should be written in precise language. Highly technical terms, however, should be avoided unless they are necessary to ensure precision or unless the item is specifically designed to assess the examinee's technical vocabulary.

5. Item stems should not be phrased negatively using qualifiers such as not or least.

6. Vague quantifiers (e.g., few, many) should be replaced by more precise terms (e.g., 10 percent, a majority).

7. Item text (and illustrations where used) should avoid racial, ethnic, or sexual bias.

The Response Alternatives

1. When an item stem is an incomplete statement, each alternative should complete the statement in a grammatically correct manner (e.g., plural verb forms in the stem may require plural nouns in the alteratives). When an item stem is a complete question, the alternatives should be grammatically correct answers to the question.

2. Key words or phrases used in the stem should not be repeated in the correct response alternative because this can serve as a clue that an alternative is correct.

3. The distractors should be thoroughly wrong or clearly incorrect, yet plausible enough to appeal to less-knowledgeable examinees.

4. The distractors should be similar in length and complexity to the correct response alternative.

5. The distractors should not contain the word never or the word always since these words are often associated with false statements.

6. A distractor generally should not be opposite in meaning to the correct response or synonymous with another distractor. (Certain Mechanical Comprehension items will require violations of this principle--as in the use of right versus left or up versus down.)

7. "None of the above" should not be used as an alternative unless other plausible answers do not exist (as in the case of certain Mechanical Comprehension items) and each of the other alternatives can be clearly and unambiguously identified as correct or incorrect. "All of the above" should not be used as an alternative.

8. Alternatives which vary along a single quantitative or qualitative dimension should be ordered along that dimension. Alternatives which are single letters (A-E) or numbers (1-5) should be placed in their nominal positions.

9. The position of the correct alternative should vary so that the correct alternative is placed in each position about 20% of the time. Assignment of the position of the correct alternative should be guided by a random process.

APPENDIX B

DEVELOPMENT OF RTC TO MEPS PARAMETER TRANSFORMATIONS

The pretesting of the items yielded parameter estimates computed from a sample of examinees at RTCs. These estimates were on a metric which assumed that ability was standard (0,1) in the RTC population. The population of interest in this study was examinees in the MEPS. Relative to the MEPS population, the RTC sample is restricted with respect to ability because the recruits tested in the RTCs are selected from the MEPS population, at least in part, on the basis of their ability as assessed by the AFQT. Thus, a transformation for the RTC parameter estimates is required to estimate the values of the parameters on the MEPS ability metric. The following development describes the relationship between the RTC and the MEPS metrics.

If $\underline{u}$ is the score of an individual in the unrestricted (i.e., MEPS) sample, the standard score for that individual in the unrestricted sample is given by

$$\theta_u = (u - \bar{u}) / S_u \tag{1}$$

where θ_u = the standardized score,

u = the raw score,

$\bar{u}$ = the mean of the raw scores in the unrestricted sample, and

S_u = the standard deviation of the raw scores in the unrestricted sample.

By the definition of standard scores, the mean ability in the unrestricted sample will be zero and the standard deviation will be one. If the same transformation is applied to scores from the restricted (i.e., RTC) sample, the mean and standard deviation of the transformed scores will be equal to

$$\bar{\theta}_r = E\left[(r - \bar{u}) / S_u \right] \tag{2}$$

$$= (\bar{r} - \bar{u}) / S_u$$

and

$$S_{\theta_r} = S_r / S_u \tag{3}$$

where $\bar{\theta}_r$ = the mean of the transformed scores of the restricted sample,

S_{θ_r} = the standard deviation of the transformed scores of the restricted sample,

r = a raw score from the restricted sample,

$\bar{r}$ = the mean of the raw scores from the restricted sample, and

S_r = the standard deviation of the raw scores from the restricted sample.

Item calibration procedures typically scale the ability metric to (0,1). In so doing, both the MEPS and the RTC samples are assumed to have the same ability distribution, even though one is a selected subset of the other. By the mechanisms of IRT, however, the two ability scales differ only by a linear transformation. Having called ability in the unrestricted sample θ, we will now call ability in the restricted sample Γ. Levels of ability on the two metrics can be considered equivalent if they lead to equal predicted probabilities of correct responses. This is true when the logits are equal. Thus, the following must hold for θ and Γ:

$$a\,(\theta - b) = a^*\,(\Gamma - b^*) \qquad (4)$$

where a = the discrimination parameter for the unrestricted sample,

a^* = the discrimination parameter for the restricted sample,

b = the difficulty parameter for the unrestricted sample, and

b^* = the difficulty parameter for the restricted sample.

