CAT-ASVAB Technical Bulletin #1

Personnel Testing Division
Defense Manpower Data Center

March 2006

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **22 MAR 2006**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **CAT-ASVAB Technical Bulletin #1**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Personnel Testing Division Defense Manpower Data Center**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**293**	

Acknowledgements

This Technical Bulletin includes some material that has been published by the American Psychological Association in *Computerized Adaptive Testing: From Inquiry to Operation*, 1997, edited by William A. Sands, Brian K. Waters, and James McBride. The chapters presented herewith were prepared by government employees and are, therefore, in the public domain:

Chapter 2
Segall, D. O., Moreno, K. E., & Hetter, R. D.
"Item Pool Development and Evaluation" (pp. 117-130)

Chapter 3
Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D.
"Psychometric Procedures for Administering CAT-ASVAB" (pp. 131-140)

Chapter 4
Hetter, R. D., & Sympson J. B.
"Item Exposure Control in CAT-ASVAB" (pp. 141-144)

Chapter 5
Rafacz, B., & Hetter, R. D.
"ACAP Hardware Selection, Software Development, and Acceptance Testing" (pp. 145-156)

Chapter 6
Hetter, R. D., Segall, D. O., & Bloxom, B. M.
"Evaluating Item Calibration Medium in Computerized Adaptive Testing" (pp. 161-168)

Chapter 7
Moreno, K. E., & Segall, D. O.
"Reliability and Construct Validity of CAT-ASVAB" (pp. 169-174)

Chapter 8
Wolfe, J. H., Moreno, K. E., & Segall, D. O.
"Evaluating the Predictive Validity of CAT-ASVAB" (pp. 175-180)

Chapter 9
Segall, D. O.
"Equating the CAT-ASVAB" (pp. 181-198)

Chapter 10
Moreno, K. E.
"CAT-ASVAB Operational Test and Evaluation" (pp. 199-206)

Chapter 11
Unpingco, V., Hom, I., & Rafacz, B.
"Development of a System for Nationwide Implementation" (pp. 209-218)

Chapter 12
Segall, D. O.
"The Psychometric Comparability of Computer Hardware" (pp. 219-226)

CAT-ASVAB Technical Bulletin — Table of Contents

CAT-ASVAB Technical Bulletin — Table of Tables

CAT-ASVAB Technical Bulletin — Table of Figures

Chapter 1

Introduction to CAT-ASVAB

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 were 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.

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 paper-and-pencil version of the ASVAB (P&P-ASVAB) was administered to all 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 P&P-ASVAB was believed susceptible to compromise and coaching (Maier, 1993). Historically, the P&P-ASVAB program has offered continuous, on-demand scheduling opportunities at nearly 1,000 testing sites located in geographically disperse areas.

Further, both applicants and recruiters have had strong incentives to exchange information on operational test questions because (a) high-scoring applicants can qualify for Service enlistment bonuses, educational benefits, and desirable job assignments; and (b) performance standards for recruiters are based on the number of high-scoring applicants they enlist. Around the time of the P&P-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 P&P-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 (CAT) 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 applicants and recruiters less profitable, since each applicant would receive items tailored to his or her specific ability level.

Part way through a Marine Corps Exploratory Development Project, the Department of Defense (DoD) initiated a Joint-Service Project for development and further evaluation of the feasibility of implementing a 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. Perry. By this time, there was a strong interest in a 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 testing. 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: (a) contractor delivery-system development (hardware and software to administer CAT-ASVAB), and (b) 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 examinee 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. 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 and 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 (3PL) 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 1980s, 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 (a) the true ability value, and (b) 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 precalculated 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 bases 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). These items were pre-tested on samples of military recruits, and items with low discrimination were removed from the pools. The remaining items were administered in P&P booklets to over 100,000 military applicants (providing about 1,500 responses per item). IRT item parameter estimates were obtained using a 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 P&P booklets be suitable for use with these same items administered in a CAT format? Given the large numbers of examinees 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).

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, & Sympson, 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).

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 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. To help control for the influence of 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: (a) the operational P&P-ASVAB taken prior to enlistment, (b) the experimental CAT-ASVAB taken during basic training, and (c) selected P&P-ASVAB tests also taken during basic training. 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.

With the resolution of hardware and software issues came a re-evaluation and eventual resolution of the 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, (a) 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 (b) 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.

Over the past few 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 DoD 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 Military Services selection and classification system.

References

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 Associates.

Maier, M. H. (1993). *Military aptitude testin: 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., & 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.

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

Segall, D. O., Moreno, K. E., Kieckhaefer, 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.

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.

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

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.

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.

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 7606). Princeton, NJ: Educational Testing Service.

Chapter 2

CAT-ASVAB ITEM POOL DEVELOPMENT AND EVALUATION

By the mid-1980s, an item pool had been constructed for use in an experimental computerized adaptive testing of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) system (Wolfe, McBride, & Sympson, 1997) and had been administered to a large number of subjects participating in research studies. However, this pool was ill-suited for operational use. First, many items had been taken from the retired paper-and-pencil (P&P) ASVAB forms (8, 9, and 10). Using these items in an operational CAT-ASVAB would degrade test security since these items had broad exposure through the P&P testing program. In addition, the experimental CAT-ASVAB system contained only one form. For re-testing purposes, it is desirable to have two parallel forms (consisting of non-overlapping item pools) to accommodate applicants who take the battery twice within a short time interval. To avoid practice and compromise effects, it is desirable for the second administered form to contain no common items with the initial form.

This chapter summarizes the procedures used to construct and evaluate the operational CAT-ASVAB item pools. The first section describes the development of the primary and supplemental item banks. Additional sections discuss dimensionality, alternate form construction, and precision analyses. The final section summarizes important findings with general implications for CAT item pool development.

Development and Calibration

Primary Item Banks

The primary item banks for CAT-ASVAB Forms 1 and 2 were developed and calibrated by Prestwood, Vale, Massey, and Welsh (1985). The P&P-ASVAB Form 8A was used to outline the content of items written in each area. However, important differences between the development of adaptive and conventional (paper-and-pencil) item pools were noted, which led to several modifications in P&P-ASVAB test specifications:

- **Increased range of item difficulties.** Domain specifications were expanded to provide additional easy and difficult items.

- **Functionally independent items**. The Paragraph Comprehension test (as measured in P&P-ASVAB) typically contains reading passages followed by several questions referring to the same passage. Items of these types are likely to violate the assumption of local independence made by the standard unidimensional IRT model. Consequently, CAT-ASVAB items were written to have a single question per passage.

- **Unidimensionality.** In the P&P-ASVAB, auto and shop items are combined into a single test. However, to help satisfy the assumption of unidimensionality, Auto and Shop Information were treated as separate content areas: large non-overlapping pools were written for each, and separate item calibrations were conducted.

About 3,600 items (400 for each of the nine content areas) were written and pretested on a sample of recruits. The pretest was intended to screen about half of the items for inclusion in a large-sample item calibration study. Items administered in the pretest were assembled into 71 booklets, with each booklet containing items from a single content area. Examinees were given 50 minutes to complete all items in a booklet. Data from about 21,000 recruits were gathered, resulting in about 300 responses per item. IRT item parameters were estimated for each item using the ASCAL (Vale & Gialluca, 1985) computer program. (ASCAL is a joint maximum-likelihood/modal-Bayesian item calibration program for the three-parameter logistic item response model.)

For each content area, a subset of items with an approximately rectangular distribution of item difficulties was selected for a more extensive calibration study. This was accomplished from an examination of the IRT difficulty and discrimination parameters. Within each content area, items were divided into 20 equally spaced difficulty levels. Approximately equal numbers of items were drawn from each level, with preference given to the most highly discriminating items.

The surviving 2,118 items (about 235 items per content area) were assembled into 43 P&P test booklets similar in construction to the pretest (each booklet containing items from a single content area, 50 minutes of testing per examinee). Data from 137,000 applicants were collected from 63 Military Entrance Processing Stations (MEPSs) and their associated Mobile Examining Team sites (METSs) during late spring and early summer of 1983. Each examinee was given one experimental form and an operational P&P-ASVAB. After matching booklet and operational ASVAB data, about 116,000 cases remained for IRT calibration analysis (providing about 2,700 responses per item). Within each content area, all experimental and operational P&P-ASVAB items were calibrated jointly using the ASCAL computer program. This helped ensure that the item parameters were properly linked across booklets and provided IRT estimates for several operational P&P-ASVAB forms on a common metric.

