



3

Development of the CAT-ASVAB

Daniel O. Segall

Kathleen E. Moreno

*Defense Manpower Data Center
United States Department of Defense*

The Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) is one of the most thoroughly researched tests of human proficiencies in modern history. Data from over 400,000 test-takers collected over a 20 year period have been used to address crucial research and development issues. In spite of its lengthy and thorough development cycle, CAT-ASVAB was the first large-scale adaptive battery to be administered in a high-stakes setting, influencing the qualification status of applicants for the U.S. Armed Forces. This chapter outlines the development of the battery, and describes some of the challenges and dilemmas faced in constructing a CAT version that is interchangeable with its preexisting paper-and-pencil (P&P) counterpart.

Since 1976, the P&P-ASVAB has been used for selection into the Military, and classification into a number of occupational specialties. Because applicants tend to have little employment history, and because large numbers apply each year, scores on the ASVAB play an important role in determining an applicant's qualification status. The ASVAB has proven to be a good predictor of future training success. Because it can be administered to large groups of applicants and takes about 3 hours, it has proven to be an efficient means of obtaining information for predicting future success. In 1990 about one million applicants took the ASVAB. In recent years, the annual administration rate has dropped to about 600,000. The numbers fluctuate from year-to-year, depending on the economy, and on Military manpower needs.

In the years prior to 1976, the Army, Air Force, Navy, and Marine Corps each administered unique classification batteries to their respective applicants. Beginning in 1976, a joint-Service ASVAB was administered to all

Chapter appears in F. Drasgow & J. B. Olson-Buchanan (Eds.). (1999). *Innovations in Computerized Assessment* (pp. 35–65). Hillsdale, NJ: Lawrence Erlbaum Associates.

Military applicants. The battery was formed primarily from a collection of Service specific tests. The use of a common battery among Services facilitated manpower management, standardized reporting on accession quality to Congress, and enabled applicants to shop among the Services without taking several test batteries.

Virtually from its inception, the ASVAB was believed susceptible to compromise and coaching (Maier, 1993). Historically, the ASVAB program has offered continuous on-demand scheduling opportunities, with nearly 1000 testing sites located in geographically disperse areas. Both applicants and recruiters have strong incentives to exchange information on operational test questions. High scoring applicants can qualify for service, enlistment bonuses, educational benefits, and desirable job assignments. Performance standards for recruiters are based on the number of high-scoring applicants they enlist. Around the time of the ASVAB implementation in 1976, additional compromise pressures were brought to bear by the difficulty Services had in meeting their goals in the all-volunteer service of the post-Vietnam era. In fact, Congressional hearings were held to explore alternative solutions to ASVAB compromise. Although other solutions were identified and later implemented (i.e. the introduction of additional test forms), one solution proposed during this era was implementation of a computerized adaptive testing version of the ASVAB. The computerization of test questions was believed to make them less prone to physical loss than P&P test booklets. Additionally, the adaptive nature of the tests was believed to make sharing item content among test-takers and recruiters less profitable, since applicants receive items tailored to their specific ability level.

3.1 EARLY RESEARCH AND DEVELOPMENT

Just one year after the implementation of the ASVAB, concerns over test compromise and other practical issues led to the Marine Corps Exploratory Development Project (McBride, 1997). The purpose of the project was to answer practical questions related to computerized adaptive testing. First, could an adaptive-testing delivery system suitable for military personnel tests be developed? At the time, two classes of computers existed: mainframes and minicomputers. Mainframes were expensive to purchase, operate, and maintain. Previous attempts with CAT using mainframes were largely unsuccessful because of the unpredictable response times associated with time-sharing (Weiss, 1975). Similar concerns existed about minicomputers (which bear little resemblance to the personal computers of today). Minicomputers cost between \$50,000 and \$100,000, and there was some concern that they would not be powerful enough to handle the computations necessary to support multiple users in an adaptive testing environment.

A second question addressed the correspondence between empirical re-

sults (from live subjects) and those predicted on the basis of theoretical and simulation analyses. Specifically, would advantages claimed on the basis of theoretical studies be confirmed by empirical data obtained from military recruits? Although a great deal of theoretical work was suggestive of the relative advantages of CAT over P&P, empirical validation of these findings was largely absent from the testing literature.

These early practical questions were addressed by two studies. Both studies used CAT algorithms based on the three parameter logistic model (3PL) and Owen's (1969, 1975) Bayesian sequential procedure for item selection and scoring. The tests were administered on remote terminals controlled by a time-shared minicomputer system. The first study (McBride & Martin, 1983) compared the relative efficiency of experimental verbal tests administered in adaptive and conventional modes by computer. Data from 466 Marine Corps recruits were gathered and used to compute reliability and construct validity coefficients at different test lengths. The results corroborated the theoretical advantage of superior CAT efficiency manifested by reduced test lengths required to achieve a desired level of precision.

The second study (Moreno, Wetzel, McBride, & Weiss, 1984) expanded the investigation from one to three content areas: Word Knowledge, Paragraph Comprehension, and Arithmetic Reasoning¹. Data from 356 Marine Corps recruits were gathered on three versions of each test: (1) a CAT version, (2) an operational P&P version taken prior to enlistment, and (3) a second P&P version taken after enlistment. For each test, the CAT version administered roughly half the items as were contained in the P&P versions. Results were consistent with prior beliefs that a shorter CAT could measure the same constructs as a P&P test, with equivalent or higher precision.

Taken together, these two studies corroborated results between the empirical findings and those predicted from theory concerning CAT's increased efficiency. However, the suitability of computer hardware for the purpose of CAT was still questionable. Each minicomputer used in the study was only able to support a small number of users without distracting response time delays. Still other deficiencies, such as the inability to display graphics items and easily move computers between test sites made them unsuitable for large-scale operational use.

Part way through the Marine Corps Exploratory Development Project, the Department of Defense initiated a Joint Service Project for development and further evaluation of the feasibility of implementing CAT (Martin & Hoshaw, 1997). A tasking memo was cosigned on 5 January 1979 by the Under Secretary of Defense for Research and Engineering, later Secretary of Defense, William J. Pery. By this time, there was a strong interest in

¹These content areas are included in the Armed Forces Qualification Test (AFQT) selection composite used by the services to qualify applicants for admission into the military.

CAT among the Services as a potential solution to several testing problems. This enthusiasm was partly generated by the possibility of addressing test-security concerns and partly by a litany of other possible benefits over P&P. These potential benefits included: shorter tests, greater precision, flexible start/stop times, online calibration, the possibility of administering new types of tests, standardized test administration (instructions/time-limits), and reduced scoring errors (from hand or scanner scoring).

From the outset, the Joint Service CAT-ASVAB project had an ambitious and optimistic research and development schedule. Because of this compressed timeline, the effort was split into two parallel projects: (1) contractor delivery system development (hardware and software to administer CAT-ASVAB), and (2) psychometric development and evaluation of CAT-ASVAB. In 1979 micro-computing was in its infancy; no off-the-shelf system was capable of meeting the needs of CAT-ASVAB, including portability, high fidelity graphics, and fast processing capability to avoid distracting delays to test-taker input. Several contractors competed for the opportunity to develop the delivery system, and by 1984 three contractors had developed prototypes that met all critical needs. By this time however the microcomputer industry had advanced to the point where off-the-shelf equipment was less expensive and more suitable for CAT-ASVAB use. Consequently, the contractor delivery system was abandoned, and off-the-shelf computers were selected as a platform for CAT-ASVAB. Meanwhile, during the period 1979-1984, psychometric evaluation proceeded apace, with the development and validation of an experimental CAT-ASVAB version.

3.1.1 The Experimental CAT-ASVAB System

The experimental CAT-ASVAB system was developed to collect empirical data for studying the adequacy of proposed adaptive testing algorithms and test development procedures. The intent was to develop a full battery CAT version that measured the same dimensions as the P&P-ASVAB that could be administered in experimental settings. Several substantial efforts were required to construct the system, including psychometric development, item pool development, and delivery system development.