For the parameters to be comparable, either the θ or the Γ metric must be adopted. In this case, the θ metric (which is standard in the unrestricted population) is preferred. The relationship between θ and Γ (where both are standard in their appropriate samples) must therefore be:

$$\Gamma = (\theta - \bar{\theta}_r) \,/\, S_{\theta_r}. \qquad (5)$$

Substituting Equation 5 into Equation 4 and rearranging terms, the θ-metric parameter equivalents on the Γ-sample parameters are:

$$a = a^*\,(S_u \,/\, S_r) \qquad (6)$$

and

$$b = \bar{\theta}_u + [\, b^*\,(S_r \,/\, S_u)\,] \qquad (7)$$

$$= [\, \bar{r} - \bar{u} + b^*\,(S_r)\,] \,/\, S_u.$$

APPENDIX C

ITEM-POOL INFORMATION FUNCTIONS

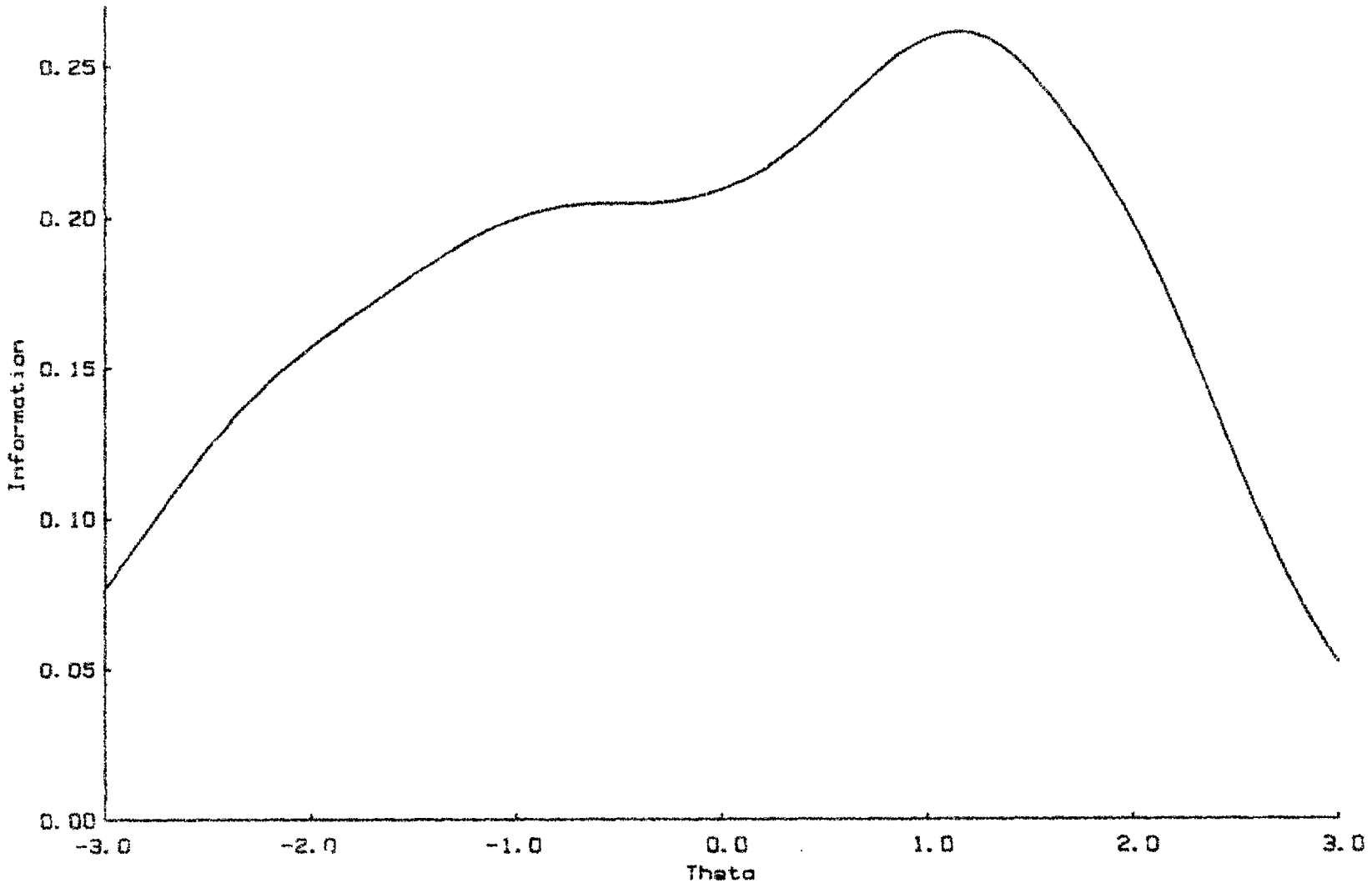

Figure C-1. Average Item Information Function for General Science Items Based on the Joint-Calibration Parameters.

