The Computerized Adaptive Testing System Development Project

James R. McBride and J. B. Sympson
Navy Personnel Research and Development Center

All four Armed Services use ASVAB (the Armed Services Vocational Aptitude Battery) to assess potential enlisted personnel, and to make decisions regarding personnel selection and classification. The Computerized Adaptive Testing (CAT) project is a joint-service coordinated effort to develop and evaluate a system for automated, adaptive administration of the ASVAB. If approved, the CAT system will be used by the Military Entrance Processing Command (MEPCOM) to conduct operational testing of applicants for military enlistment. Navy Personnel Research and Development Center (NPRDC) is the lead service laboratory for this research and development effort.

Project Background

A Description of CAT

CAT is a system for administering personnel tests. It differs from conventional test administration in two major respects:

1. Automated test administration using a computer terminal, rather than printed booklets and answer sheets;

2. Adaptive (tailored) sequencing of test questions, rather than the lock-step sequencing inherent in printed tests.

Automated test administration. In a CAT system, test questions are displayed on the video screen of a computer terminal. The examinee answers each question using an input device such as a keypad or a light pen. The test is automatically scored by a computer program.

Although a conventional computer terminal could be used for CAT, specially designed test administration terminals are preferable for several reasons, including: (1) the need for graphics display capabilities, since many questions include pictures and drawings; and (2) the fact that most examinees are unfamiliar with the use of computer terminals and typewriter-style keyboards.

Adaptive question sequencing. Examinees taking a conventional, printed test all answer the same questions, printed in the same sequence. Usually, easy questions are printed first and the more difficult ones are printed last. This lock-step sequencing wastes considerable amounts of time, because little useful

information is gained by asking highly able examinees easy questions, or by challenging the less able examinees with questions far too difficult for them. From an aptitude measurement point of view, it is far more efficient to tailor the choice of test questions to the ability level of each individual examinee.

CAT does this. CAT makes successive approximations to the examinee's ability, updating the approximation after each question is answered. Each question in the test sequence is selected from a large bank of pre-calibrated questions to match the difficulty of the next question to the estimated ability of the examinee. This process of sequentially tailoring test difficulty to examinee ability is the essence of adaptive testing.

The current state-of-the-art in psychometrics will support the application of adaptive testing only to what are called "power" tests of cognitive aptitude; i.e., tests in which speed of the response is not a significant factor. Forms 8, 9, and 10 of ASVAB each contain eight power tests. Two other subtests (Numerical Operations and Coding Speed) are highly speeded; these speeded tests can be automated, but cannot presently be made adaptive.

Advantages of a CAT System

CAT has a number of potential advantages over conventional printed testing. Some of its advantages are due to automation; others are attributable to the adaptive nature of CAT.

1. Efficiency: CAT can reduce the length of many ASVAB tests by as much as 50%, without loss of measurement precision. This is a direct result of the adaptive sequencing of test questions.

2. Precision: ASVAB scores are precise only at mid-range ability levels. At the low and high extremes of ability, ASVAB is inherently imprecise. Because CAT matches test difficulty to the examinee's ability level, CAT scores will be substantially more precise than ASVAB in the extremes, and just as precise in the mid-range.

3. Accuracy: In the past, ASVAB tests were often manually scored. The raw test scores were transformed to standard scores, and then combined into aptitude-area composite scores, also manually. Finally, raw scores and composite scores were typed onto enlistment processing forms by clerks. All of these manual operations are susceptible to clerical errors, resulting in inaccurate data in the personnel record.

4. Security: Any printed or conventional test is susceptible to compromise. The probability of compromise increases directly with the frequency of administration and the duration of the operational life of a specific version of the test. In the past, ASVAB security violations have occurred; future compromise attempts seem inevitable as long as there is a significant incentive for applicants to perform well on military selection tests. CAT will eliminate the two major features that make printed ASVAB susceptible to compromise: pilferable test booklets, and predictable sequences of test questions. When implemented, the CAT system will also incorporate several

other features that will decrease the probability of test compromise.