Psychometric procedures (item selection, scoring, and item pool development) of the experimental system were based on item response theory (IRT). Earlier attempts at adaptive tests using Classical Test Theory did not appear promising (Lord, 1971; Weiss 1974). The three parameter logistic model was selected from among other alternatives (one and two parameter normal ogive and logistic models) primarily because of its mathematical tractability, and its superior accuracy in modeling response probabilities of multiple choice test questions.

By the early 1980's two promising adaptive strategies had been proposed in the testing literature, one based on maximum likelihood (ML) estimation theory (Lord, 1980), and another based on Bayesian theory (Owen, 1969,

1975; Urry, 1983). The principle difference between the procedures involves the use of prior information—the ML procedure defines estimated ability in terms of the value which maximizes the likelihood of the observed response pattern. The Bayesian procedure incorporates both the likelihood and prior information about the distribution of ability. The two procedures also differ in their characterizations of uncertainty about the true ability value, and how the potential administration of candidate items might reduce this uncertainty.

Differences between the approaches had practical advantages and disadvantages in the context of CAT. The ML item selection and scoring procedure enables the use of pre-calculated information tables to improve the speed of item selection, however provisional ability estimates required for item selection may be undefined or poorly defined early in the test (e.g. for all correct or incorrect patterns). The Owen's Bayesian item selection and scoring procedure provides adequately defined and rapidly computed provisional ability estimates (regardless of the response pattern), but computations required for item selection taxed the capabilities of available processors at the time. The net result of these differences led to the development of a hybrid method (Wetzel & McBride, 1983) which combined the strengths of both procedures. The hybrid method uses Owen's Bayesian procedure to compute provisional and final ability estimates, and based item selection on ML information tables. In a simulation study of alternative methods, Wetzel and McBride found the hybrid procedure to compare favorably to the pure ML and Owen's Bayesian procedures in terms of precision and efficiency.

Large item pools were written and calibrated for the experimental system (Wolfe, McBride, & Sympson, 1997). Over 4,000 items were written (about 450 for each of nine content areas). These items were pre-tested on samples of military recruits (providing about 300 responses per item). Items with low discrimination were removed from the pools, and the remaining items were administered in paper-and-pencil booklets to over 100,000 military applicants (providing about 1500 responses per item). IRT item parameter estimates were obtained using joint maximum likelihood procedure implemented by the computer program LOGIST (Wood, Wingersky, & Lord, 1976).

There was some concern about the calibration medium used to estimate the necessary item parameters. Specifically, would the IRT item parameters estimated from responses obtained on paper-and-pencil booklets be suitable for use of these same items administered in a computerized adaptive testing format? Given the large numbers of test-takers required, calibration of these items from computerized administration was not feasible. Some assurance concerning the suitability of P&P item parameters was given by the favorable results of other adaptive tests which had relied on P&P calibrations (McBride & Martin, 1983; Urry, 1974). A systematic treatment of this issue was conducted for the development of the operational CAT-

ASVAB forms several years later by Hetter, Segall, and Bloxom (1994) who found that the medium of item calibration has no practical impact on the psychometric properties of adaptive test scores.

While the primary hardware/software system for nationwide implementation was under development by contractors, another delivery system was constructed in-house specifically for use in low-stakes experimental research (Wolfe, McBride, & Sympton, 1997). This experimental system had many important features, including the ability to present items with graphical content, capability of rapid interaction when processing examinee input, portability, and psychometric flexibility (in terms of item selection, scoring, and time-limits). By 1982, Apple III personal computers that could meet these requirements were commercially available. The experimental system consisted of up to eight Apple computers (with 256K of random access memory) networked with a single 10-megabyte hard disk drive². The system included a modified keyboard, where all but six lettered keys (labeled A, B,C, D, E, and HELP) were covered. Additional keys were labeled “yes,” “no,” and “erase,” which served to confirm and enter responses.

The experimental software, written in PASCAL, used the hybrid item-selection and scoring strategy, fixed length adaptive tests (15 items for each power subtest except Paragraph Comprehension, which administered 10 items), and incorporated two non-adaptive speed tests. The test-lengths were about 40 to 50 percent shorter than their P&P counterparts, and were consistent with both theoretical and empirical findings that demonstrated adaptive tests to be about twice as efficient as their P&P counterparts (McBride & Martin, 1983). To deal with potential delays to examinee input, it also incorporated a look-ahead procedure. While examinees were reading test questions, the system performed the necessary item-selection calculations to determine the most appropriate follow-on question for both correct and incorrect responses to the current item. Thus when the answer was input and scored, the next question appeared almost instantly on the computer monitor.

3.1.2 Joint-Service Validity Study

From 1982-1984, the experimental CAT-ASVAB system was used in a large-scale validity study to answer a fundamental question concerning the exchangeability of CAT and P&P versions of the ASVAB (Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997). Specifically, could a short adaptive version of the ASVAB have the same validity as its longer P&P counterpart for predicting success in training? Because the prediction of training success is a central function of the ASVAB, a direct answer to this issue

²Individual computers had no internal hard-drive because of the prohibitive expense. In 1982, a 20-megabyte hard drive cost about \$5,500.

was of primary importance. Previous studies had not examined criterion related CAT validity, and only examined the construct validity of limited content areas. In addition, no empirical data were available on the performance of speeded (conventional) tests administered by computer, and their equivalence with P&P versions.

Predictor data were gathered from 7,518 recruits scheduled for training in one of 23 military occupational specialties. There was some concern that the collection of predictor data from recruits (rather than applicants) would distort the outcome of the validity analysis. First, the time interval between the collection of predictor (ASVAB) and criterion (success in training) data was somewhat compressed for recruits. Thus, maturation in predictor-abilities may tend to inflate the validity estimates. Second, the non-operational recruit testing environment is somewhat more variable and less motivating than the operational applicant testing environment; these factors tending to lower the validity estimates. However, applicant testing was not efficient or practical since only a small (unknown) portion of the applicant population would eventually be assigned to the 23 schools included in the study. To help control for the influence of these extraneous factors, recruits were tested on both CAT-ASVAB and P&P-ASVAB versions under similar experimental conditions. Consequently, three sets of predictors were available for analysis: (1) the operational P&P-ASVAB taken prior to enlistment, (2) the experimental CAT-ASVAB taken during basic training, and (3) selected P&P-ASVAB subtests also taken during basic training.

Occupational specialties were chosen to ensure that: (1) a broad spectrum of Service training programs were represented, (2) all P&P-ASVAB tests were included in school predictor composites, and (3) within-specialty sample sizes would be large enough to make meaningful validity comparisons between CAT- and P&P-ASVAB. Criteria data for study participants were collected several months later at the end of course instruction. These criteria consisted of final course grade, completion time, or a composite of mid-term test scores. After removing subjects that did not complete course training, the average sample size for each school was about 327. (Actual sample sizes ranged from a low of 69, to a high of 456.)

For each of the 23 schools, multiple correlations (R 's) were computed between criteria and optimally weighted subtests. Separate R 's were computed for each of the three predictor sets (CAT-ASVAB, operational P&P-ASVAB, and non-operational P&P-ASVAB.) Multiple correlations based on optimally weighed subtests were used rather than the standard unit weighting because of scaling differences between the experimental CAT-ASVAB (which computed scores in the IRT θ -metric) and P&P-ASVAB (number-correct metric)³. Significance tests comparing the validity of CAT-

³The P&P-ASVAB uses unit weighted composites computed from standardized subtest scores for predicting success in training. Since an equating between the CAT and P&P versions was unavailable at the time, CAT-ASVAB scores could not be placed on

ASVAB to each of the two P&P-ASVAB versions were performed. Among the 56 comparisons, only one significant difference was found. This difference favored CAT-ASVAB. These results in general suggested that CAT-ASVAB and P&P-ASVAB predict school performance equally well.