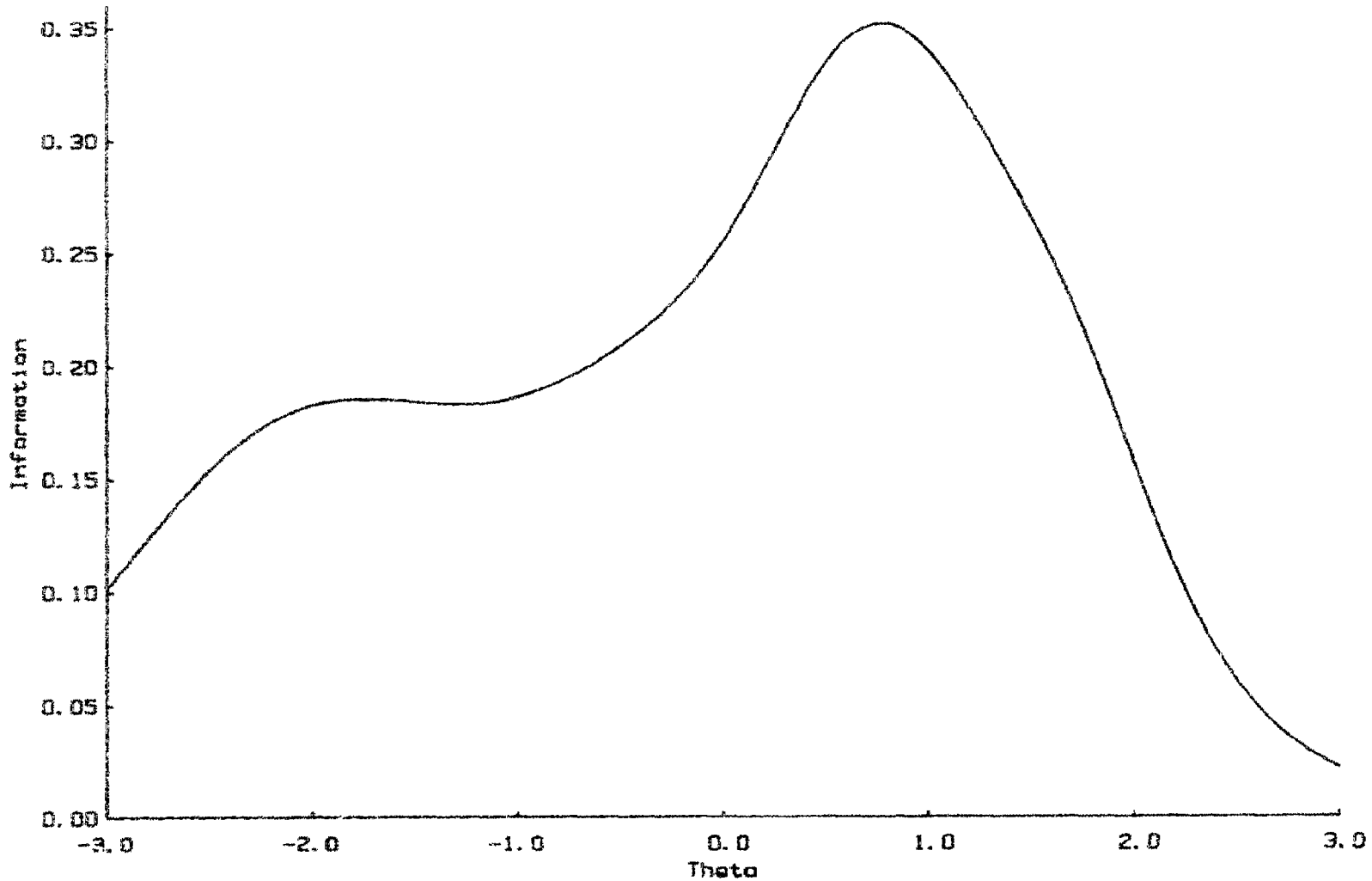

Figure C-2. Average Item Information Function for Arithmetic Reasoning Items Based on the Joint-Calibration Parameters.