Supplemental Item Bank

An analysis of the primary item banks (described below) indicated that two of the content areas, Arithmetic Reasoning (AR) and Word Knowledge (WK), had lower than desired precision over the middle ability range. Therefore, the item pools for these two content areas were supplemented with additional items taken from the experimental CAT-ASVAB system (166 AR items and 195 WK items). Sympson and Hartmann (1985) used a modified version of LOGIST 2.b to calibrate the supplemental items. Data for these calibrations were obtained from a MEPS' administration of P&P booklets. Supplemental item parameters were transformed to the "primary item-metric" using the Stocking and Lord (1983) procedure. The linking design is shown in Table 2-1.

Table 2-1. Linking Design

Calibration	**P&P-ASVAB Form**							
	8A	**8B**	**9A**	**9B**	**10A**	**10B**	**10X**	**10Y**
			Common Forms					
Primary			X	X	X	X	X	X
Supplemental	X	X	X	X	X	X		

The primary calibration included six P&P-ASVAB forms; the supplemental calibration included a different but overlapping set of six P&P-ASVAB forms. The two sets of parameters were linked through the four forms common to both calibrations: 9A, 9B, 10A, and 10B. The specific procedure involved the computation of two test characteristic curves (TCCs), one based on the primary item calibration, and another based on the supplemental item calibration. The linear transformation of the supplemental scale that minimized the weighted sum of squared differences between the two TCCs was computed. The squared differences at selected ability levels were weighted by an $N(0,1)$ density function. This procedure was repeated for both AR and WK. All AR and WK supplemental IRT discrimination and difficulty parameters were transformed to the primary metric, using the appropriate transformation of scale.

Item Reviews

Primary and supplemental items were screened using several criteria. First, an Educational Testing Service (ETS) panel performed sensitivity and quality reviews. The panel recommendations were then submitted to the Service laboratories for their comments. An Item Review Committee made up of researchers at the Navy Personnel Research and Development Center (NPRDC) reviewed the

Service laboratories' and ETS's reports and comments. When needed, the committee was augmented with additional NPRDC personnel having expertise in areas related to the item content under review. The committee reviewed the items and coded them as (a) unacceptable, (b) marginally unacceptable, (c) less than optimal, and (d) acceptable, in each of the two review categories (sensitivity and quality).

Item keys were verified by an examination of point-biserial correlations that were computed for each distracter. Items with positive point-biserial correlations for incorrect options were identified and reviewed.

The display suitability of the item screens was evaluated for (a) clutter (particularly applicable to PC), (b) legibility, (c) graphics quality, (d) congruence of text and graphics (do words and pictures match?), and (e) congruence of screen and booklet versions. In addition, items on the Hewlett Packard Integral Personal Computer (HP-IPC) screen were compared to those in the printed booklets. Displayed items were also examined for (a) words split at the end of lines (no hyphenation allowed), (b) missing characters at the end of lines, (c) missing lines or words, (d) misspelled words, and (e) spelling discrepancies within the booklets. After the items were examined on the HP-IPC, reviewers presented their recommendations to a review group which made final recommendations.

Options Format Study

The primary item pools for AR and WK consisted of multiple-choice items with five response alternatives, while the supplemental items had only four alternatives. If primary and supplemental items were combined in a single pool, examinees would probably receive a mixture of four- and five-choice items during the adaptive test. There was concern that mixing items with different numbers of response options within a test would cause confusion or careless errors by the examinee, and perhaps affect item difficulties.

The authors conducted a study to examine the effect of mixing four- and five-option items on computerized test performance. Examinees in this study were 1,200 male Navy recruits at the Recruit Training Center, San Diego, CA. The task for each examinee was to answer a mixture of four- and five-option items. These included 32 WK items followed by 24 PC items administered by computer using a conventional non-adaptive strategy.

Subjects were randomly assigned to one of six conditions. Specific items administered in each condition for WK are displayed in Table 2-2. Examinees assigned to Conditions A or B received items of one type exclusively: examinees assigned to Condition A received items 1–32 (all five-option items), and examinees assigned to Condition B received items 33–64 (all four-option items). Items in Conditions A and B were selected to span the range of difficulty. Note that four- and five-option items were paired {1,33}, {2,34}, {3,35},... so that items in the same position in the linear sequence would have similar item response functions (and consequently similar difficulty and discrimination levels). Examinees assigned to Condition C received alternating sequences of five- and four-choice items (5, 4, 5, 4,...). Examinees assigned to Condition D received a test in which every fourth item was a four-option item (5, 5, 5, 4, 5, 5, 5, 4,....). In Condition E, every eighth item administered was a four-option item. Finally, in Condition F, an equal number of randomly selected four- and five-option items were administered to each examinee. The first item administered was randomly selected from {1 or 33}, the second item was selected from {2 or 34}, etc. An example assignment for this condition is given in the last column of Table 2-2. Note that for this condition, assignments were generated independently for each examinee. An identical design was used for PC, except that only 24 items were administered to each examinee. Three different outcome measures were examined to assess the effects of mixing item formats: item difficulty, test difficulty, and response latency.

Item Difficulty. For Conditions C, D, E, and F, item difficulties (proportion of correct responses) were compared with those of the corresponding items in the Control Conditions (A or B). For example, comparison of difficulty values in Condition C included pairs {Condition C, Item 1} with {Condition A, Item 1}, {Condition C, Item 34}, etc. The significance of the difference between pairs of item difficulty values was tested using a 2 × 2 chi-square analysis. For WK, only seven of the 160 comparisons (about 4.4%) produced significant differences (at the .05 alpha level). For PC, only one of the 120 comparisons of item difficulty was significant.

Table 2-2. Options Format Study: WK Item Lists Presented in Control and Experimental Conditions

Control		Experimental			
Condition A	**Condition B**	**Condition C**	**Condition D**	**Condition E**	**Condition F**
Five-Option	**Four-Option**	**Mixed: 1:1**	**Mixed: 3:1**	**Mixed: 7:1**	**Random: 1:1**
1	**33**	*1*	*1*	*1*	*1*
2	**34**	**34**	*2*	*2*	*2*
3	**35**	*3*	*3*	*3*	*3*
4	**36**	**36**	**36**	*4*	**36**
5	**37**	*5*	*5*	*5*	**37**
6	**38**	**38**	*6*	*6*	*6*
7	**39**	*7*	*7*	*7*	**39**
8	**40**	**40**	**40**	**40**	**40**
9	**41**	*9*	*9*	*9*	**41**
10	**42**	**42**	*10*	*10*	*10*
11	**43**	*11*	*11*	*11*	**43**
12	**44**	**44**	**44**	*12*	*12*
13	**45**	*13*	*13*	*13*	*13*
14	**46**	**46**	*14*	*14*	*14*
15	**47**	*15*	*15*	*15*	**47**
16	**48**	**48**	**48**	**48**	*16*
17	**49**	*17*	*17*	*17*	**48**
18	**50**	**50**	*18*	*18*	**50**
19	**51**	*19*	*19*	*19*	**51**
20	**52**	**52**	**52**	*20*	*20*
21	**53**	*21*	*21*	*21*	*21*
22	**54**	**54**	*22*	*22*	**54**
23	**55**	*23*	*23*	*23*	**55**
24	**56**	**56**	**56**	**56**	*24*
25	**57**	*25*	*25*	*25*	**57**
26	**58**	**58**	*26*	*26*	**58**
27	**59**	*27*	*27*	*27*	*27*
28	**60**	**60**	**60**	*28*	*28*
29	**61**	*29*	*29*	*29*	*29*
30	**62**	**62**	*30*	*30*	*30*
31	**63**	*31*	*31*	*31*	**63**
32	**64**	**64**	**64**	**64**	**64**

Test Difficulty. For examinees in Conditions C, D, and E, two number-right scores were computed: one based on four-option items, and another based on five-option items. Number-right scores from corresponding items were computed for examinees in the Control conditions A and B. The number of items entering into each score for each condition are displayed in the second and fifth columns of Table 2-3. The significance of the difference between mean number-right scores across the Experimental and Control groups was tested using an independent group *t* statistic. The results are displayed in Table 2-3. None of the comparisons displayed significant results at the .05 alpha level.

Response Latencies. For examinees in Conditions C, D, and E, two latency measures were computed: one based on four-option items, and another based on five-option items. Latency measures were also computed from corresponding items in the Control conditions A and B. Mean latencies were compared across the Experimental and Control groups (Table 2-3). None of the comparisons displayed significant results at the .05 alpha level.