In addition to the predictive validity analysis, the construct equivalence of the CAT and P&P versions were examined via factor analysis, based on correlations between the experimental CAT-ASVAB and the operational ASVAB test scores. Four factors spanning Verbal, Technical, Quantitative, and Speed dimensions were extracted. The pattern of factor loadings across CAT and P&P versions was very similar, and most CAT-ASVAB tests had equivalent or higher factor loadings (consistent with equivalent or higher measurement precision).

The results of the experimental validity study were very encouraging: equivalent construct and predictive validity could be obtained by computerized adaptive tests which administered about 40 percent fewer items than their P&P counterparts. These results provided powerful evidence in support of the operational implementation of CAT-ASVAB.

3.2 OPERATIONAL CAT-ASVAB DEVELOPMENT

The 1984 decision to abandon customized hardware led to the selection of an off-the-shelf system of networked personal computer hardware, and software was developed by project staff. With the resolution of hardware and software issues came a reevaluation and eventual resolution of psychometric aspects of the CAT-ASVAB system. Although the experimental CAT-ASVAB system was a useful research tool, in many respects it was ill-suited for operational use. Before CAT-ASVAB could be administered operationally to military applicants, substantial research and development efforts were needed in the areas of item pool development, psychometric procedures, and delivery system. The high-stakes nature and large volume of Military applicant testing raised the burden of proof for the adequacy of CAT-ASVAB to an extraordinarily high level. Policy guidance from military leadership insisted that in spite of the promising outcomes of the previous empirical studies and many potential benefits of CAT, it was essential for CAT-ASVAB to match or exceed the high standards set by the P&P-ASVAB, and that there should be a very high degree of confidence among researchers and policy makers that these standards have been met. Work on the operational CAT-ASVAB system occurred from about 1985 to 1990.

the same metric as the P&P-ASVAB.

3.2.1 *Item Pool Development*

Items for the first two operational forms of CAT-ASVAB were written and calibrated by Prestwood, Vale Massey, and Welsh (1985). The P&P reference form (8A) was used to outline item content, but differences existed between the test-specifications of the adaptive and conventional versions. The adaptive pools had an increased range of item difficulties and functionally independent items. For the Paragraph Comprehension subtest, this meant asking a single question per passage (for CAT), as opposed to multiple questions for P&P-ASVAB. This functional independence was necessary to help satisfy the IRT assumption of local independence.

About 3,600 items (400 for each of nine content areas) were written and pre-tested on a sample of recruits (providing about 300 responses per item). IRT item parameter estimates were obtained and used to select a subset of highly discriminating items (with an approximately rectangular distribution of difficulties) for more extensive calibration study. The surviving 2,118 items (about 235 items per content area) were assembled into 43 booklets and administered to about 137,000 military applicants. After editing, about 2,700 responses per item were available for item-calibration. All items within a content area were calibrated jointly using ASCAL (Vale & Gialluca, 1985), along with operational P&P-ASVAB items taken for enlistment purposes. This design ensured that item parameter estimates were placed on the same scale across all experimental booklets.

One concern in the development of these pools was whether CAT items must be calibrated from data collected in a computerized administration, or if equally accurate results can be obtained by calibrating items from data collected in a P&P administration (Hetter, Segall, & Bloxom, 1994, 1997). If computer administration of CAT items is required for calibration purposes, then the item pool development effort would be increased substantially. Data from 2,955 military recruits were gathered to estimate two types of calibration-medium effects: (a) whether items calibrated by computer produced adaptive test scores with greater precision than adaptive scores computed from a P&P based calibration, and (b) whether calibration medium effects adaptive test scores in a systematic or non-systematic way.

Examinees were randomly assigned to one of three groups. Shortened pools (for General Science, Arithmetic Reasoning, Word Knowledge, and Shop Information) were constructed from a subset of power test items with high expected usage rates. These pools were administered in fixed blocks by computer to two groups, and by P&P to the third group. Separate IRT calibrations were obtained using data from one computer group and the P&P group. Then each calibration was used to estimate IRT adaptive scores for the remaining computer group. This was accomplished by applying the adaptive item selection and scoring algorithms post-hoc to a subset of responses made by the second computer group. Calibration

medium effects on the measured construct and on the reliability of the test scores were assessed by comparative analyses of the ability estimates using the alternative calibrations. Calibration medium effects on the score scale were assessed by comparing IRT difficulty parameters from computer-based and P&P-based calibrations. Results indicated that item parameter estimates obtained by P&P calibration produced adaptive test scores that have the same reliability and measured the same construct as scores produced from item parameters obtained by computer calibrations. The descriptive analyses of difficulty parameters suggested little or no effect of calibration medium on the score scale.

The IRT model used for CAT-ASVAB (i.e. the 3PL) assumes that each subtest is unidimensional (i.e., all items measure the same, single ability). Violations of this assumption may have serious implications for validity and test fairness. Three approaches were considered for dealing with this problem (Segall, Moreno, & Hetter, 1997): (a) Unidimensional treatment (apply unidimensional adaptive item selection and scoring algorithms without special item-content constraints), (b) Content Balancing (place constraints on the numbers of items administered from targeted content areas), and (c) Pool Splitting (construct and calibrate separate item pools for targeted content areas and measure each from separately administered adaptive tests). For each item pool, a number of analyses were considered in determining the most suitable approach, including: factor analysis of items, the statistical significance of additional factors, factor interpretation, item difficulties, and factor intercorrelations. In accordance with the above guidelines, six of the eight ASVAB power tests were treated as unidimensional, one (General Science) was content balanced, and another (Auto-Shop Information) was split into two pools, and measured by separately administered adaptive tests.

Available items were divided into two parallel item pools (Moreno, 1986). Pairs of items with similar measurement properties were assigned to alternate pools⁴. Through a series of simulation studies, the precision of CAT-ASVAB was compared to that of the P&P-ASVAB at a number of ability levels; it was desirable for CAT-ASVAB to have higher or equal precision at all ability levels. However, for two subtests (Arithmetic Reasoning and Word Knowledge), the precision of CAT-ASVAB fell below that of the P&P-ASVAB over the middle ability ranges. Item pools for these subtests were supplemented with additional items (from the experimental system), and the precision of supplemented pools matched or exceeded that of the P&P-ASVAB at all ability levels.

⁴For test security and retesting purposes, it is necessary for CAT-ASVAB to have at least two forms consisting of unique (non-overlapping) item pools.

3.2.2 Psychometric Procedures

The development of the operational system provided an opportunity to review and revise psychometric procedures. The success of the experimental CAT-ASVAB system made it a useful starting point. However, the high-stakes nature of applicant testing and the operational testing environment imposed additional requirements not specifically addressed in the experimental system development. These requirements dealt with item exposure, final scoring, time-limits, scoring incomplete and speeded tests, collecting data on tryout items, and user interface issues.

Exposure Control

After a review of data from the experimental CAT-ASVAB system, it became apparent that some items had an extremely high exposure rate. Many items were administered to a large proportion of sample. The experimental system used the 5-4-3-2-1 strategy (Wetzel & McBride, 1985), which was intended to guard against one specific type of compromise: remembering response sequences. If item selection were based solely on the hybrid strategy (maximizing information at the provisional Bayesian ability estimate) then a given response pattern (e.g. ADECA) would lead to a deterministic set of presented items, and a predictable test score. Accordingly, once the correct pattern was known, a high score could be obtained by simply remembering the ideal response pattern. The 5-4-3-2-1 strategy guards against this strategy by randomly selecting the first item from among the five most informative, the second item from among the four most informative, and so forth. Item selection for the fifth and subsequent items are based solely on maximum information. This procedure was found to have little decrement in final score precision when compared to optimal (non-random) item selection. However, results from the experimental validity study indicated that the procedure may be susceptible to a different compromise strategy: sharing item-content among test-takers. Some highly informative items of moderate difficulty levels were administered to nearly all subjects. The 5-4-3-2-1 strategy guarded against one type of compromise strategy (remembering response sequences), but did not protect against another (sharing item-content among test-takers).