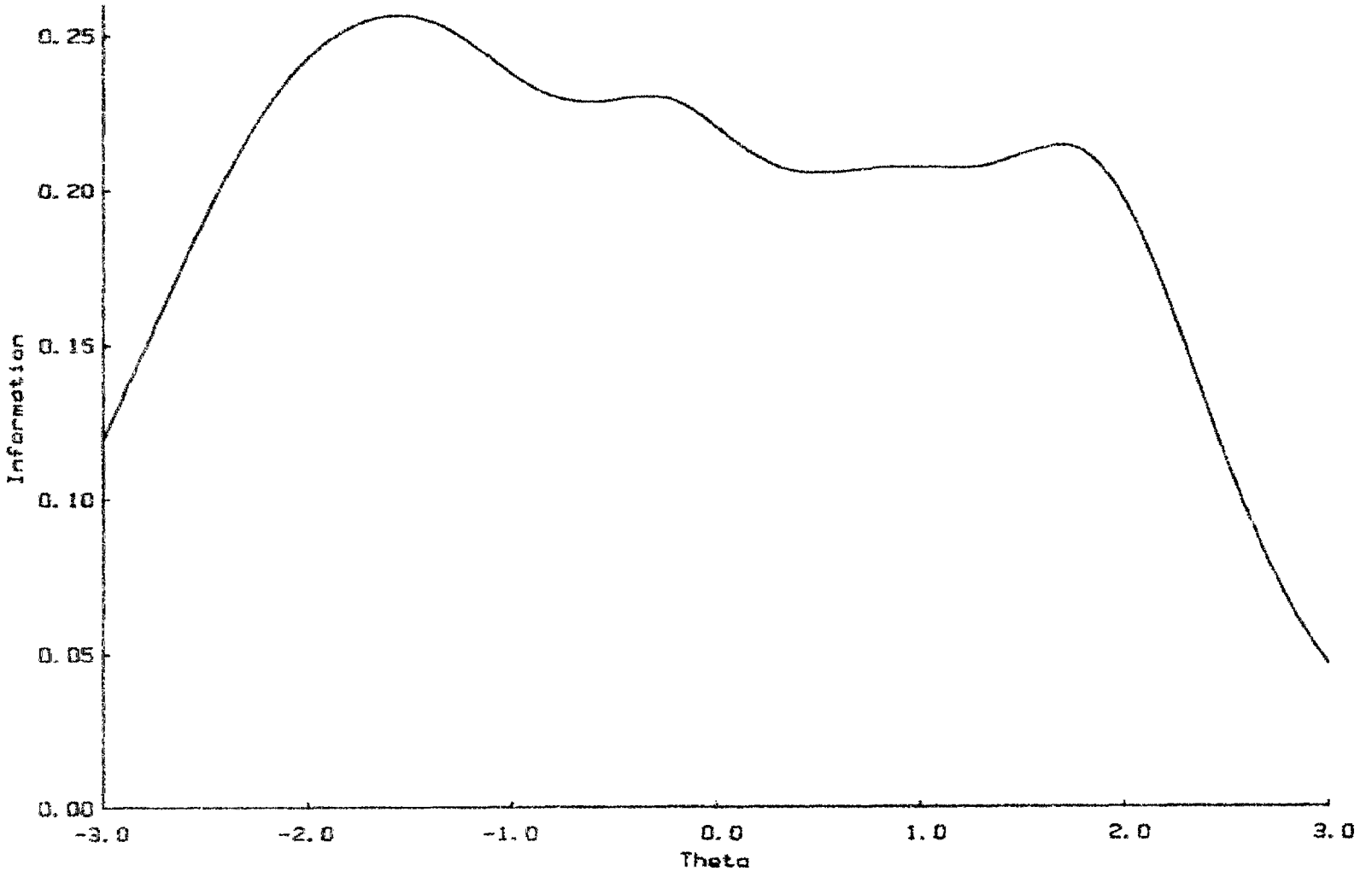

Figure C-3. Average Item Information Function for Word Knowledge Items Based on the Joint-Calibration Parameters.

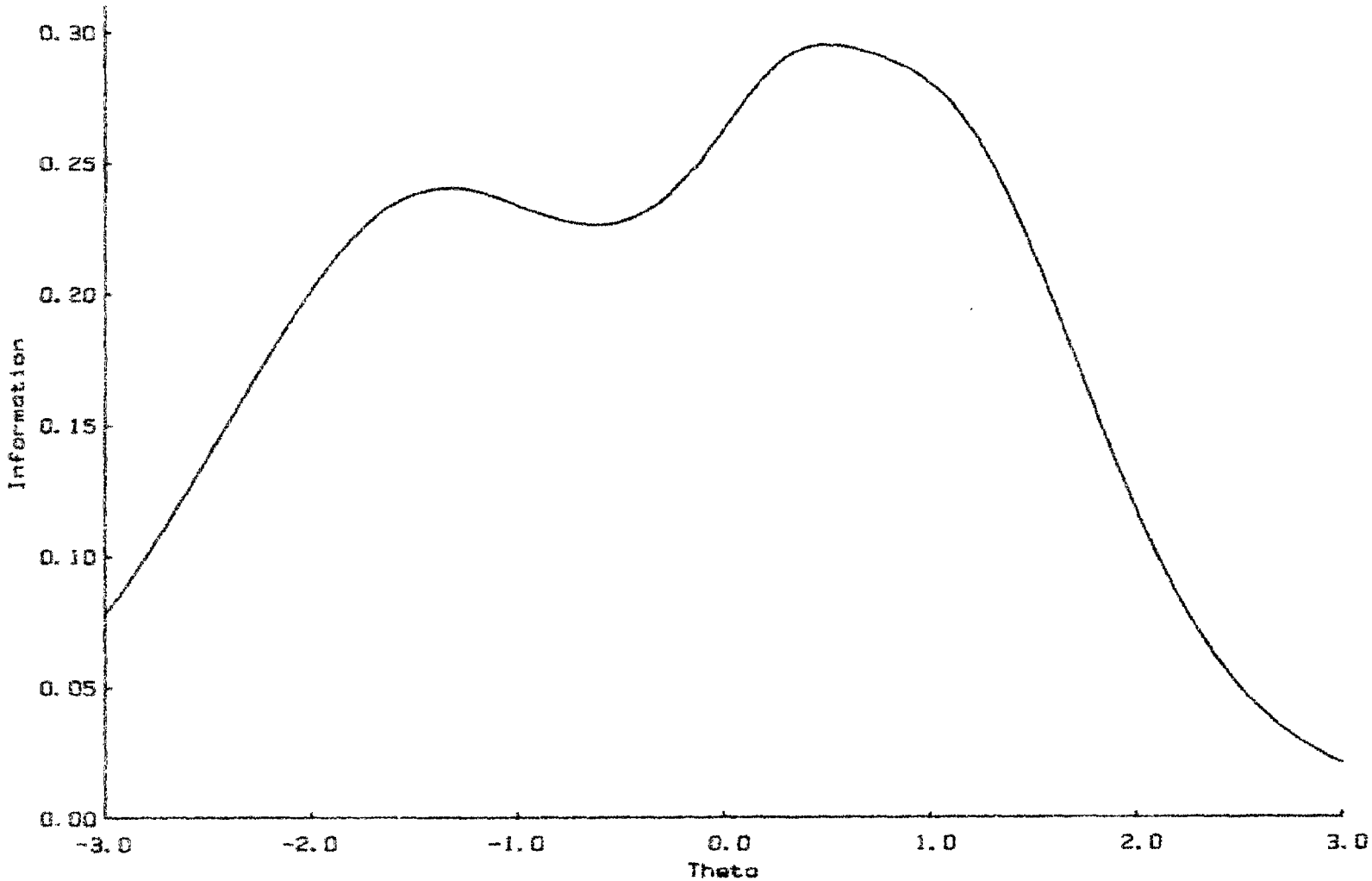

Figure C-4. Average Item Information Function for Paragraph Comprehension Items Based on the Joint-Calibration Parameters.