Table 2-3. Options Format Study: Significance Tests for Test Difficulties and Response Latencies

	Word Knowledge			Paragraph Comprehension		
		t-value			*t*-value	
Condition	**No. Items**	**Difficulty**	**Latency**	**No. Items**	**Difficulty**	**Latency**
Comparison with Five-Option Control						
C	16	.06	–.85	12	–.08	–1.77
D	24	–1.09	.47	18	–.21	–.64
E	28	–.24	–.98	21	–1.82	.67
Comparison with Four-Option Control						
C	16	–1.83	1.49	12	1.30	–.72
D	8	–1.35	1.84	6	–.98	–1.92
E	4	1.35	–.07	3	–1.40	–.28

Discussion. Mixing items with different numbers of response options produced no measurable effects on item or test performance. This result differed from those reported by Brittain and Vaughan (1984), who studied the effects of mixing items with different numbers of options on a P&P version of the Army Skills Qualification Test. They predicted errors would increase when an item with *n* answer options followed an item with more than *n* answer options, where errors were defined as choosing non-existent answer options. Consistent with their hypothesis, mixing items with different numbers of answer options caused an increase in errors.

Likely explanations for the different findings between the current study and the Brittain and Vaughan (1984) study involve differences in medium (computer verses P&P). In the Brittain and Vaughan study, examinees answered questions using a standard five-option answer sheet for all items, making the selection of a non-existent option possible. However, in the current study, software features were employed which helped eliminate erroneous responses. (These software features are common to both the current study and the CAT-ASVAB system.)

First, after the examinee makes a selection among response alternatives, he or she is required to confirm the selection. For example, if the examinee selects option "D", the system responds with

```
If “D” is your answer press ENTER.
Otherwise, type another answer.
```

That is, the examinee is informed about the selection that was made and given an opportunity to change the selection. This process would tend to minimize the likelihood of careless errors.

A second desirable feature incorporated into the CAT-ASVAB software (and included in the options format study) was the sequence of events following an "invalid-key" press. Suppose, for example, that a particular item had only four response alternatives (A, B, C, and D) and the examinee selects "E" by mistake. The examinee would see the messages

```
You DID NOT type A, B, C, or D.
Enter your answer (A, B, C, or D).
```

Note that if an examinee accidentally selects a nonexistent option (i.e.,"E"), the item is not scored incorrect; instead, the examinee is given an opportunity to make another selection. This feature would also reduce the likelihood of careless errors. These software features, along with the empirical results of the options format study, addressed the major concerns about mixing four- and five-choice items.

Dimensionality

One major assumption of the IRT item-selection and scoring procedures used by CAT-ASVAB is that performance on items within a given content area can be characterized by a unidimensional latent trait or ability. Earlier research showed that IRT estimation techniques are robust against minor violations of the unidimensionality assumption and that unidimensional IRT parameter estimates have many practical applications in multidimensional item pools (Reckase, 1979; Drasgow & Parsons, 1983, Dorans & Kingston, 1985). However, violations of the unidimensional adaptive testing model may have serious implications for validity and test fairness. Because of the adaptive nature of the test, and the IRT scoring algorithms, multidimensionality may lead to observed scores that represent a different mixture of the underlying unidimensional constructs than intended. This could alter the validity of the test. Furthermore, the application of the unidimensional model to multidimensional item pools may produce differences in the representation of dimensions among examinees. Some examinees may receive items measuring primarily one dimension, while others receive items measuring another dimension. This raises issues of test fairness. If the pool is multidimensional, two examinees (with the same ability levels) may be administered items measuring two largely different constructs and receive widely discrepant scores.

In principle, at least three approaches exist for dealing with multidimensional item pools (Table 2-4). These approaches differ in the item selection and scoring algorithms, and in the item calibration design.

Table 2-4. Treatment Approaches for Multidimensional Item Pools

Approach	Calibration	Item Selection	Scoring
Unidimensional Treatment	Combined calibration containing items of each content type	No constraints placed on item content for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
Content Balancing	Combined calibration containing items of each content type	Constraints placed on the number of items drawn from each content area for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
Pool Splitting	Separate calibrations for items of each content	Separate adaptively tailored tests for each content area	Separate IRT ability estimates for each content area

1. **Unidimensional Treatment.** This option essentially ignores the dimensionality of the item pools in terms of item calibration, item selection, and scoring. A single item calibration containing items spanning all content areas is performed to estimate the IRT item parameters. No content constraints are placed on the selection of items during the adaptive sequence—items are selected on the basis of maximum information. Intermediate and final scoring are performed according to the unidimensional IRT model, and a single score is obtained based on items spanning all content areas.

2. **Content Balancing.** This approach balances the numbers of administered items from targeted content areas. A single item calibration containing items spanning all content areas is performed to estimate the IRT item parameters. During the adaptive test, items are selected from *content-specific* subpools in a

fixed sequence. For example, the content balancing sequence for General Science could be LPLPLPLPLPLPLPL (L = Life Science, P = Physical Science). Accordingly, the first item administered would be selected from among the candidate Life Science items, the second item administered would be selected from the physical science items, and so forth. Within each targeted content area, items are selected on the basis of IRT item information. Intermediate and final scores are based on the unidimensional ability estimator computed from items spanning all content areas.

3. **Pool Splitting.** Item pools for different dimensions are constructed and calibrated separately. For each content area, separate adaptive tests are administered and scored. It is then usually necessary to combine final scores on the separate adaptive tests to form a single composite measure that spans the separately measured content areas.

For each item pool, a number of criteria were considered in determining the most suitable dimensionality-approach, including (a) statistical factor significance, (b) factor interpretation, (c) item difficulties, and (d) factor intercorrelations. The relationship between these criteria and the recommended approach is summarized in Table 2-5.

Table 2-5. Decision Rules for Approaches to Dimensionality

Case	Statistical Factor Sig.	Interpretable Factors	Overlapping Item Difficulties	Factor Correlations	Approach
1	No	—	—	—	Unidimensional
2	Yes	Yes	Yes	High	Content Balance
3	Yes	Yes	Yes	Low	Split Pool
4	Yes	Yes	No	—	Unidimensional
5	Yes	No	Yes	—	Unidimensional
6	Yes	No	No	—	Unidimensional

Statistical Factor Significance

The first, and perhaps most important, criterion for selecting the dimensionality approach is the factor structure of the item pool. If there is empirical evidence to suggest that responses of an item pool are multidimensional, then content balancing or pool splitting should be considered. In the absence of such evidence, item pools should be treated as unidimensional. Such empirical evidence can be obtained from factor analytic studies of item responses using one several available approaches, including TESTFACT (Wilson, Wood, & Gibbons, 1991) and NOHARM (Fraser, 1988). The full item-information procedure used in TESTFACT allows the statistical significance of multidimensional solutions to be tested against the unidimensional solution using a hierarchical likelihood ratio procedure.

All adaptive testing programs do not share this strong empirical emphasis recommended here. The adaptive item-selection algorithm used in the CAT-GRE (Stocking & Swanson, 1993) incorporates both item information and test plan specifications. The test plans are based on expert judgments of content specialists. Accordingly, there is likely to be a disconnect between the test plan specifications and the empirical dimensionality of the item pools. This can lead to situations where constraints are placed on the presentation of items that are largely unidimensional. In general, overly restrictive content-based constraints on item selection will lead to the use of less informative items and, ultimately, to test scores with lower precision.

Factor Interpretation

According to a strictly empirical approach, the number of factors could be determined by statistical considerations, and items could be allocated to areas based on their estimated loadings. Items could be balanced with respect to these areas defined by the empirical analysis. However, a major drawback with this approach is the likelihood of meaningless results, both in terms of the number of factors to be balanced and in the allocation of items to content areas. Significance tests applied to large samples would almost certainly lead to high-dimensionality solutions, regardless of the strength of the factors. Furthermore, there is no guarantee that the rotated factor solution accurately describes the underlying factors.

The alternative judgmental approach noted above would divide the pool into areas on the basis of expert judgments. The major problem with this approach is that without an examination of empirical data, it is not possible to determine which content areas affect the dimensionality of the pool. Choice of content areas could be defined at several arbitrary levels. As Green, Bock, Humphreys, Linn, & Reckase (1982) suggest, “There is obviously a limit to how finely the content should be subdivided. Each item is to a large extent specific."