Recognizing this deficiency, Sympson and Hetter (1985) developed a procedure to guard against both types of strategies. Specifically, the exposure control algorithm was designed to: (a) place an upper limit on the exposure rate of the most informative items, and (b) reduce the predictability of item presentation. The algorithm is probabilistic, and controls item selection during adaptive testing through the use of previously computed parameters associated with each item (Hetter & Sympson, 1997). Items are selected on the basis of maximum information at the provisional ability level, but before an item is administered, a random uniform number is generated and compared to the item specific exposure control parameter.

If the random number is less than the parameter, the item is administered. Otherwise the item is set aside, and not considered again for administration to the test-taker. Exposure control parameters are determined in advance through simulations. Highly informative items of moderate difficulty levels tend to have their usage restricted; items of extreme difficulty or lesser discrimination tend to have little or no usage restrictions. In simulation studies, Hetter and Sympson found the procedure resulted in only modest loss of precision (primarily over the middle ability ranges) when compared to optimal unrestricted item selection.

Stopping Rules

Both variable and fixed-length stopping-rules were considered for use in the operational CAT-ASVAB. In variable length adaptive testing, additional test questions are administered until the examinee's standard error of measurement falls below some pre-specified target level. In fixed length adaptive testing, each test-taker receives a fixed number of items, regardless of the estimated precision of the test score. Fixed length testing was selected for use in CAT-ASVAB primarily because of its increased efficiency over variable length testing. The results of simulations showed that for examinees at the extreme ability ranges (where few informative items exist) the incremental value of each additional item quickly reaches the point of diminishing returns, leading to an inefficient use of the test-takers time and effort.

Scoring

In the experimental CAT-ASVAB system, Owen's Bayesian ability estimate was used to update provisional scores after each administered item, and at the end of the test to provide a final score. As a final score, it has one undesirable property: the score depends on the order in which the items are administered. It is possible for two examinees to receive the same items, provide the same responses, but receive different final Owen's ability estimates! This could occur if two examinees received the items in different sequences. To avoid this possibility, the Bayesian mode ability estimate was selected for use as a final score in the operational system. The Bayesian mode is unaffected by the order of item administration, and, as suggested by simulation studies, provides slightly greater precision than Owen's estimator. Other Bayesian ability estimates were also considered for final scores, but very little difference in precision among the methods was found. Among the alternatives studied, the Bayesian mode required the fewest numbers of computations.

Time-Limits

Administrative requirements forced the imposition of time-limits on each of the adaptive power tests for the operational CAT-ASVAB. There was some

concern about the use of time limits since the standard IRT model does not make allowances for the effects of time pressure on item functioning. Much discussion ensued about the most desirable method for specification of time-limits. One idea was to set the CAT-ASVAB subtest time-limits from the per-item time allowed on the P&P-ASVAB⁵. Another proposal was to examine the distribution of subtest completion-times from the untimed experimental version (using joint Service validity data), and set time-limits at the 95-th percentile. These two methods produced very different time-limits for the nine adaptive tests, and there was substantial pressure among policy makers to use the shorter (P&P-ASVAB based) limits, since they were likely to save additional testing time. However, data collected from an untimed pilot study (Vicino & Moreno, 1997) supported the use of the longer limits. Results indicated that (for reasoning tests) high ability examinees generally took longer than low-ability examinees, indicating that these examinees would be most effected by shorten time-limits. The explanation for this finding was that high ability examinees received more difficult questions, which required additional time to answer. This trend is exactly opposite of anticipated response times from traditional P&P tests. In traditional testing, motivated low ability examinees are generally expected to take longer than high ability examinees.

Ironically, even the longer time-limits based on the untimed experimental pilot study were later found to be too short. The problem was noted during the early stages of data collection for an equating study (Segall, 1997a). For one subtest (Arithmetic Reasoning) fewer than 87% of the test-takers complete all 16 items. Other subtests also had completion rates lower than the 95% target. These early data were used to revise the time-limits.

Penalty for Incomplete Tests

The imposition of time-limits also led to a penalty procedure for incomplete tests. The Bayesian scoring procedure contains a bias: generally estimates are too close to the population mean, and this bias is inversely related to test length. A low ability test-taker could use this property to his or her advantage. Below-average applicants could increase their score by answering the minimum number of items allowed.

To discourage this potential compromise strategy, a penalty procedure was developed for scoring incomplete adaptive tests (Segall, 1987). The procedure provides a final score that is equivalent (in expectation) to the score obtained by guessing at random on the unfinished items. In practice, penalty functions were determined through a series of simulations, and separate functions were determined for each subtest and each possible number of unanswered questions. The procedure has several desirable qualities: (a) the size of the penalty is related to the number of unfinished items; (b)

⁵CAT-ASVAB time limits would be prorated to adjust for differences in test lengths.

applicants who have answered the same number of items and have the same provisional ability estimate will receive the same penalty, and (c) the penalty rule eliminates “coachable” test-taking strategies with respect to answering or not answering test items. Since the penalty procedure may punish high ability test-takers disproportionately (especially when applied to reasoning tests) care was taken to ensure that very few test-takers were penalized. Accordingly, generous time-limits were used, allowing sufficient time to permit over 98 percent of all test-takers to complete each adaptive subtest within the allotted time.

Seeding Tryout Items

The operational CAT-ASVAB administers unscored experimental items for tryout and calibration purposes. Experimental items are administered as the 2nd, 3rd, or 4th item in the adaptive sequence. The experimental item is given early in the sequence where variation in item difficulty among successive items is high. Thus the administration of an experimental item of inappropriate difficulty is likely to be less disruptive when administered towards the beginning of the test than towards the end where item difficulties tend to center tightly around the estimated ability. The position is randomly determined so that it is not apparent to the examinee specifically which item is non-operational. This method of collecting data on new items has several advantages over traditional tryout methods, which collect data under non-operational conditions, require special printing of test booklets, and typically require special data collection studies. By seeding new items among operational items in CAT-ASVAB, tryout data from highly motivated applicants can be obtained without additional data collection efforts.

Speeded Subtests

The ASVAB contains two speeded tests: Numerical Operations and Coding Speed. Variation among scores on these tests is primarily attributed to variation in speed among examinees, rather than variation in response accuracy. (Nearly all attempted items are answered correctly, but examinees differ substantially in the number of reached items.) Consequently CAT-ASVAB speeded tests are not modeled by IRT. Like the P&P-ASVAB, the experimental CAT-ASVAB scored these tests by number-correct. Given equivalent time-limits across versions, it was found that CAT-ASVAB test-takers scored higher than P&P test-takers, primarily because pressing an answer key on computer was faster than marking an answer-sheet bubble with P&P. Thus in the experimental system, an adjustment was made to the speeded-test time limits.

Partly to avoid time-limit specification issues, and partly due to other potential benefits (such as increased precision and desirable score-distribution properties), the operational CAT-ASVAB scores its speeded tests by rate score. One proposal (Greaud & Green, 1986) defined scores as proportion

correct divided by the average response latency. Prior to operational use however, it was discovered that this score was susceptible to a possible compromise strategy: pressing answer keys rapidly, and at random. The positive numerator (about .25) and the small denominator produced very high scores, higher than was possible by reading and properly answering the question. Consequently, the operational system incorporates a correction for guessing in the numerator, so that the expected rate score is zero for random guessing.

Administrative Requirements

A number of administrative requirements were imposed on the system to make it suitable for operational use, including easy procedures for changing and confirming power test answers, implicit help calls (brought about by repeatedly pressing invalid keys, or not responding to an item), explicit help calls (to address questions raised by the examinee), and a clock (displayed on the lower right-hand corner of the screen showing the number of items and time remaining on the subtest). Because self-paced test-takers begin and end each timed section at different intervals, a standard wall clock or timer would be cumbersome. The computerized clock provides individualized information, allowing test-takers to pace themselves and use allotted time efficiently. Test-takers are not allowed to skip items, or to review previously answered items. The CAT-ASVAB branching feature requires a response to each item. If allowed, omitting would lead to less optimal scoring, and possibly to particular compromise strategies.