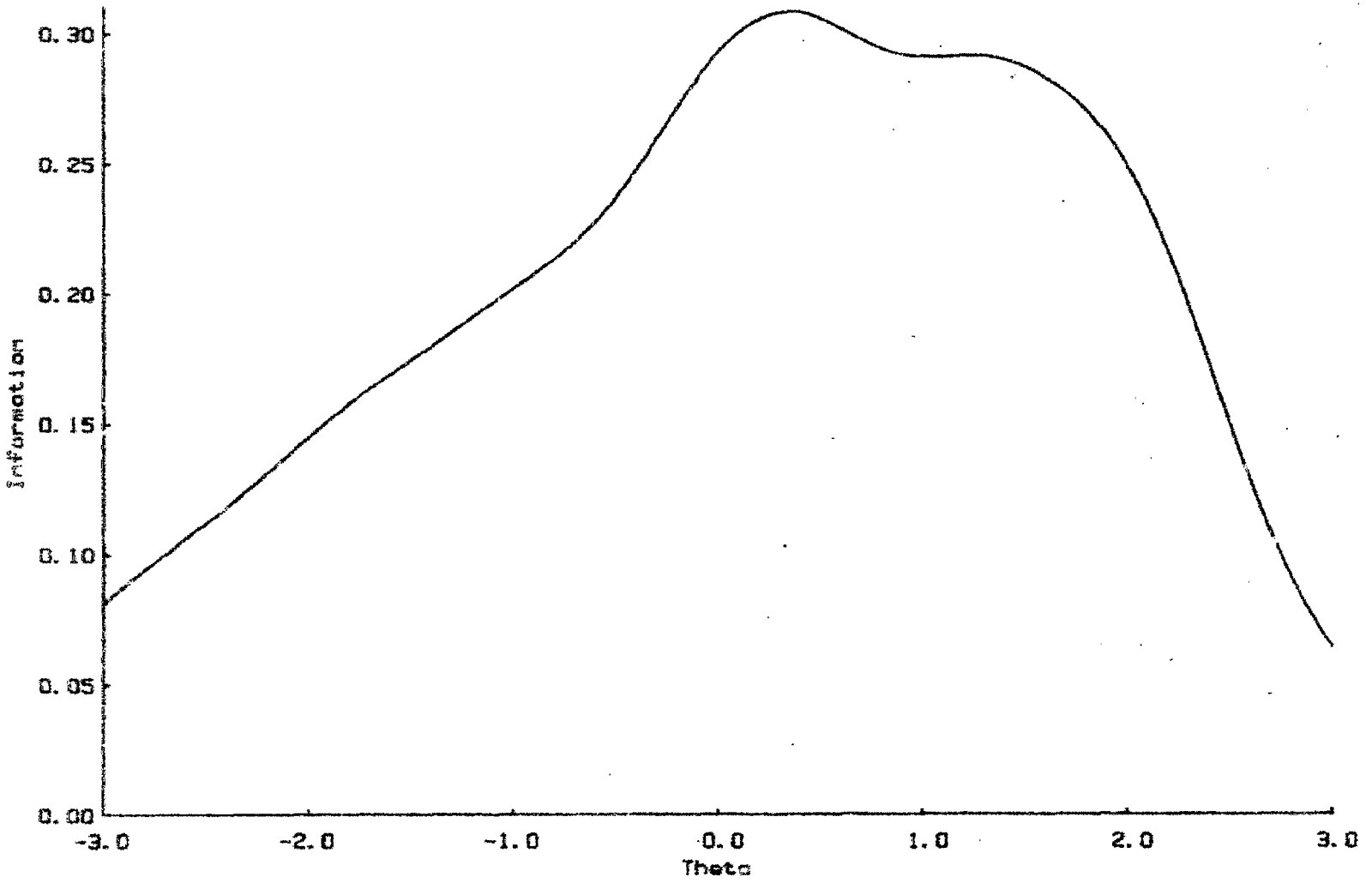

Figure C-5. Average Item Information Function for Auto Information Items Based on the Joint-Calibration Parameters.