In CAT-ASVAB development, we formed a decision rule based on a compromise between the empirical and judgmental approaches. If a pool was found to be statistically multidimensional, items loading highly on each factor were inspected for similarity of content. If agreement between factor solutions and content judgments was high, then balancing was considered; otherwise, balancing was not considered.

Item Difficulties

Another important criterion for selecting among dimensionality approaches concerns the overlap of item difficulties associated with items of each content area. The overlap of item difficulties can provide some clues about the causes of the dimensionality and suggest an appropriate remedy. Lord (1977) makes an important observation:

> Suppose, to take an extreme example, certain items in a test are taught to one group of students and not taught to another, while other items are taught to both groups. This way of teaching increases the dimensionality of whatever is measured by the test. If items would otherwise have been factorially unidimensional, this way of teaching will introduce additional dimensions. (p. 24)

If a pool contains some items with material exposed to the entire population (say non-academic content), and other items with material taught to a sub-population (in school—academic content), then we would expect to find statistically significant factors with easy items loading on the non-academic factor and moderate to difficult items loading on the academic factor. Application of the unidimensional item selection and scoring algorithms would result in low- ability examinees receiving easy (non-academic) items and moderate-to-high ability examinees receiving academic items. Thus the unidimensional treatment would appropriately tailor the content of the items according to the standing of the examinee along the latent dimension. Note that content balancing in this situation could substantially reduce the precision of the test scores. For example, if an equal number of items from each content area were administered to each examinee, then low-ability examinees would receive a large number of uninformative difficult items; conversely, high ability examinees would receive a large number of uninformative easy items.

We would expect to observe a different pattern of item difficulty values if substantially non-overlapping subgroups were taught different material. In this instance, we would expect to observe two or more factors defined by items with overlapping difficulty values (falling within a common range). Here, an appropriate remedy would involve content balancing or pool splitting, since different dimensions represent knowledge of somewhat independent domains.

Factor Correlations

A final consideration for selecting among dimensionality approaches concerns the magnitude of the correlation between latent factors. Different approaches might be desirable depending on the correlation between factors estimated in the item factor analysis. If factors are highly correlated, then content balancing may provide the most satisfactory results. In this instance, the unidimensional model used in conjunction with content balancing is likely to provide an adequate approximation for characterizing item information and for estimating latent ability.

If the correlations among factors are found to be low or moderate, then the usefulness of the unidimensional model for characterizing item information and estimating latent abilities is questionable. When the factors have low correlations, pool splitting is likely to provide the best remedy. Separate IRT calibrations should be performed for items of each factor; separate adaptive tests should be administered; and final adaptive test scores can be combined to form a composite measure representing the standing among examinees along the latent composite dimension.

Choosing Among Alternative Approaches

Table 2-5 summarizes different possible outcomes and the recommended approach for each. If an item factor analysis provides no significant second, or higher order, factors, then the pool should be treated as unidimensional (Case 1). If statistically significant higher order factors are identified (these factors relate to item content), and item difficulties of each content span a common range, then consideration should be given to content balancing (Case 2, if the factor intercorrelations are high) or to pool splitting (Case 3, if the factor intercorrelations are low to moderate). For reasons given above, if the statistical factors are not interpretable (Cases 5 and 6), or if the item difficulty values of each content area span non-overlapping ranges (Cases 4 and 6), then unidimensional treatment may provide the most useful approach.

Results and Discussion

In earlier studies of the Auto-Shop content area, a decision was made to apply the pool-splitting approach; this content area was split into separate auto and shop item pools (Case 3, Table 2-5). As described in an earlier section, these pools were calibrated separately. The decision to split these pools was based on the moderately high correlation among the auto and shop dimensions. In the analysis described below, the auto and shop pools were examined separately and subjected to the same analyses as other pools.

The first step in the dimensionality analysis involved factor analyses using item data (Prestwood, Vale, Massey & Welsh, 1985). Empirical item responses were analyzed using the TESTFACT computer program (Muraki, 1984) which employs full information item factor analysis based on IRT (Bock & Aitkin, 1981). While the program computes item difficulty and item discrimination parameters, guessing parameters are treated as known constants and must be supplied to the program. For these analyses, the guessing parameters estimated by Prestwood et al. were used. For all analyses, a maximum of four factors were extracted, using a stepwise procedure. An item pool was considered statistically multidimensional if a change in chi-square (between the one-factor solution and the two-factor solution) was statistically significant (at the .01 alpha level). If the change in chi-square for the two-factor solution was significant, the three- and four-factor solutions were also examined for significant changes in chi-square. Since items within a pool were divided into separate booklets for data collection purposes, all items within a pool could not be factor analyzed at once. Therefore, subsets of items (generally, all items in one booklet) were analyzed. The number of statistically significant factors found across booklets was not necessarily identical. In such cases, the factor solutions examined were the number found in the majority of the booklets. The number of statistically significant factors found for each item pool is summarized in Table 2-6. For those item pools showing statistical evidence of multidimensionality, items were reviewed to determine whether the pattern of factor loadings was related to content, mean difficulty parameters were computed by content area, and factor intercorrelations were examined. These results are displayed in Table 2-6.

Table 2-6. Dimensionality of CAT-ASVAB Item Pools

Item Pool	No. Significant Factors	Interpretable Factors	Overlapping Item Difficulties	Factor Correlations	Case	Approach
GS	4	Yes	Yes	High	2	Content Bal.
AR	2	Yes	No	—	4	Unidimensional
WK	2	Yes	No	—	4	Unidimensional
PC	1	—	—	—	1	Unidimensional
AI	2	Yes	No	—	4	Unidimensional
SI	2	Yes	No	—	4	Unidimensional
MK	4	No	Yes	—	5	Unidimensional
MC	1	—	—	—	1	Unidimensional
EI	2	Yes	No	—	4	Unidimensional

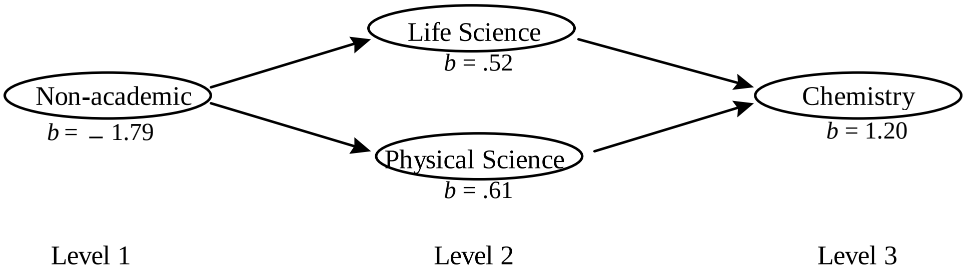
Based on the factor analyses, PC and MC were found to be unidimensional (Case 1, Table 2-5). All other item pools were multidimensional, with GS and MK having four factors and AR, WK, AI, SI, and EI each having two factors. For those areas having two factors, the pattern of factor loadings was readily apparent. Items that loaded highly on the first factor were non-academic items (i.e., taught to the whole group through everyday experiences). Items that loaded highly on the second factor were academic items (i.e., taught to a subgroup through classroom instruction or specialized experience). Means of IRT difficulty parameters for academic and non-academic items are displayed in Table 2-7. As indicated, the mean difficulty values for non-academic items were much lower than those for academic items. Accordingly, AR, WK, AI, SI, and EI were treated as unidimensional item pools (Case 4, Table 2-5).

Table 2-7. Mean IRT Item Difficulty (*b*) Parameters

Item Content	AR	WK	AI	SI	EI
Non-academic	–2.37	–2.30	–2.28	–2.15	–1.51
Academic	.30	.47	.48	.57	.61

The GS pool appeared, in part, to follow a different pattern than the five pools discussed above. An examination of the factor solutions and item content provided some evidence for a four-factor solution interpreted as (a) non-academic, (b) life science, (c) physical science, and (d) chemistry. This interpretation is supported by the fact that many high schools offer a multiple-track science program (Figure 2-1). At Level 1, students have little or no formal instruction. At Level 2, some students receive training in life science, while others receive physical science training. Finally, at Level 3, some members of both groups are instructed in chemistry. Notice that each higher level contains only a subset of students contained in the levels directly below it. For example, not everyone completing a life science or a physical science course will receive instruction in chemistry. The mean IRT item difficulty values (displayed in Figure 2-1) also support this interpretation of dimensionality. The life science and physical science items are of moderate (and approximately equal) difficulty. The chemistry items appear to be the most difficult and non-academic items least difficult. These findings support balancing content among life and physical science items (Case 2, Table 2-5). Non-academic and chemistry items should be administered to examinees of appropriate ability levels. (See Chapter 3 in this technical bulletin for additional details on the GS content balancing algorithm.)