Operational Delivery System

In 1984, the contractor effort to develop a customized hardware platform was abandoned in favor of an off-the-shelf system consisting of networked personal computers (Rafacz & Hetter, 1997). At this time, micro-computing was still in its infancy, and only a few standards were available. The Hewlett Packard (HP) Integral Computer was selected primarily because of its superior portability (17 pounds), large random access memory (1.5 megabytes), fast CPU (8 MHz 6800 Motorola), and advanced graphics display capability (9 inch monitor with electroluminescent display and resolution of 512 by 255 pixels)⁶. The HP operating system was UNIX based, and supported the C programming language. Each station contained a floppy diskette drive, had no internal hard drive, and cost about \$5,000.

The system design emphasized the use of RAM, and used a network to download software and items, and upload test-response data. To increase test security, no hard drive storage was used. All software and data were stored in RAM at each station. This allowed each station to operate inde-

⁶Although these specifications seem pathetic by today's standards, they were advanced relative to other personal computers of the time.

pendently of every other station so that network traffic did not slow system response. In the case of hardware failure, the test-taker response data was recorded on diskette, and could be transferred to any other station in the network for completion of the battery. Software was designed and written by in-house personnel to implement the psychometric procedures and item pools for operational use.

Although software development is an obviously important step in system development, an equal, but not so obviously important step is acceptance testing. Here, we make a distinction between software testing (performed by software programmers) and acceptance testing (performed by an independent group, preferably those who are most familiar with the system requirements). In development of the operational CAT-ASVAB system, the time and effort dedicated to acceptance testing matched or exceeded that spent by programmers developing and debugging code.

Several useful refinements to acceptance testing procedures developed over the course of the project. First, it is useful to prioritize errors into level of severity, into those that effect: (I) what the examinee observes (item selection, familiarization, and practice screens), (II) final test scores, and (III) other software functions (networking, failure recovery, database accuracy, etc.). The severity of the error will often determine the necessary course of action, and the timing of this action. Almost without exception, Type I errors must be fixed since they have psychometric implications for test accuracy and validity. However, if the system is used non-operationally for low-stakes research, then some errors of Type II and III errors can be tolerated, since final scores can be computed after-the-fact, and not all data and networking functions are necessary. If the system is operational, then some Type III errors may be tolerated, since not all database fields and software functions are crucial. The course of action taken for each discovered error depends on the software use and the specific nature of the error.

Another useful acceptance-testing refinement included strict configuration management controls. To ensure that the proper software components were evaluated, all source code was compiled on a dedicated configuration management computer. The resulting computer program executable was compared to that provided by the software development group. This step was especially important when large numbers of individuals were involved in source code development, and changes were being made on a frequent basis.

A catalog of testing scenarios was developed over the course of the project to test important software functions. As errors were discovered and refinements made, additional scenarios were added to ensure software compliance. Two types of scenarios were developed: (a) automated (where test-taker input was read from a file), and (b) manual (where input was entered manually at the keyboard). Automated checks were used to verify item selection and scoring; manual scenarios were used to verify the accuracy of

item presentation, system timing, database integrity, and failure-recovery procedures. Because of the complex interdependencies among software components, the entire catalog of scenarios was examined after a software modification, regardless of how small or trivial the change may have appeared.

3.3 KEY RESEARCH AND OUTCOMES

A number of studies played central roles in the eventual decision to implement CAT-ASVAB nationwide. These studies examined the comparability of the CAT and P&P versions of the ASVAB, the utility of adding new computerized predictors, and the economic benefits derived from CAT. This research provided policy makers with the reassurance necessary to make dramatic changes to the Armed Forces selection and classification system.

3.3.1 *Human Factors*

During the 1970's and early 1980's, the use of computers among young men and woman was limited primarily to those with specialized interests. There was some concern that lack of computer experience among the majority of youth would be an impediment to accurate and valid CAT-ASVAB measurement. It was also believed that the favorable results obtained with the experimental CAT-ASVAB concerning clarity of instructions may not generalize to military applicants. First, the instructions for the operational CAT-ASVAB system had undergone extensive revisions to accommodate necessary changes for administration to applicants. Second, all previous studies had been conducted with recruits who had taken the ASVAB (P&P version) prior to enlistment, and who had scored in the middle or upper ranges. Consequently, it was important to evaluate instructions on a broad range of test-takers who did not have the benefit of prior ASVAB exposure.

Data were gathered from a sample of 231 military applicants and 73 high school students to address issues relating to computer familiarity, instruction clarity, and attitudes towards CAT-ASVAB. Data were collected during October and November of 1986. After completing CAT-ASVAB, each participant received a 42-item questionnaire. In addition, about 90 test-takers participated in structured interviews. In general, both computer-naïve and experienced test-takers felt very comfortable using the HP-computer, exhibited positive attitudes towards CAT-ASVAB, and preferred a computerized test over P&P. Test-takers strongly agreed that the instructions were easy to understand, although some examinees did not understand particular words in the instructions (like "proctor"). Based on the pilot study results, the reading grade level of some words and phrases was lowered to make the directions comprehensible to the target applicant population. The

only negative outcome was the finding that most test-takers were bothered by not being able to review and modify previously answered questions. Because of the requirements of the adaptive testing algorithm, this aspect of CAT-ASVAB was not altered.

3.3.2 Reliability and Construct Validity

The constructs measured by an adaptive test, and the precision of scores depends on a number of factors, including contents and quality of the item pool, item selection, scoring, and exposure algorithms, and the clarity of test instructions. Although IRT provides a basis for making theoretical predictions about these psychometric properties, most assumptions on which these predictions are based are violated, at least to some degree. Consequently before CAT-ASVAB scores were used operationally in high-stakes testing, an empirical verification of its precision and construct equivalence with the P&P-ASVAB was conducted. If CAT and P&P versions measured the same constructs, and the CAT-ASVAB version provided equal or greater precision, then the massive amount of predictive validity evidence accumulated on the P&P version would be directly applicable to CAT-ASVAB. Construct equivalence would also support the exchangeability of the two versions, allowing both versions to be used concurrently for selection and classification.

To study reliability and construct validity, military recruits were administered two alternate ASVAB forms (Moreno & Segall, 1997). One group ($N = 1033$) received two P&P-ASVAB forms; another randomly equivalent group ($N = 1057$) received two CAT-ASVAB forms. All participants' operational P&P-ASVAB scores taken prior to enlistment were also analyzed. Reliability coefficients for each medium were estimated from the correlations between liked named subtests of alternate forms. Evidence of construct equivalence was obtained from disattenuated correlations between CAT-ASVAB and operational P&P-ASVAB versions.

In comparison to the P&P-ASVAB, seven of the ten CAT-ASVAB tests displayed significantly higher alternate-forms reliability coefficients. The other three tests displayed non-significant differences. Nine of the ten disattenuated correlations between CAT and the operational P&P were about 1.0 (with a very narrow confidence interval due to the large sample sizes). Only one speeded test (Coding Speed) displayed a disattenuated correlation substantially less than one (.86), which may be attributed to the differences in instruction clarity between the written (CAT) and oral (P&P) versions. The low disattenuated correlation for Coding Speed was not considered problematic because selection composites that contain this subtest had high disattenuated CAT-P&P correlations approaching 1.0. In general, these results confirmed the expectations based on theoretical IRT predictions: that CAT-ASVAB measured the same constructs as its P&P counterpart with equivalent or greater precision.