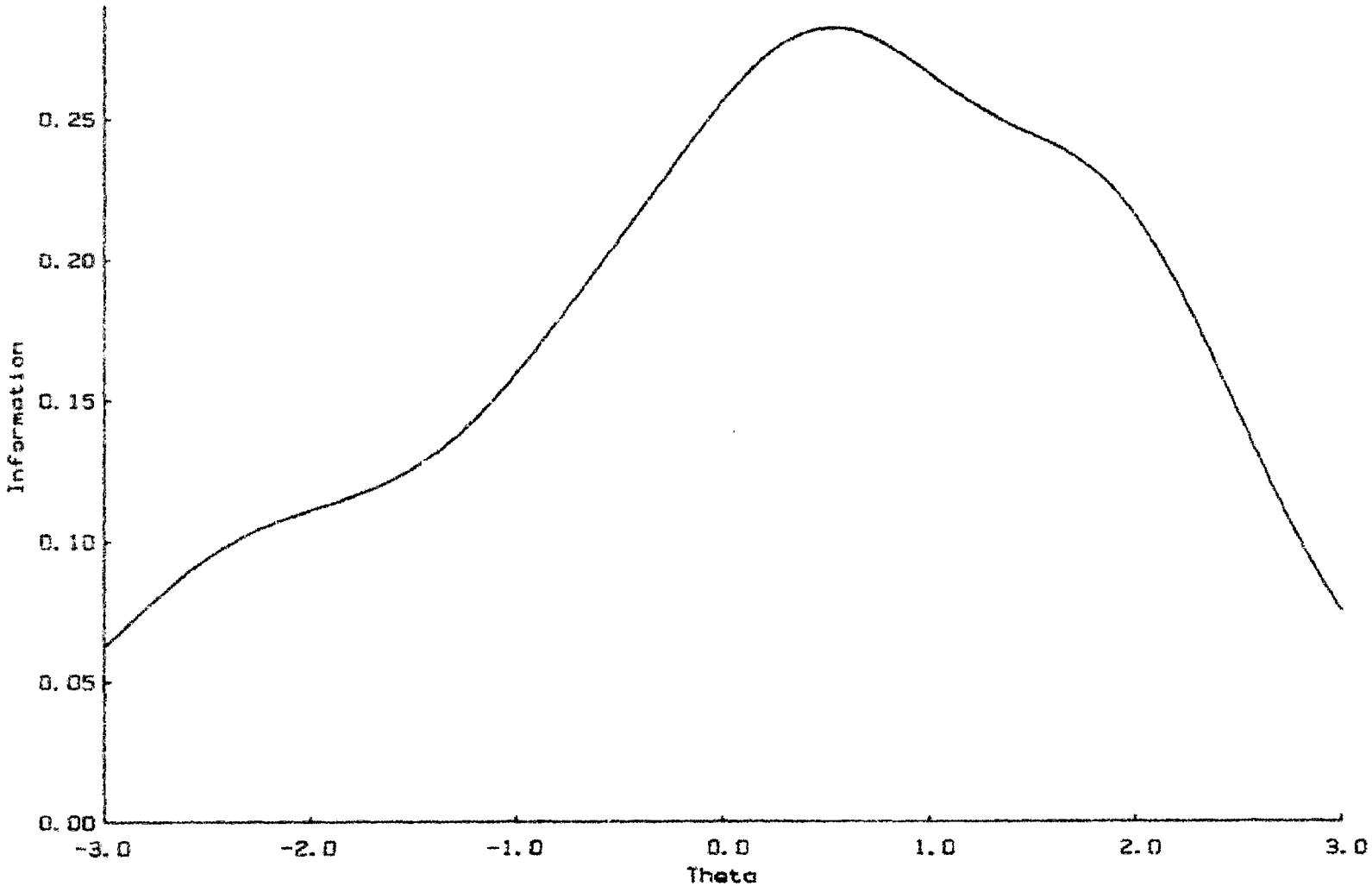

Figure C-6. Average Item Information Function for Shop Information Items Based on the Joint-Calibration Parameters.