Figure 2-1
General Science Dual Track Instruction



For MK, the pattern of factor loadings associated with the two-, three-, or four-factor solutions could not be associated with item content. Consequently, the MK item pool was treated as unidimensional (Case 5, Table 2-5).

Alternate Forms

In developing the item pools for CAT-ASVAB, it was necessary to create two alternate test forms so applicants could be re-tested on another form of CAT-ASVAB. Once the item-screening procedures were completed, items within each content area were assigned to alternate pools. Pairs of items with similar information functions were identified and assigned to alternate pools. The primary goal of the alternate form assignment was to minimize the weighted sum-of-squared differences between the two pool information functions. (A pool information function was computed from the sum of the item information functions.) The squared differences between pool information functions were weighted by an *N*(0,1) density.

The procedure used to create the GS alternate forms differed slightly from the other content areas because of the content balancing requirement. GS items were first divided into physical, life, and chemistry content areas. Domain specifications provided by Prestwood, Vale, Massey, & Welsh (1985) were used for assignment to these content areas. Once items had been assigned to a content area, alternate forms were created separately for each of the three areas.

Precision Analyses

Precision is an important criterion for judging the adequacy of the items pools, since it depends in large part on the quality of the pools. Precision analyses were conducted separately for the 22 item pools displayed in Table 2-8. The content area and form are listed in columns two and four. The target exposure rate (for the battery, i.e, across the two forms) is provided in the last column. This target was used to compute exposure-control parameters according to the Sympson-Hetter algorithm (Chapter 4 in this technical bulletin). The fifth column shows whether the pool included supplemental items. The third column provides a descriptive label for each condition used in the text and tables.

As would be expected, the results of any precision analysis would show various degrees of precision among the CAT-ASVAB tests. But how much precision is enough? The precision of the P&P-ASVAB offers a useful baseline. It is desirable for CAT-ASVAB to match or exceed P&P-ASVAB precision. Accordingly, precision criteria were computed for both P&P-ASVAB and CAT-ASVAB.

It is important to evaluate the impact of using the CAT-ASVAB item selection and scoring algorithm on precision, since the precision of adaptive test scores depends on both (a) the quality of the item pools, and (b) the adaptive testing procedures. The specific item selection and scoring procedures used are described in Chapter 3 of this technical bulletin. For each adaptively administered test, the precision of the Bayesian modal estimate was evaluated. For each item pool, two measures of precision were examined: (a) score information, and (b) reliability.

Table 2-8. Item Pools Evaluated in Precision Analyses

Condition	Con-tent Area	Label	For m	Supplemented	Target Exposure Rate
1	GS	GS-1	1	No	1 / 3
2	GS	GS-2	2	No	1 / 3
3	AR	AR-1	1	No	1 / 6
4	AR	AR-2	2	No	1 / 6
5	AR	AR_s-1	1	Yes	1 / 6
6	AR	AR_s-2	2	Yes	1 / 6
7	WK	WK-1	1	No	1 / 6
8	WK	WK-2	2	No	1 / 6
9	WK	WK_s-1	1	Yes	1 / 6
10	WK	WK_s-2	2	Yes	1 / 6
11	PC	PC-1	1	No	1 / 6
12	PC	PC-2	2	No	1 / 6
13	AI	AI-1	1	No	1 / 3
14	AI	AI-2	2	No	1 / 3
15	SI	SI-1	1	No	1 / 3
16	SI	SI-2	2	No	1 / 3
17	MC	MC-1	1	No	1 / 3
18	MC	MC-2	2	No	1 / 3
19	MK	MK-1	1	No	1 / 6
20	MK	MK-2	2	No	1 / 6
21	EI	EI-1	1	No	1 / 3
22	EI	EI-2	2	No	1 / 3

Score Information

Score information functions provide one criterion for comparing the relative precision of the CAT-ASVAB with the P&P-ASVAB. Birnbaum (1968, Section 17.7) defines the information function for any score y to be

$$I\{\theta, y\} \equiv \frac{\left(\dfrac{d}{d\theta}\mu_{y|\theta}\right)^2}{\mathrm{Var}(y|\theta)} \quad . \tag{2-1}$$

This function is by definition inversely proportional to the square of the length of the asymptotic confidence interval for estimating ability θ from score y. For each content area, information functions can be compared between the CAT-ASVAB and the P&P-ASVAB. The test with greater information at a given ability level will possess a smaller asymptotic confidence interval for estimating θ.

CAT-ASVAB Score Information Functions. The score information functions (SIFs) for each CAT-ASVAB item pool were approximated from simulated test sessions. For a given pool, simulations were repeated independently for 500 examinees at each of 31 different θ levels. These θ levels were equally spaced along the [–3, +3] interval. At each θ level, the mean m and variance s^2 of the 500 final scores were computed. The information function at each selected level of θ can be approximated from these results, using (Lord, 1980, eq. 10-7)

$$I\{\theta,\hat{\theta}\} \approx \frac{[m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1})]^2}{(\theta_{+1}-\theta_{-1})^2 s^2(\hat{\theta}|\theta_0)} \tag{2-2}$$

where θ_{-1}, θ_0, θ_{+1} represent the successive levels of θ. However, the curve produced by this approximation often appears jagged, with many local variations. To reduce this problem, information was approximated by

$$I\{\theta,\hat{\theta}\} \approx \frac{\left[\dfrac{m(\hat{\theta}|\theta_{+1}) + m(\hat{\theta}|\theta_{+2})}{2} - \dfrac{m(\hat{\theta}|\theta_{-1}) + m(\hat{\theta}|\theta_{-2})}{2}\right]^2}{\left[\dfrac{\theta_{+1}+\theta_{+2}}{2} - \dfrac{\theta_{-1}+\theta_{-2}}{2}\right]^2 \left[\dfrac{1}{5}\sum_{k=-2}^{+2} s(\hat{\theta}|\theta)\right]^2} \tag{2-3}$$

$$= \frac{25\left[m(\hat{\theta}|\theta_{+2}) + m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1}) - m(\hat{\theta}|\theta_{-2})\right]^2}{(\theta_{+2}+\theta_{+1}-\theta_{-1}-\theta_{-2})^2\left[\sum_{k=-2}^{+2} s(\hat{\theta}|\theta_k)\right]^2} \tag{2-4}$$

where θ_{-2}, θ_{-1}, θ_0, θ_{+1}, θ_{+2} represent successive levels of θ. This approximation results in a moderately smoothed curve with small local differences.

P&P-ASVAB Score Information Functions. The P&P-SIF for a number right score *x* was computed by (Lord, 1980, eq. 5-13)

$$I\{\theta, x\} = \frac{\left[\sum_{i=1}^{n} P_i'(\theta)\right]^2}{\sum_{i=1}^{n} P_i(\theta) Q_i(\theta)}$$

(2-5)

This function was computed for each content area by substituting the estimated P&P-ASVAB (9A) parameters for those assumed to be known in Equation 2-5.

A special procedure was used to compute the SIF for AS since that test is represented by two tests in CAT-ASVAB. The AS-P&P (9A) test was divided into AI and SI items. SIFs (eq. 2-5) were computed separately for these AI-P&P and SI-P&P items to simplify comparisons with the corresponding CAT-ASVAB SIFs. Parameters used in the computation of these SIFs were taken from the joint calibrations of P&P-ASVAB and CAT-ASVAB items. In these calibrations, AS-P&P items were separated and calibrated among CAT-ASVAB items of corresponding content (i.e., AI-P&P items were calibrated with AI-CAT, and SI-P&P with SI-CAT items). However, two AS-P&P (9A) items appeared to overlap in AI/SI content and appeared in both AI and SI calibrations. For computations of score information, these two items were included in both AI-P&P and SI-P&P information functions. This represents a conservative approach (favoring the P&P-ASVAB), since we are counting these two items twice in the computations of the P&P-ASVAB SIFs.

Score Information Results. CAT-ASVAB SIFs were computed for each of the 22 conditions listed in Table 2-8. For comparison, the P&P-ASVAB SIF (for 9A) was computed. The SIFs for the CAT-ASVAB equaled or exceeded the P&P-ASVAB SIFs for all but four conditions: 3, 4, 7, and 8. These four exceptions involved the two pools of AR and WK that consisted of only primary items. When these pools were supplemented with additional items (see conditions 5, 6, 9, and 10), the resulting SIFs equaled or exceeded the corresponding P&P-ASVAB SIFs.