3.3.3 *Equating CAT and P&P Versions*

Equating is a psychometric analysis designed to place scores from two versions of a test (typically an older reference form and a new form) on the same scale. Equating CAT-ASVAB to the P&P scale was seen as a major psychometric hurdle to implementation. Historically, qualification standards for entrance into the military and into occupational specialties had been specified relative to the P&P-ASVAB number-correct score-scale. The CAT-ASVAB produces scores on the IRT ability metric. Before CAT-ASVAB could be used in high-stakes testing, an equating procedure for placing CAT scores on the P&P metric was required. The objective of the equating was to provide a transformation of the CAT-ASVAB scale so that its score distribution would match the P&P version. This transformation, when applied to CAT-ASVAB, would allow scores on the versions to be used interchangeably, without disrupting applicant qualification rates.

The equating study (Segall, 1997a) addressed three concerns. First, how could qualification rates associated with the existing P&P-ASVAB cut-scores be preserved for CAT-ASVAB? Several equating procedures were considered and rejected. Ultimately, an equipercentile equating procedure based on observed test scores from test-takers was used to obtain the required transformations. Distribution smoothing procedures were used to increase the precision of the transformations and the equivalence of CAT-P&P composite distributions were verified to ensure that the use of CAT-ASVAB would not disrupt flow rates. (Although equating was performed at the subtest level, qualification scores are based on composites.)

A second concern dealt with disadvantaged subgroups: subgroup members taking CAT-ASVAB should not be placed at a disadvantage relative to their subgroup counterparts taking the P&P-ASVAB. Although it is desirable to match distributions for subgroups as well as the entire group, this may not be possible for a variety of reasons. First, differences in precision between the CAT and P&P versions may magnify existing differences between subgroups. Second, small differences in dimensionality, such as the verbal loading of a test, may cause differential subgroup performance. The issue of subgroup differences was addressed by applying the equating transformation (based on the entire group) to subgroup members taking CAT-ASVAB, and comparing their distribution to their P&P counterparts. For specified subgroups (Blacks and females), the difference between CAT and P&P means was used to assess relative advantage/disadvantage among CAT-ASVAB test-takers. Although some statistically significant subgroup differences were observed from the equating study data, their practical significance on qualification rates was small.

The final concern addressed by the equating study dealt with the effects of motivation and other population characteristics on the equating transformation. Specifically, who should participate in the study, and under what conditions should they be tested? To eliminate the effects of motivation on

the final equating transformation, the study was conducted in two phases. The first phase (Score Equating Development) was used to obtain a provisional equating based on data collected under non-operationally motivated conditions. The second phase (Score Equating Verification) was used to obtain an equating transformation based on operationally motivated applicants, whose CAT-ASVAB scores were transformed to the P&P metric using the provisional equating. For P&P-ASVAB equating studies, the first phase is typically conducted on a sample of convenience—military recruits. However, for this first attempt at CAT-ASVAB equating, the provisional equating was based on military applicants. Applicants were chosen over recruits primarily to provide the full range of scores necessary to obtain a precise estimate of the transformation over the lower ranges. The choice of applicants resulted in an especially burdensome data collection effort from both the applicant and recruiter’s perspective. It meant a full day of testing, rather than the normal half day. (In addition to taking a non-operational CAT or P&P version, each applicant was required to take an operational P&P version for enlistment purposes.) The choice of applicants was especially fortuitous since a later study conducted with recruits found the equating transformation to differ across recruit and applicant populations (Segall, 1997c).

The equating data collection and analysis spanned a four year period from 1988 to 1992. For the Score Equating Development phase, data were gathered from over 8,000 military applicants. Three randomly equivalent groups of about 2700 each were administered either a non-operational P&P-ASVAB form, or one of two non-operational CAT-ASVAB forms. Scores on the non-operational CAT and P&P versions did not effect the applicant’s eligibility for enlistment, and were used to construct the provisional equating⁷. For the second Score Equating Verification Phase, additional data were gathered from about 10,400 applicants. Three randomly equivalent groups of about 3,500 each were administered either an operational P&P-ASVAB form, or one of two operational CAT-ASVAB forms. In this phase, scores on the CAT and P&P versions did effect the applicant’s eligibility for enlistment, and were used to construct the final equating for CAT-ASVAB.

The results of the equating studies were noteworthy from several respects. First, although some statistically significant subgroup differences were observed, their practical significance on qualification rates was small. Second, the provisional and final transformations obtained from the non-operational and operational data collections were very similar, indicating that data collection and analytic procedures were consistent and reliable.

⁷Non-operational tests were administered first, followed by an operational P&P-ASVAB used for selection and classification. By using only data from the first (non-operational) test to develop the equating, levels of fatigue and practice were expected to be equivalent across CAT and P&P versions, and were expected to closely match those occurring under operational testing conditions.

Third, the beginning of the Score Equating Verification phase in September of 1990 marked a milestone in the CAT-ASVAB project: for the first time CAT-ASVAB was administered operationally in a high stakes environment to qualify applicants for Military Service.

3.3.4 Economic Analysis I: Enhancing ASVAB Content

Ironically, as data were collected to address the last major psychometric obstacle (equating), the future prospects for CAT-ASVAB implementation had reached an all time low. This pessimism was based primarily on a costs-benefits analysis (Automated Sciences Group, and CACI, 1988). This analysis weighed cost associated with CAT-ASVAB (e.g. computer hardware) against the potential savings accrued by improved selection and classification. The benefits of CAT-ASVAB were measured in terms of improved prediction of job success using a formulation developed by Cronbach and Gleser (1965). At the heart of this approach was the notion that the increased precision of CAT-ASVAB would lead to improved predictive validity. Unfortunately, at this point in the development cycle, the emphasis in efficiency led to a design which stressed time-savings (i.e. short tests), as opposed to increased precision. CAT-ASVAB was designed to be about half as long as its P&P counterpart, with equal or slightly greater precision. Consequently, the gain in predictive validity (correlation of test performance with training success) obtained by CAT-ASVAB was estimated to be a mere .005. Results indicated that the benefit based upon the dollar value of improved person-job match was not great enough to offset the costs of computers. It was believed that a significant increase in predictive validity (above that provided by CAT-ASVAB) would be required to make CAT cost effective. Accordingly the CAT-ASVAB program was redirected toward a Joint-Service validation of new computerized cognitive and psychomotor tests. Support for this redirection came from a prediction made by Schmidt, Hunter, and Dunn (1987) who indicated that adding perceptual speed and psychomotor tests to the ASVAB could result in hundreds of millions of dollars worth of personnel performance improvements annually.

The enhanced CAT (or ECAT) validity study (Wolfe, Alderton, Larson, Bloxom, & Wise, 1997) was intended to provide an empirical verification of the increased validity associated with new types of computerized tests. Specifically, the study was designed to identify the aptitude constructs that are likely to make the greatest contribution to increased validity and to provide estimates of their relative validity gains. Nine tests measuring four constructs were included in the experimental battery: non-verbal reasoning, spatial ability, psychomotor skill, and perceptual speed. Six of the nine tests required computer administration, whereas three tests had P&P

counterparts⁸. These measures can be classified as tests of fluid intelligence, rather than as measures of crystallized intelligence assessed by most ASVAB tests. As measures of fluid intelligence, it was believed (and later confirmed with empirical data) that these measures were likely to display less adverse impact for educationally disadvantaged subgroups.

Predictor and criteria data were gathered from over 11,000 recruits attending one of 18 schools. Criteria included quizzes, homework assignments, and laboratory/shop exercises. Where possible, performance criteria were used in preference to written tests, since it was expected that the new tests would display the largest incremental validities with hands-on measures. The incremental validity of the new measures over ASVAB was estimated to be .031 using these performance based criteria. According to the utility analysis, this gain should be sufficient to offset capital investment in computers. However, for several reasons, these results were not viewed as justification for CAT-ASVAB's benefits.

First, to obtain the full validity increment, the entire battery of nine tests (taking about 3 hours) would need to be added to the ASVAB. This was considered impractical from a recruiting and processing standpoint. With room to add only a small number of tests, the resulting incremental validity was not considered sufficiently large. Because of other practical constraints, tests were not considered for inclusion unless they could also be administered by P&P⁹. Among the remaining tests, Assembling Objects had the broadest application (in terms of incremental validity) and was tentatively added to the ASVAB, pending results of more extensive validity studies.