5. Economy: The initial cost of creating the CAT system will be offset by substantial cost savings in several categories as the printed ASVAB is phased out. The cost of test administrators' labor should be significantly reduced, since CAT will take half as long as ASVAB to administer. Public burden costs (civilian test-takers' time) will be similarly reduced. Printing, distribution, and storage costs associated with the paper-and-pencil tests will be eliminated. And the cost of developing new forms of the tests will be substantially reduced.

6. Ease of Revision: Developing replacement forms of ASVAB has taken three to five years in the past, and has been very expensive and logistically cumbersome, due to the need for large-scale administration of experimental tests. In contrast, every administration of a CAT test can incorporate a small number of experimental test questions embedded in the operational test. These experimental questions will be unobtrusive, and will not interfere with the operational testing routine. This process will result in frequent periodic updating of the CAT item banks; i.e., revision of the tests will occur constantly, unobtrusively, and rapidly.

Psychometric Development of CAT

Psychometric methods and procedures to be employed in the CAT hardware/software system are being developed through a combination of contract and in-house research. Development is underway in the following areas:

1. Constructing calibrated item banks for the CAT system;
2. On-line calibration research;
3. Equating CAT with ASVAB subtests and service composites;
4. Validation of CAT as a measurement technique;
5. Evaluation of CAT's utility for predicting performance;
6. Meeting established professional standards for tests.

Constructing Calibrated Item Banks

Successful implementation of any CAT system requires the development of carefully calibrated item banks. Here, the term "calibration" refers to data analysis procedures that provide estimates of item response theory (IRT) parameters for each test question. IRT item parameters serve to describe the operating characteristics of test questions that are used to measure ability. The choice of which item to administer to an examinee at any point in an adaptive test is determined by a complex mathematical function of these parameters.

In connection with the joint-services CAT project, two major efforts have been undertaken in the area of item calibration. These may be identified as (1)

experimental CAT system item bank development and (2) operational CAT system item bank development.

Experimental item bank development. In order to evaluate the reliability, construct validity, and criterion-related validity of CAT prior to its operational implementation, NPRDC has created an experimental CAT system. This small experimental system uses seven Apple III microcomputers, all connected to a Corvus hard-disk unit. The disk unit serves as a storage medium for the item banks, item parameters, and instruction files that are used by the experimental system. It also stores the data that are generated in connection with each examinee's testing session. At this time, development of the item banks needed for this experimental system is partially complete. Information about the calibrations that have been completed so far is presented in the paper following this one.

Operational item bank development. In order to have an adequate number of high-quality test questions available at the time that CAT is likely to become operational, Air Force Human Resources Laboratory (AFHRL) has contracted with Assessment Systems Corporation of St. Paul, MN, to develop and calibrate additional questions in the content areas associated with ASVAB "power" subtests. These questions, and possibly some of the better questions from the experimental item banks, will be used in the initial item banks for the operational CAT system. A target of 200 questions per adaptive-test item bank has been established.

On-Line Calibration Research

The two item-calibration efforts just described involve the calibration of test questions that have been administered in printed test booklets. There may be some risk in using item parameters obtained under this type of testing condition in the operational CAT system, since the medium for item presentation and examinee response will be quite different. An obvious solution for this potential problem would be to collect item calibration data on computer terminals. Unfortunately, item banks and subject samples as large as those required for the operational CAT system make this currently impossible. However, once the operational CAT system is in place, it will be a simple matter to insert two or three newly developed items into each CAT subtest that an examinee completes. Data collected in this manner (i.e., "on-line") could be used to calibrate new items very rapidly. Moreover, if research does indicate that the medium for test delivery has an influence on obtained parameter estimates, items appearing in the initial operational item banks could themselves be re-calibrated on-line.

At this time, well-developed statistical procedures for on-line item calibration do not exist. Some straightforward generalizations of current procedures could be implemented, but they probably would not be optimal. In order to foster the development of efficient on-line item calibration methods, and to make it likely that these methods will be available by the time that CAT becomes operational, the Office of Naval Research (ONR), NPRDC, AFHRL, and MEPCOM are planning to co-fund an ONR research contract in which several IRT researchers address the issues involved in on-line calibration. This research effort should result in one or more procedures that can be implemented in the operational CAT system.