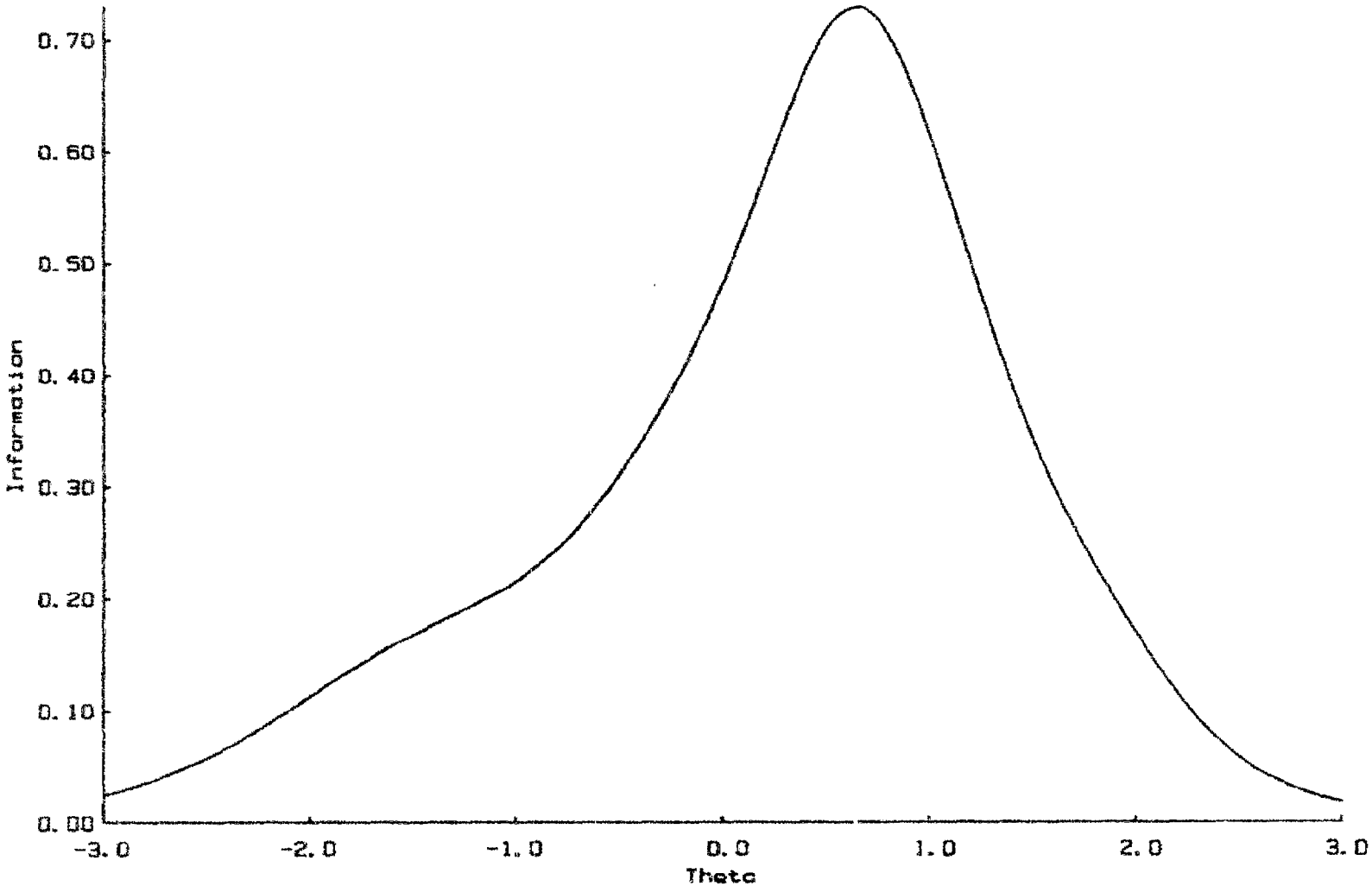

Figure C-7. Average Item Information Function for Mathematics Knowledge Items Based on the Joint-Calibration Parameters.