With ECAT validity data in hand, a second economic analysis was conducted to assess the feasibility of implementing CAT-ASVAB.

3.3.5 *Economic Analysis II: Effects on Applicant Processing*

In 1992, a confluence of events shaped the next economic analysis, and the eventual decision to implement CAT-ASVAB. During this time frame, the ECAT validity study was nearing completion, and results were only moderately supportive of sufficient incremental validity. The equating study had just been completed, overcoming the last hurdle to operational use. And preparations were underway for an updated economic analysis. In framing the approach for the second economic analysis, it was apparent that the use

⁸CAT-ASVAB provided two practical opportunities for expanding ASVAB content. First the computerized platform allowed types of measures not possible in P&P format. Second, the time savings resulting from the shortened adaptive power tests provided extra time for computerized versions of P&P assessments.

⁹The ASVAB concept of operation specified that the battery would be administered in both CAT and P&P formats, and that the same constructs should be measured by both versions.

of hypothetical utility dollars did not provide sufficiently compelling justification to policy makers who would be asked to divert real dollars from other sources to pay for computer hardware. Consequently, the revised utility analysis was expanded to include the economic effects of CAT-ASVAB on applicant processing.

The operational impact of CAT-ASVAB was examined in an Operational Test and Evaluation (OT&E) study conducted at five geographically disperse locations (Moreno, 1997). Together these sites tested about 7 percent of all military applicants. Several types of data were collected, including CAT-ASVAB, on-site observations, interviews (with testing personnel and recruiters), and questionnaires (recruiter and applicants). Several key findings with economic ramifications were identified. First, flexible start times (the ability for test-takers to start the test at individualized times) and the shorter CAT-ASVAB test length led to a reduction in the estimate of the number of required computers. Second, CAT-ASVAB enabled some sites to conduct one-day processing, where the candidate could complete all processing and screening activities in a single day. Because of fewer meal and lodging expenditures, applicant processing costs were substantially reduced.

The limited implementation of the OT&E study also marked an important turning point in project support. Exceptional system design and performance led to the enthusiastic support of CAT-ASVAB by testing and recruiting personnel. It also eliminated any remaining concerns from Policy and Technical staff about the operational impact of the system. In fact, users of CAT-ASVAB were so enthusiastic that upon completion of the study all expressed an extreme reluctance to revert back to P&P testing. Consequently, CAT-ASVAB testing continued at these locations until the final system was implemented nationally several years later.

Armed with data from the OT&E study, a second economic analysis was conducted (Wise, Curran, & McBride, 1997). The primary objectives of the study were to determine how CAT-ASVAB would be used operationally, and if its benefits justified hardware and other incremental costs. This study held the quality of selection and classification decisions constant across P&P and CAT versions, and rather focused on benefits relating to the reduction in recruiting and enlistment processing costs. The length (time) of the battery influenced these later costs. Thus the primary strength of CAT-ASVAB (shorter testing time) played a significant role in the outcome, unlike the first economic analysis which did not consider operational savings. Based in large part on the feasibility of one-day processing with CAT-ASVAB, the economic analysis indicated that savings in recruiting and processing costs would exceed the hardware costs after just one year. This outcome played a significant role in the decision to implement CAT-ASVAB at all high-volume locations nationwide, which together test about half of all applicants. Other applicants are tested at Mobile Examining Team Sites (METS), where it is impractical, for the most part, to

use desktop computers. The feasibility of implementing CAT-ASVAB at METS will be examined in more detail to identify the most cost-effective strategy.

3.4 NATIONWIDE IMPLEMENTATION

Before CAT-ASVAB could be administered on a national scale, a new delivery system was required. The HP computers used previously were no longer manufactured. The use of new machines raised issues about the psychometric comparability of different hardware. Additional studies were conducted to address this and other issues concerning test compromise and the development of new forms.

3.4.1 Delivery System

The requirements of the nationwide delivery system were based on the capabilities of the HP and experiences from the OT&E study (Unpingco, Hom, & Rafacz, 1997). A number of factors were considered in hardware selection, including portability, microprocessor, random access memory, disk storage, monitor, networking capability, and input device. An attempt was made to select a system constructed from commonly used components. This would likely reduce maintenance costs, provide for future growth, and delay system obsolescence. Fortunately, nearly all performance standards of the HP system (designed in the mid 1980's) were met or exceeded by personal computers available in 1993. The selected system was an Intel-based compatible, with a 33Mhz or faster microprocessor. Only 4MB of RAM were required, with a 14-inch SVGA video monitor. Additional requirements included an 80MB hard drive, and an ethernet-networking card. Many of the psychometric routines for test administration and scoring were transported to the new system, although about 80 percent of the code was rewritten and designed specifically for the MS-DOS environment.

3.4.2 Hardware Effects

The use of a new delivery system led to concerns about the psychometric comparability of different hardware. It was conceivable that differences among computer hardware (monitor size and resolution, keyboard layout, physical dimensions) could influence item functioning. There was some evidence that speeded tests contained in the ASVAB were especially sensitive to small changes in test presentation format, more so than the adaptive power tests. A study was conducted to provide some insight into the exchangeability of different hardware—whether machines of different makes and models can be used interchangeably, and which hardware character-

istics must remain constant among testing platforms to ensure adequate precision and score interpretation (Segall, 1997b). The study was designed to examine three psychometric characteristics, including score-scale, precision, and construct validity. Data were gathered from 3,062 subjects recruited from the San Diego area. Each subject was randomly assigned to one of 13 conditions. The effects of several hardware characteristics were studied, including input device, color scheme, monitor type, CPU speed, and portability. The outcome of the study indicated that adaptive power tests were robust to differences among computer hardware, whereas speed tests are likely to be effected by several hardware characteristics.

Results of the hardware effects study demonstrated the sensitivity of speeded test to hardware differences, supporting the need for an additional equating study using the new hardware. As in the previous study, this equating was conducted in two phases. The first phase used recruits to develop a provisional transformation; it used a random groups design with about 2,500 respondents per form. The second phase tested applicants using the provisional transformation to provide operational scores. This second data collection effort was also a random groups design with about 10,000 test-takers per form.

After completion of the first phase, there was suspicion that the provisional equating was flawed. This was based on a comparison of transformations (for the same forms) estimated from previous studies using the HP. It was hypothesized that for recruits there were different levels of motivation/fatigue between CAT and P&P groups, and this resulted in a biased estimate of the provisional equating transformation. The difference was in a direction that suggested that CAT examinees were more motivated than P&P examinees (possibly due to shorter test lengths or novel/interactive medium). Consequently, the prior equating based on the HP system was used provisionally (for power tests), and the equating based on the new hardware was used for the speed tests. A later analysis of operational applicant data confirmed suspicions that the recruit equating was flawed. Findings suggested that the results of a cross-medium equating may differ depending upon whether the respondents are motivated or unmotivated. In future equatings, this problem may be circumvented by performing only within medium equatings when a sample's degree of motivation is in doubt.

3.4.3 *Test Compromise*

Several concerns about CAT-ASVAB test-security led to a re-evaluation of the effects of possible compromise. First, expected item usage rates resulting from the Simpson-Hetter algorithm are based on an assumed ability distribution, and may depart from those obtained with the actual ability distribution, especially for homogeneous subgroups. Second, each adaptively administered item may influence the final score more than an item on the P&P test, because CAT tests tend to be shorter than their P&P

counterparts. Thus, knowledge of a single CAT item may result in a larger score gain than knowledge of a single item administered in a conventional P&P test. These concerns about the susceptibility of CAT to compromise motivated a simulation study (Segall, 1995) which examined the expected score gains resulting from six different compromise strategies. Strategies differed across three dimensions: the transmittal mechanism (sharing among friends or item banking), the correlation between the cheater and informant ability levels, and the method used by the informant to select items for disclosure. The dependent measure was score gain, which represented the mean gain for the group of cheaters over a group of non-cheaters for the same fixed ability level. Score gains were computed for both CAT (assuming two forms), and for P&P (assuming 6 forms). The results indicated that the score gains for CAT were larger than those for the corresponding P&P conditions, with the largest CAT-P&P gain differences occurring for cheaters at the lowest ability ranges. Results suggested that more stringent item exposure controls should be imposed on the adaptive selection process. The effect of altering several characteristics of the adaptive test on potential score gains were also investigated. It was found that increasing the number of CAT forms (from two to three) had a substantial reduction in score-gain. Using three forms of CAT provided score gains equivalent or less than those observed for six forms of the P&P-ASVAB under all compromise strategies. These results led to the decision to implement additional CAT-ASVAB forms.