Equating CAT with ASVAB

For military personnel selection and classification purposes, ASVAB subtest scores are combined into aptitude-area composites, such as the Mechanical, Administrative, General, and Electronics (M, A, G, and E) composites used by the Air Force. Each of the Services has its own composites and establishes composite qualifying scores for entry into specific occupational specialities. When CAT becomes operational, it will be essential to the continuity of the four Services' selection and assignment practices that CAT test scores be interchangeable with those of ASVAB. This will require that procedures be developed to "equate" CAT with ASVAB, i.e., to transform CAT scores onto the familiar ASVAB score scale so that a given score on a CAT test will have the same interpretive meaning and be useful for the same purposes as actual ASVAB scores. A committee of psychometric experts has been commissioned, with joint ONR/NPRDC funding, to develop methods for equating CAT to ASVAB.

Validating CAT as a Measurement Technique.

CAT is intended to measure the same aptitudes and abilities that ASVAB now measures. However, CAT is very different from ASVAB in its mode of administration and in other obvious respects. These obvious differences raise the possibility that CAT may measure something rather different from one or more of its ASVAB subtest counterparts. Consequently, it is necessary to conduct psychometric research to establish the construct validity of the CAT subtests. This research will involve administration of CAT and ASVAB tests to experimental groups of examinees, followed by statistical analyses to investigate the extent to which CAT and ASVAB measure the same aptitudes. The experimental CAT battery described above is currently being administered to military recruits in order to collect the data required for an assessment of CAT's construct validity.

Evaluating the Predictive Utility of CAT

ASVAB is useful as a tool for selection and classification of enlisted personnel by virtue of the known correlation of ASVAB subtest and composite scores with performance in occupational specialty training. Although there is every reason to think that CAT will be as useful as ASVAB for predicting training performance, it is still necessary to demonstrate CAT's predictive utility. To accomplish this, it is necessary to test recruits using CAT, and to conduct followup studies to determine the correlation between CAT test scores and training performance. Although it is not practical to do this for every occupational specialty in the four Services, it is essential to demonstrate that CAT has predictive utility across a broad spectrum of job types in each of the Services. Each of the four Services has been requested to identify six technical training courses to be involved in the demonstration of CAT's predictive utility. Recruits designated for subsequent assignment to those courses will be tested using the experimental CAT battery, and will be tracked through training in order to collect the criterion data required to evaluate CAT's predictive relationships with training performance data.

Meeting Established Professional Standards

Aptitude tests that are well designed and well developed can be enormously valuable to the institution that uses them, while at the same time being fair and equitable to examinees. Over the years, professional standards for the development and use of tests have been established. The best known of these are the "Standards for Educational and Psychological Tests," jointly published by the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education. These standards are currently undergoing revision, with the revised standards due to be published in 1984.

It is essential that the CAT system comply with the highest professional standards for test use, in order to insure CAT's credibility as a valid, useful, and fair instrument for personnel assessment and selection. Meeting such standards will require that a variety of psychometric research studies be conducted and documented. The preferred method of documentation is the publication of a professional "test manual," which will contain all of the results of the research into CAT's psychometric characteristics and benefits.

Under an ONR contract, Bert Green of Johns Hopkins University has studied the current ASVAB test battery and the proposed CAT system, and has developed recommendations for evaluation of the CAT system; he was assisted in his efforts by psychometricians Darrell Bock, Lloyd Humphreys, Robert Linn, and Mark Reckase. The evaluation plan developed by this panel supplements the current APA/AERA/NCME Standards, for the special case of the CAT system. NPRDC is currently designing and conducting research to address the major points presented in the evaluation plan proposed by Green et al., and intends to report the results of this research in a test manual for the CAT system.

Criteria for CAT Delivery System Evaluation

The CAT system will become operational only after a milestone decision is made to implement it. That decision cannot be affirmative unless the CAT hardware/software system meets a number of criteria. Since the CAT system is being developed in several stages, intermediate evaluations can be conducted during the system development process, well in advance of the final evaluation that precedes the decision whether to implement the system. The following eight major criteria will be used in evaluating the CAT delivery system.