Table 2-9 lists the number of items used in selected SIF analyses. The number of times (across simulees) that an item was administered was recorded for each SIF simulation. The values in Table 2-9 represent the number of items that were administered at least once during the 15,500 simulated test sessions. A separate count for primary and supplemental items is provided for AR and WK.

Table 2-9. Number of Used Items in CAT-ASVAB Item Pools

		Number of Used Items					
		Form 1			Form 2		
Content Area	Exposure Rate	Primary	Supp.	Total	Primary	Supp.	Total
GS	1/3	72	—	72	67	—	67
AR	1/6	62	32	94	53	41	94
WK	1/6	61	34	95	55	44	99
PC	1/6	50	—	50	52	—	52
AI	1/3	53	—	53	53	—	53
SI	1/3	51	—	51	49	—	49
MK	1/6	84	—	84	85	—	85
MC	1/3	64	—	64	64	—	64
EI	1/3	61	—	61	61	—	61

Reliability

A reliability index provides another criterion for comparing the relative precision of the CAT-ASVAB with the P&P-ASVAB. These indices were computed for each pool and for one form (9A) of the P&P-ASVAB. The reliabilities were estimated from simulated test sessions: 1,900 values were sampled from an $N(0,1)$ distribution. Each value represented the ability level of a simulated examinee (simulee). The simulated tests were administered twice to each of the 1,900 simulees. The reliability index was the correlation between the pairs of Bayesian modal estimates of ability from the two simulated administrations. The CAT-ASVAB reliabilities were computed separately for each pool. The item selection and scoring procedures match those used in CAT-ASVAB (Chapter 3 in this technical bulletin).

The P&P-ASVAB reliabilities were computed from simulated administrations of Form 9A. The following procedure was used to generate number right scores for each of the 1,900 simulees:

STEP 1: The probability of a correct response to a given item was obtained for a simulee by substituting the (9A) item parameter estimates and the simulee's ability level into the three-parameter logistic model.

STEP 2: A random uniform value in the interval [0,1] was generated and compared to the probability of a correct response. If the random number was less than the probability value, the item was scored correct; otherwise, it was scored incorrect.

STEP 3: Steps 1 and 2 were repeated across test items for each simulee. The number right score was the sum of the responses scored correct.

Steps 1 through 3 were repeated twice to obtain two number-right scores for each simulee. The reliability index for the P&P-ASVAB was the correlation between the two number-right scores.

A special procedure was used to compute reliability indices for AS. These items on the P&P version (9A) were divided into two components: AI and SI. This split corresponded to the assignment made in the item calibration of these content areas. A reliability index was computed separately for each component.

Reliability indices were computed for each of the 22 conditions and are listed in Table 2-10. For comparison, the P&P-ASVAB reliability (for 9A) was computed and displayed in the same table. Exposure rates and test lengths are also provided. The estimated CAT-ASVAB reliability indices exceeded the corresponding P&P-ASVAB (9A) values for all 22 conditions.

Table 2-10. Simulated CAT-ASVAB Reliabilities (N=1,900)

Test	Form	Test Length	Exposure Rate	Reliability *r*
GS	CAT-1	15	1/3	.902
	CAT-2	15	1/3	.900
	ASVAB-9A	25		.835
AR	CAT-1	15	1/6	.904
	CAT-2	15	1/6	.903
	CAT_s-1	15	1/6	.924
	CAT_s-2	15	1/6	.924
	ASVAB-9A	30		.891
WK	CAT-1	15	1/6	.912
	CAT-2	15	1/6	.913
	CAT_s-1	15	1/6	.934
	CAT_s-2	15	1/6	.936
	ASVAB-9A	35		.902
PC	CAT-1	10	1/6	.847
	CAT-2	10	1/6	.855
	ASVAB-9A	15		.758
AI	CAT-1	10	1/3	.894
	CAT-2	10	1/3	.904
	ASVAB-9A	17		.821
SI	CAT-1	10	1/3	.874
	CAT-2	10	1/3	.873
	ASVAB-9A	10		.651
MK	CAT-1	15	1/6	.933
	CAT-2	15	1/6	.935
	ASVAB-9A	25		.854
MC	CAT-1	15	1/3	.886
	CAT-2	15	1/3	.897
	ASVAB-9A	25		.807
EI	CAT-1	15	1/3	.875
	CAT-2	15	1/3	.873
	ASVAB-9A	20		.768

Summary

The procedures described in this chapter formed the basis of the item pool construction and evaluation procedures. Large item pools were pre-tested and calibrated in large samples of applicants. Two item pools (WK and AR) were supplemented with additional items, and a special study was conducted to evaluate adverse consequences of mixing four-option supplemental items with other five-option items. Extensive analyses were conducted to evaluate each pool's dimensionality. For pools found to be multidimensional, these analyses aided in selecting the most appropriate approach for item selection and scoring. Finally, extensive precision analyses were conducted to (a) evaluate the conditional and unconditional precision levels of the item pools, and (b) compare these precision levels with the P&P-ASVAB.

Based on the score information analyses, the precision for the primary AR and WK pools over the middle ranges of ability was inadequate. By supplementing these pools with experimental CAT-ASVAB items, the precision was raised to an acceptable level. Why was it necessary to supplement these pools, and what lessons can be applied to the construction of future pools?

One clue comes from the distribution of difficulty parameters obtained from surviving items (those items in the pools that have a greater than zero probability of administration). An examination of this distribution indicates a bell shaped distribution, with a larger number of difficulty values appearing over the middle ranges and fewer values appearing in the extremes. Note that the target difficulty distribution for item writing and for inclusion in the calibration study was a uniform distribution. This suggests that there was actually an excess of items in the extremes (which had zero probabilities of administration), but for WK and AR there was a deficiency of items over the middle ranges. Future development efforts should attempt to construct banks of items with bell-shaped distributions of item difficulty values, similar to those constructed for P&P tests.

A bell-shaped distribution of item difficulties has at least two desirable properties for CAT. First, larger numbers of items with moderate difficulty values are likely to lead to higher precision over the middle range, since the adaptive algorithm is likely to have more highly discriminating items to choose from. This may be especially desirable if it is important to match the precision of a P&P test that peeks in information over the middle ability ranges. Second, the Sympson-Hetter exposure-control algorithm (Chapter 4 of this technical bulletin) places demands on moderately difficult items, since the administration of these items is restricted. Because of the restrictions placed on these items, more highly informative items of moderate difficulty are necessary to maintain high levels of precision.

Although CAT-ASVAB precision analyses indicated favorable comparisons with the P&P-ASVAB, many strong assumptions were made in the simulation analyses that may limit applicability of these findings to operational administrations with real examinees. Such assumptions (including unidimensionality, local independence, and knowledge of true item functioning) are almost certainly violated to some extent in applied testing situations. Therefore, it is important to examine the precision of these pools with live examinees who are administered tests using the same adaptive item selection and scoring algorithms evaluated here. Such an evaluation is described in Chapter 7 in this technical bulletin.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46,* 443-459.

Brittain, C. V., & Vaughan, P. R. (1984). Effects of mixing two-option and multi-option items in a test. *Proceedings of the 26th Annual Conference of the Military Testing Association Volume I* (369–374). Munich, Federal Republic of Germany: Psychological Service of the German Federal Armed Forces.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22,* 249-262.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Fraser, C. (1988). *NOHARM II. A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England, Center for Behavioral Studies.

Green, B. F., Jr., Bock R. D., Humphreys, L. G., Linn, R. L, & Reckase, M.D. (1982). *Evaluation plan for the Computerized Adaptive Armed Services Vocational Aptitude Battery* (NTIS No. AD-A115 334). Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Lord, F. M. (1977). A broad range tailored test of verbal ability. *Applied Psychological Measurement,* 95-100.

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

Muraki, E. (1984). *Implementing full-information factor analysis: TESTFACT program*. A paper presented at the annual meeting of Psychometric Society, University of California, Santa Barbara, CA.

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). Brooks AFB, TX: Air Force Human Resources Laboratory. (AD-A160 608)

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 1,* 201-210.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17.*

Sympson, J. B., & Hartmann, L. (1985). Item calibrations for computerized adaptive testing (CAT) item pools. In D.J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis, MN: Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota. (NTIS No. AD-A163 040)

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. (NTIS No. AD-A169 737)

Wilson, D .T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, Item Statistics, and Item Factor Analysis*. Scientific Software International, Inc. Chicago, IL.