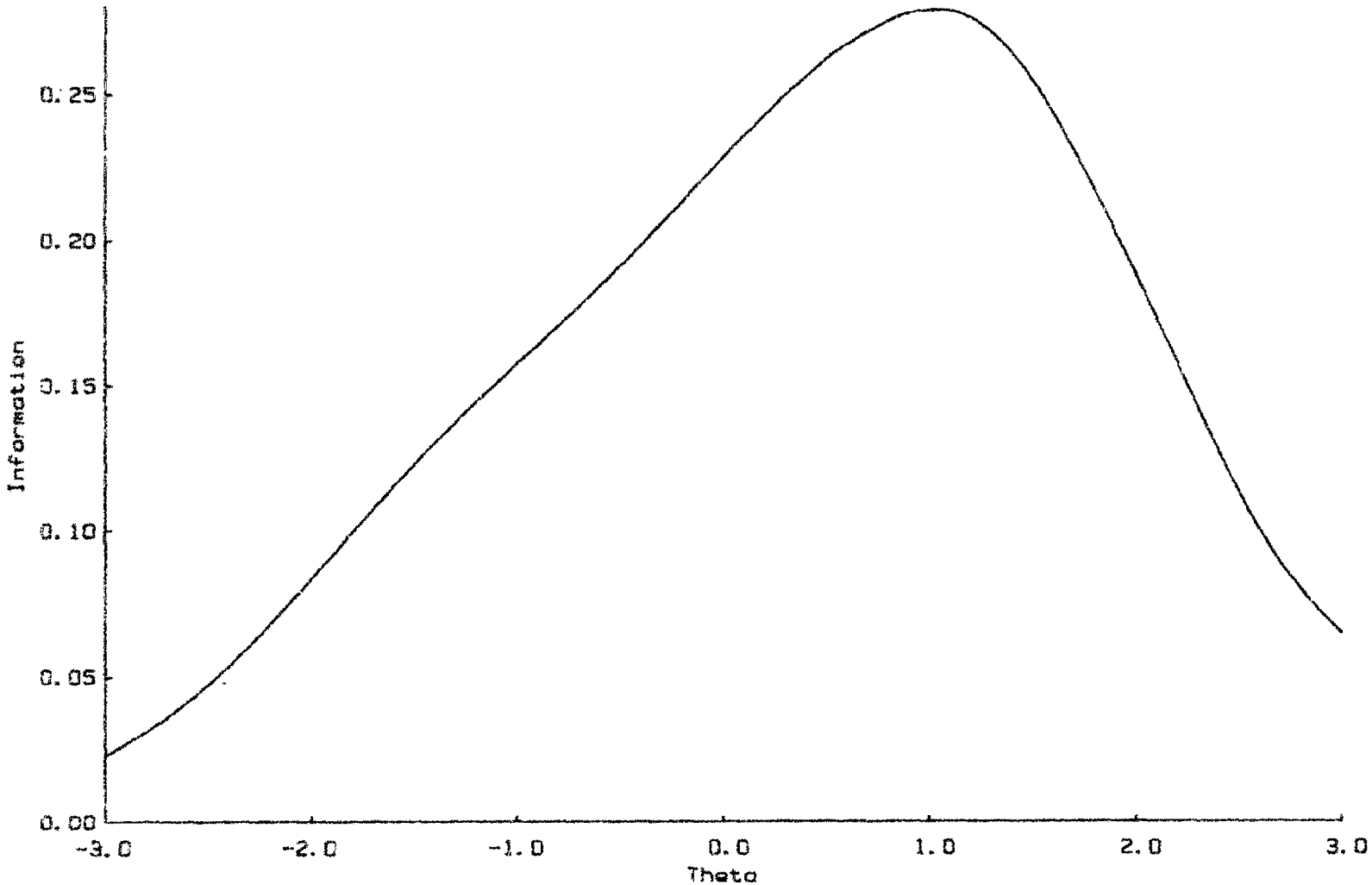

Figure C-8. Average Item Information Function for Mechanical Comprehension Items Based on the Joint-Calibration Parameters.

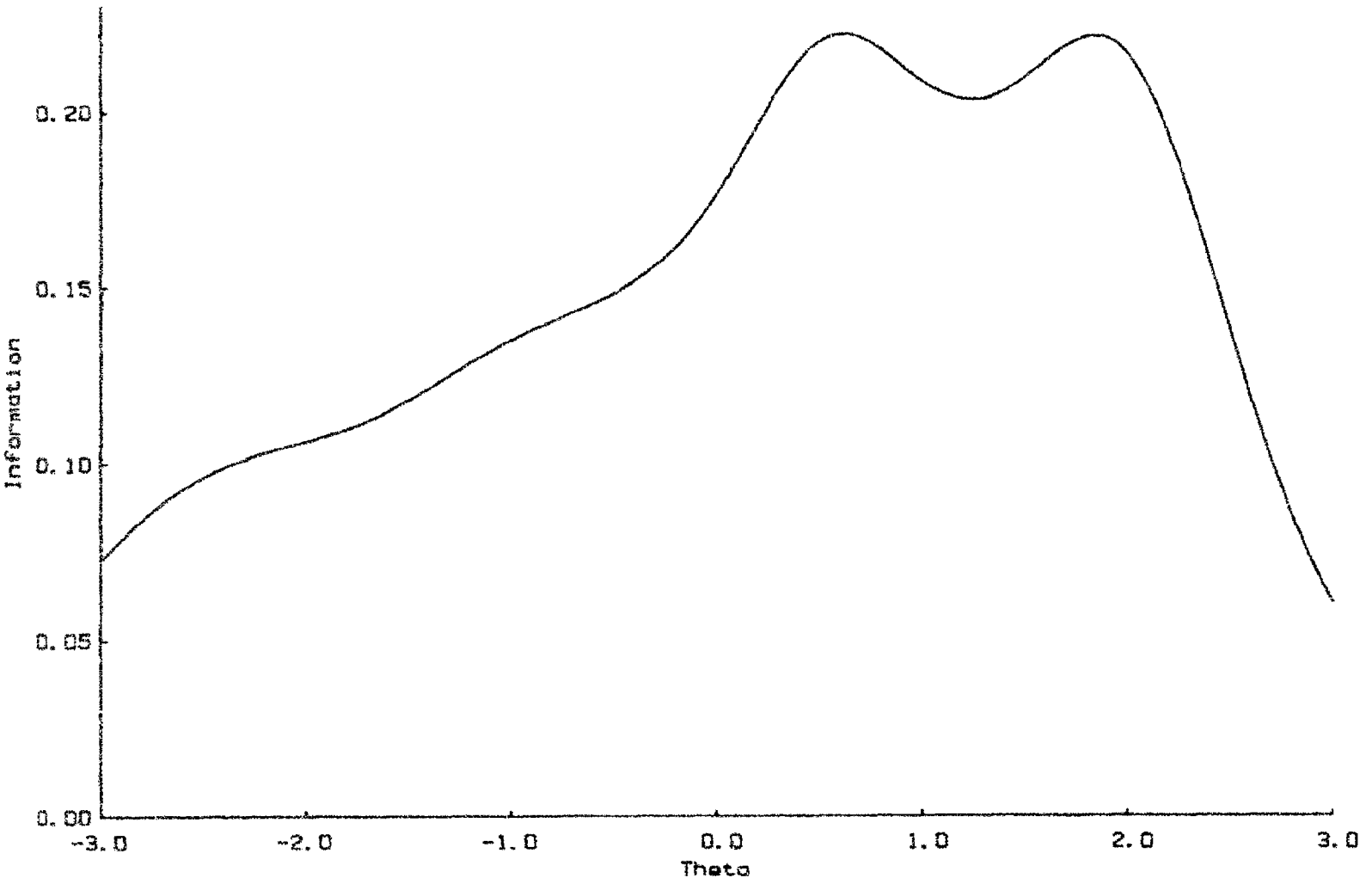

Figure C-9. Average Item Information Function for Electronics Information Items Based on the Joint-Calibration Parameters.

☆U.S. GOVERNMENT PRINTING OFFICE: 1985 569 053 20047