1. Performance. The CAT system will be evaluated as a computer system with a number of critical performance parameters. Included among these are:

   a. Speed of response: The system should respond to examinee input in less than 2 seconds. This is necessary in order to avoid distracting the examinee, and to ensure that the efficiency of adaptive testing will be reflected in reduced test administration time compared with ASVAB.

   b. Speed of display: Once the system has responded to examinee input, the next test question or dialogue frame should be completely dis-

played in less than 3 seconds, for the same reasons.

c. Display resolution: The display must provide clear, unambiguous presentation of both test and pictorial material. Alphanumerics should have a dot resolution of at least 7 × 9 for ease of reading. Pictorial material should have a resolution of at least 400 horizontal by 300 vertical picture elements for clarity.

d. Mass storage capacity: The system must provide adequate mass storage to contain system and applications programs, the large item banks used for adaptive testing, and archival records of each test administration and the test results. It should also have excess capacity, in order to be capable of growth in the number of subtests and in the size of the item banks.

e. Communications capability: The CAT system will be required to transmit large volumes of test results to a central computer site on a daily basis, and to receive software revisions and item bank updates periodically, using electronic data communications. In order to accomplish these functions economically and without interrupting the operational schedule of test administration, the system must be capable of high-speed data communications.

2. Suitability. The CAT system is intended for use in both Military Entrance Processing Stations (MEPS) and Mobile Examining Team sites (METs) without significant changes in the staffing or facilities currently available. This implies that CAT must be capable of operation by personnel of normal skill levels, without need for extensive specialized training. It must also be capable of operation in the normal office environment that characterizes MEPS and METs, without requiring special environmental controls on temperature and humidity. It should require a minimum of facility modifications. It should be portable if portability is required to serve MET sites.

3. Reliability and availability. The system must be available for testing whenever testing is scheduled; once testing has begun, all examinees must complete their testing. Little deviation from these requirements can be tolerated, due to the importance of timely completion of enlistment processing. Quantitative thresholds of 99.9% for both availability and reliability have been established as design goals. 99.9% availability means that a CAT installation would be unavailable for testing less than one day per 1,000 scheduled testing days, or about once every four years. 99.9% reliability means that less than 1 test per 1,000 would be interrupted by a system failure; redundancy features of the system will provide the capability to complete the few tests which are interrupted in this way.

4. Maintainability. The system must be designed so that there is no requirement for skilled technicians in the MEPS or METs. It should include built-in diagnostic tests to alert the test administrator of present or impending hardware or software failure. The system design should include integrated logistics support.

5. Ease of use. Because it will be used by examinees and test administrators with no computer experience, the system must be designed explicitly for ease of use. For the examinee, this means the system must use a very simple procedure for answering test questions, and must first teach the examinee how to use it. For the test administrator, the system must involve only a small number of simple operations, and must direct the test administrator as to what to do at each step in its use.

6. Security. The system must incorporate protection against unauthorized access to the CAT item banks, and to examinee records. It must make it impossible to make printed copies of test questions, and must contain no small pilferable articles whose loss could result in test compromise. It must make coaching ineffective as a means of test compromise, by eliminating predictable sequences of test questions. It must include a password access system, which not only prevents unauthorized access, but also creates an audit trail of every test administered on the system.

7. Affordability. A 10-year operating life is specified for the CAT equipment. The life-cycle cost of the CAT system over its 10-year life must be competitive with that of the printed ASVAB.

8. Flexibility. Much of the long-run potential of CAT resides in its capacity to support types of assessment that are either difficult or impossible to implement under paper-and-pencil modes of test administration. Some of these possibilities include answer-until-correct responding, open-ended responding, the use of dynamic task stimuli, and the assessment of perceptual and psychomotor skills. The new CAT system should have software/hardware capabilities that are sufficiently flexible to allow exploration and evaluation of such assessment capabilities.

Conclusion

Assuming the CAT system now under development is found to satisfy the many engineering and psychometric criteria that have been established as conditions for its acceptance, the phased implementation of CAT on a nationwide basis could begin as early as 1986. By 1990, over one million applicants for military service would be tested on the CAT system each year. This fact would undoubtedly provide a strong impetus for the application of computerized adaptive testing procedures in other areas of personnel selection and classification as well.