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. 97-102). Washington, DC: American Psychological Association.

Chapter 3

PSYCHOMETRIC PROCEDURES FOR ADMINISTERING CAT-ASVAB

This chapter describes the psychometric procedures used in the administration and scoring of the computerized-adaptive testing version of the Armed Forces Vocational Aptitude Battery (CAT-ASVAB) and summarizes the rationale for selecting these procedures. Key decisions were based on extensive discussions from the mid- to late-1980s by the staff at the Navy Personnel Research and Development Center (NPRDC) and by the CAT-ASVAB Technical Committee. For many key psychometric decisions, there was an understandable tension between two camps within the CAT-ASVAB project. One camp wanted to extensively study each decision, first by reviewing the literature, then by carefully enumerating all possible alternatives, then by studying empirically all possible alternatives from carefully designed and implemented research studies, and then, and only then, choosing from among the alternatives. The other camp was less concerned with making optimal decisions and more concerned with the efficient allocation of resources needed to field an operational system. The tension between these two camps produced an adaptive testing battery (CAT-ASVAB) that achieved a remarkable balance between scientific empiricism and the drive to produce an operational system.

The experimental system (McBride, Wetzel, & Hetter, 1997; (Wolfe, McBride, & Sympson, 1997; Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997) provided a useful and important starting point for the specification of psychometric procedures. By the mid-1980s, data from over 7,500 subjects had been collected and analyzed. These data, to a large extent, supported the usefulness of many experimental system procedures; validities for predicting success in training were as high or higher than those with the paper-and-pencil version of the ASVAB (P&P-ASVAB). However, the absence of many necessary features (test time-limits, help dialogs, item seeding, stringent exposure control, and user-friendly rules for changing and confirming answers) meant that extensive psychometric changes would be required before CAT-ASVAB could be administered operationally.

From about 1985 to 1989, NPRDC and the CAT-ASVAB Technical Committee conducted an extensive review of CAT-ASVAB psychometric procedures. Virtually every characteristic of the system having psychometric implications was studied. Because of the necessary time and resource constraints, different decisions were based on different amounts of knowledge and understanding of each issue. Many important decisions were based on extensive empirical studies conducted by project staff using live or simulated data. Other decisions were based on existing work reported in the literature. And still other choices fell into the “it doesn’t matter” category. In documenting the psychometric procedures of the CAT-ASVAB, examples of each type can be found. Although not all decisions were based on a complete and thorough investigation of the issues, it is a tribute to those involved that the fundamental decisions made during this period have withstood the test of time. In this chapter, three major areas are discussed: power test administration, speeded test administration, and administrative requirements.

Power Test Administration

All power tests contained in the CAT-ASVAB are administered using an adaptive testing algorithm. The eight basic steps involved in item selection and scoring are displayed in Figure 3-1. Details of each step are provided below.

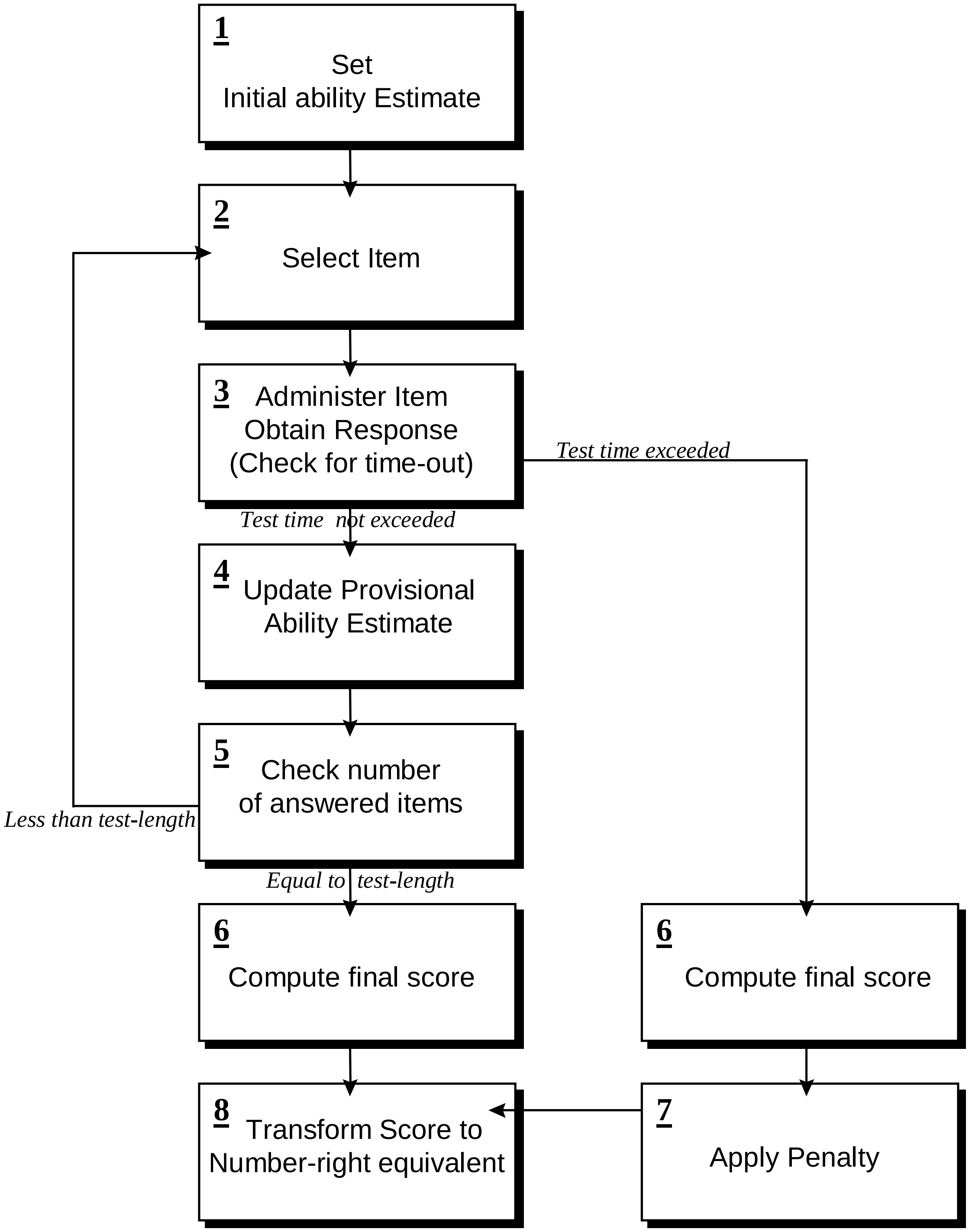
1. Initial Ability Specification

The first step in item selection is to set the initial ability estimate $\hat{\theta}_0 = 0$ (i.e., equal to the mean of the prior distribution of abilities). The mean and standard deviation of the prior were set equal to the observed moments of IRT scores (Bayesian modal estimates) calculated from the calibration sample used to estimate IRT item parameters (Chapter 2 in this technical bulletin). By specifying the initial ability estimate in this way, the first administered item will be among the most informative for average ability examinees.

2. Item Selection

Given an initial or provisional ability estimate, the second step in the adaptive algorithm is to choose the next item for presentation to the examinee. CAT-ASVAB uses item response theory (IRT) item information (Lord, 1980a, eq. 5-9) as a basis for choosing items. Selecting the most informative item for an examinee is accomplished by the use of an information table. To create the tables for each content area, items were sorted by information at each of 37 θ levels, equally spaced along the interval [–2.25, +2.25]. The use of information tables avoids the necessity for computing information values for each item in the pool between the presentation of successive items; these values are essentially computed

Figure 3-1
Steps in CAT-ASVAB Item Selection and Scoring


1 Set Initial ability Estimate
2 Select Item
3 Administer Item Obtain Response (Check for time-out)
Test time exceeded
Test time not exceeded
4 Update Provisional Ability Estimate
5 Check number of answered items
Less than test-length
Equal to test-length
6 Compute final score
6 Compute final score
8 Transform Score to Number-right equivalent
7 Apply Penalty

in advance. The General Science test is content-balanced among three content areas due to concerns about dimensionality. For this test, separate information tables were created for each of the three content areas: Life Science, Physical Science, and Chemistry.