Two additional CAT-ASVAB forms were subsequently developed (Thomason, 1996), one for use in the operational testing program and another for use in a national norming study. The operational form (along with two other previously developed forms) will be used to help guard against compromise. The second form will become the new ASVAB reference form, to be administered to a nationally representative sample of young men and women. This form will be withheld from routine use, and only administered in future equating studies. For these new forms, about 5,200 items were developed. About half the items were calibrated on a sample of over 100,000 military applicants, providing about 1500 responses per item. Analyses found that dividing the item bank into two forms resulted in less precise ability estimates in some areas than the P&P-ASVAB. Consequently, some content areas were supplemented with highly informative items (originally intended for new P&P-ASVAB forms) to meet the precision criterion. These new forms will undergo equating studies similar in design to previous forms.

The implementation of CAT-ASVAB is expected to lead to additional challenges and refinements (Segall & Moreno, 1997). CAT-ASVAB data are being collected as part of the 1997 Profile of American Youth Study, which is a national norming study conducted jointly by the Department of Labor and the Department of Defense. It will serve two important purposes. First it will provide information about the availability of high quality

young men and women. This information can be used by force planners to determine the levels of advertising and enlistment incentives required to attract the necessary numbers of qualified applicants to fill jobs of increasing complexity. These data will also provide an opportunity to develop a new score-scale based on the natural IRT metric. This score scale may have improved measurement properties over the existing number-right scale based on the P&P-ASVAB.

The implementation of CAT-ASVAB will also enable considerable streamlining of new form development. DoD is considering the possibility of eliminating all special form-development data-collection studies by replacing them with online calibration and equating. Currently, new item data is being collected by seeding experimental items among operational items. These data will be used to estimate IRT item parameters. These parameters can in turn be used to construct future forms, and possibly estimate provisional equating transformations. These provisional (theoretical) equatings could then be updated after they are used operationally to test randomly equivalent groups. Thus, the entire cycle of form development can, in principle, be seamlessly integrated into operational test administrations.

3.5 SUMMARY

Over the last two decades, many benefits of computerized adaptive testing to the U.S. Armed Forces have been enumerated, studied, and placed into practice. As the world's largest employer of young men and women, the Department of Defense ensured that the CAT-ASVAB matched or exceeded the high standards set by the P&P-ASVAB before making an implementation decision. This assurance was provided by numerous theoretical and empirical studies, and along the way to implementation, a number of important contributions to the field of psychometrics were made. In the years to come, inevitable ASVAB changes and refinements will likely add even greater efficiencies to this important component of the Armed Services selection and classification system.

3.6 References

- Automated Sciences Group & CACI. (1988). *CAT-ASVAB program: Concept of operation and cost/benefit analysis*. Fairfax, VA: Author.
- Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.
- Greaud, V. A., & Green, B. G. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement*, *10*, 23-34.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). Evaluating item calibration mode in computerized adaptive testing. *Applied Psychological Measurement, 18*, 197-204.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 161-167). Washington, DC: American Psychological Association.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Maier, M. H. (1993). *Military aptitude testing: The past fifty years*. (Report No. 93-007). Monterey, CA: Defense Manpower Data Center.

Martin, C. J., & Hoshaw, R. (1997). Policy and program management perspective. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 11-20). Washington, DC: American Psychological Association.

McBride, J. R. (1997). The Marine Corps exploratory development project: 1977-1982. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 59-67). Washington, DC: American Psychological Association.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive verbal ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223-235). New York, NY: Academic Press.

Moreno, K. E. (1986). *A procedure for producing parallel item pools for computerized adaptive testing*. Unpublished manuscript, San Diego, CA: Navy Personnel Research and Development Center.

Moreno, K. E. (1997). CAT-ASVAB operational test and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 199-205). Washington, DC: American Psychological Association.

Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169-174). Washington, DC: American Psychological Association.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. *Applied Psychological Measurement*, *8*, 155-163.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (RB-69-92). Princeton, NJ: Education Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: development of an adaptive item pool* (TR-85-19). San Antonio, TX: Air Force Systems Command, Brooks Air Force Base.

Rafacz, B., & Hetter, R. D. (1997). ACAP hardware selection, software development, and acceptance testing. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 145-156). Washington, DC: American Psychological Association.

Schmidt, F. L., Hunter, J., & Dunn, W. (1987). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery* (TN-95-5). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1987). *A procedure for scoring incomplete adaptive tests*. Unpublished manuscript, San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1995, May). *The effects of item compromise on computerized adaptive test scores*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Segall, D. O. (1997a). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181-198). Washington, DC: American Psychological Association.

Segall, D. O. (1997b). The psychometric comparability of computer hardware. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 219-226). Washington, DC: American Psychological Association.

Segall, D. O. (1997c, March). *The effects of motivation on equating adaptive and conventional tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Segall, D. O., & Moreno, K. E. (1997). Current and future challenges. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized*

adaptive testing: From inquiry to operation (pp. 257-269). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117-130). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 131-140). Washington, DC: American Psychological Association.

Segall, D. O., Moreno, K. E., Kieckhafer, W. F., Vicino, F. L., & McBride, J. R. (1997). Validation of the experimental CAT-ASVAB system. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 103-114). Washington, DC: American Psychological Association.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive tests*. Paper presented at the Annual Conference of the Military Testing Association. San Diego, CA: Military Testing Association.

Thomasson, G. L. (1996). *Item pool development for CAT-ASVAB Forms 3 and 4*. Unpublished manuscript, Monterey, CA: Defense Manpower Data Center.

Unpingco, V., Hom, I., & Rafacz, B. (1997). Development of a system for nationwide implementation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 209-218). Washington, DC: American Psychological Association.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, *34*, 253-269.

Urry, V. W. (1983). *Tailored testing and practice: A basic model, normal ogive models, and tailored testing algorithms* (NTIS No. AD-A133385). Washington, DC: Office of Personnel Management.

Vale, C. D., & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (RR ONR 85-4). St. Paul, MN: Assessment Systems Corp.

Vicino, F. L., & Moreno, K. E. (1997). Human factors in the CAT system: a pilot study. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 157-160). Washington, DC: American Psychological Association.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (RR 74-5). Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota.

Weiss, D. J. (1975). *Computerized adaptive ability measurement*. Paper presented at the Annual Conference of the Military Testing Association, Fort Benjamin Harrison, IN.

Wetzel, C. D., & McBride, J. R. (1983). *The influence of fallible item parameters on test information during adaptive test* (TR 83-15). San Diego, CA: Navy Personnel Research and Development Center.

Wetzel, C. D., & McBride, J. R. (1985). Reducing the predictability of adaptive item sequences. *Proceedings of the Annual Conference of the Military Testing Association*, 1, 43-48.

Wise, L. L., Curran, L. T., & McBride, J. R. (1997). CAT-ASVAB cost and benefit analyses. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 227-236). Washington, DC: American Psychological Association.

Wolfe, J. H., McBride, J. R., & Sympson, J. B. (1997). Development of the experimental CAT-ASVAB system. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 97-101). Washington, DC: American Psychological Association.

Wolfe, J. H., Alderton, D. L., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 239-249). Washington, DC: American Psychological Association.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST-A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton, NJ: Education Testing Service.