An item is chosen from the appropriate information table, and selection is based on the provisional ability estimate (denoted by $\hat{\theta}_n$) calculated from the *n* previously answered items. The mid-point of the θ interval in the information table closest to the provisional estimate is located, and items with the greatest information in that θ interval are considered, in turn, for administration. The selection of an item within a given θ interval of the information table is subject to two criteria. First, the item must not have been previously administered to the examinee during the test session. Second, item selection is conditional on the application of the exposure-control procedure (see Chapter 4 in this technical bulletin). According to this exposure-control algorithm, once an item is considered for administration, the system generates a random number between 0 and 1 and compares this random number to the exposure-control parameter for the item. If the value of the exposure control parameter is greater than, or equal to, the random number for the item, the item is administered. If the value of the exposure-control parameter is less than the random number, the item is not administered; it is marked as having been selected, and it is not considered for administration at any other point in the test for that examinee. In this case, the next most informative item in the interval is considered for administration, and a new random number is generated. This process is repeated until an item passes the exposure-control screen. This procedure places a ceiling on the exposure of the pool's most informative items.

The General Science test follows this same procedure, except that the allocation administers roughly the same proportion of each content area as found in the reference P&P ASVAB form (8A). The following allocation vector is used to determine the information table from which to select the next item:

$$L, P, L, P, L, P, L, P, L, P, L, P, L, P, L, C,$$

where L = Life Science, P = Physical Science, C = Chemistry. Accordingly, the first item administered in the General Science test is selected from the Life Science information table, the second item administered is selected from the Physical Science information table, and so on, with only one Chemistry item selected for an examinee.

3. Item Administration

Once the item has been selected, the third step is to display the item and obtain the examinee’s response. The administrative requirements involved in item presentation and gathering responses are described in a following section. Each adaptive test has an associated time limit (Table 3-1). If this time limit is reached before the examinee has answered the last item, the test is terminated, a final score is computed (Figure 3-1, Step 6), and a scoring penalty is applied (Figure 3-1, Step 7).

Ideally, pure power tests should be administered without time limits. This is especially true of adaptive power tests that are scored using IRT methods that do not explicitly consider the effects of time pressure on response choice. However, the imposition of time limits on all tests was necessary for administrative purposes. When scheduling test sessions and paying test administrators, it would not be practical to allow some examinees to take as long as desired. The power test time limits were initially based on response times of recruits in a Joint-Service validity study (see Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997). Those time limits were later modified from test finishing times gathered from about 400 applicants participating in the Score Equating Development (SED) study (Chapter 8 in this technical bulletin). The time limits were set so that over 95 percent of the examinees taking the test would complete all items without having to rush. In practice, each adaptive test displays completion rates of over 98 percent.

Table 3-1. Time Limits (minutes) and Test Lengths* for CAT-ASVAB Tests

	GS	AR	WK	PC	NO	CS	AI	SI	MK	MC	EI
Time limit	8	39	8	22	3	7	6	5	18	20	8
Test length	16	16	16	11	50	84	11	11	16	16	16

*For all power tests, the test lengths include one experimental item. Therefore, the number of items used to score the test is the test length minus one.

4. *Provisional Scoring*

After the presentation of each item, the scored response is used to update the provisional ability estimate. A sequential Bayesian procedure (Owen, 1969; 1975) is used for this purpose. This updated ability estimate is used to select the next item for administration (Figure 3-1, Step 2). This procedure was selected for intermediate scoring because it is computationally efficient compared to other Bayesian estimators and because it provided favorable results in empirical validity studies (Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997).

5. Test Termination

Each CAT-ASVAB test is terminated after an examinee has completed a fixed number of items or reaches the test time limit, whichever occurs first. The fifth step in the adaptive algorithm is to check to determine if the examinee has answered the prescribed number of items for the test (Table 3-1). If the examinee has, then a final score is computed (Step 6); otherwise, a new item is selected (Step 2) and administered (Step 3).

A number of rationales support the decision to use fixed-length testing in CAT-ASVAB, as opposed to variable-length testing in which additional items are administered until a pre-specified level of precision has been obtained. First, simulation studies have shown that fixed-length testing is more efficient than variable-length testing. Highly informative items are typically concentrated over a restricted range of ability. In variable-length testing, examinees falling outside this range tend to receive long tests, with each additional item providing very little information. For these examinees (usually at the high- and low-ability levels), the incremental value of each additional item quickly reaches the point of diminishing returns, leading to a very inefficient use of the examinees' time and effort. Also, with fixed-length testing, test-taking time is less variable across examinees, making the administration of the test and the planning of post-testing activities more predictable. Administering the same number of items to all examinees avoids the public-relations problem of explaining to non-experts why different numbers of items were administered.

6. Final Scoring

A final Owen's estimate can be obtained by updating the estimate with the response to the final test item. However, the Owen's estimate, as a final score, has one undesirable feature: the final score depends on the order in which the items are administered. Consequently, it is possible for two examinees to receive the same items and provide the same responses but receive different final Owen's ability estimates; this could occur if the two examinees received the items in different sequences. To avoid this possibility, the mode of the posterior distribution (Bayesian mode) is used at the conclusion of each power test to provide a final ability estimate. This estimator is unaffected by the order of item administration and provides slightly greater precision than the Owen's estimator.

In selecting a procedure for computing the final ability estimate, various alternatives were considered. The posterior mode was chosen for the following reasons:

- Although the posterior median gives estimates that are slightly more precise in simulations, the posterior mode is more established in the research literature.

- After transformation to the number-right metric, the score based on the posterior mode correlates .999-1.000 with the posterior mean number right obtained by numerical integration.

- Iterative computation of the posterior mode (with Owen's approximation to the posterior mean as the initial estimate) is more rapid than computation of the posterior mean obtained by adaptive quadrature numerical integration.

- Maximum likelihood (ML) estimation was not used because of the possible bimodality of the likelihood function, and also because it is undefined for all correct or incorrect response patterns. Also, ML estimates had lower validity for predicting success in training. This latter result was obtained by recomputing final scores with ML estimates for subjects participating in a Joint-Service validity study (Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997) and by computing the corresponding validity coefficients. These values were lower than validity coefficients computed from final scores based on Bayesian procedures.

7. Penalty for Incomplete Tests

The Bayesian modal estimator (BME) has one property that is problematic in the context of incomplete tests. As with Bayesian estimators in general, the BME contains a bias that draws the estimate toward the mean of the prior. This bias is inversely related to test length. That is, the bias is larger for short adaptive tests and smaller for long adaptive tests. A low-ability examinee could use this property to his or her advantage. If allowed, a low-ability examinee could obtain a score at, or slightly below, the mean by answering only one or two items. Even if the items were answered incorrectly, the strong positive bias would push the estimator up toward the mean of the prior. Consequently, below-average applicants could use this strategy to increase their score by answering the minimum number of items allowed.

To discourage the use of this strategy, a penalty procedure was developed for use in scoring incomplete tests (Segall, 1988). The fact that the tests are timed almost ensures that some examinees will not finish, whether intentionally or not. In general, it is desirable for a penalty procedure to have the following properties:

- *The size of the penalty should be related to the number of unfinished items*. That is, applicants with many unfinished items should generally receive a more severe penalty than applicants with one or two unfinished items.

- *Applicants who have answered the same number of items and have the same provisional ability estimate should receive the same penalty.*

- *The penalty rule should eliminate "coachable" test-taking strategies (with respect to answering or not answering test items).*

The penalty procedure used in CAT-ASVAB satisfies the above constraints by providing a final score that is equivalent (in expectation) to the score obtained by guessing at random on the unfinished items. The sizes of the penalties for different test lengths, tests, and ability levels were determined through a series of 240 simulations. The following example provides the basic steps used in determining penalty functions.

Example Penalty Simulation:
Electronics Information Form 2
Penalty for two unanswered items

- Sample 2,000 true abilities from the uniform interval [-3, +3].

- For each simulee, generate a 13-item adaptive test; obtain a provisional score on the 13-item test with the BME, denoted as $\hat{\theta}_{13}$.

- For each simulee, provide random responses for the remaining two items, with the probability of a correct response equal to *p* = .2; then re-score using all 15 responses with the BME. Denote this final estimate as $\hat{\theta}_{15}$.

- Regress $\hat{\theta}_{15}$ on $\hat{\theta}_{13}$, and fit a least-squares line predicting $\hat{\theta}_{15}$ from $\hat{\theta}_{13}$. This regression equation becomes the penalty function for EI Form 2 with 13 answered items.

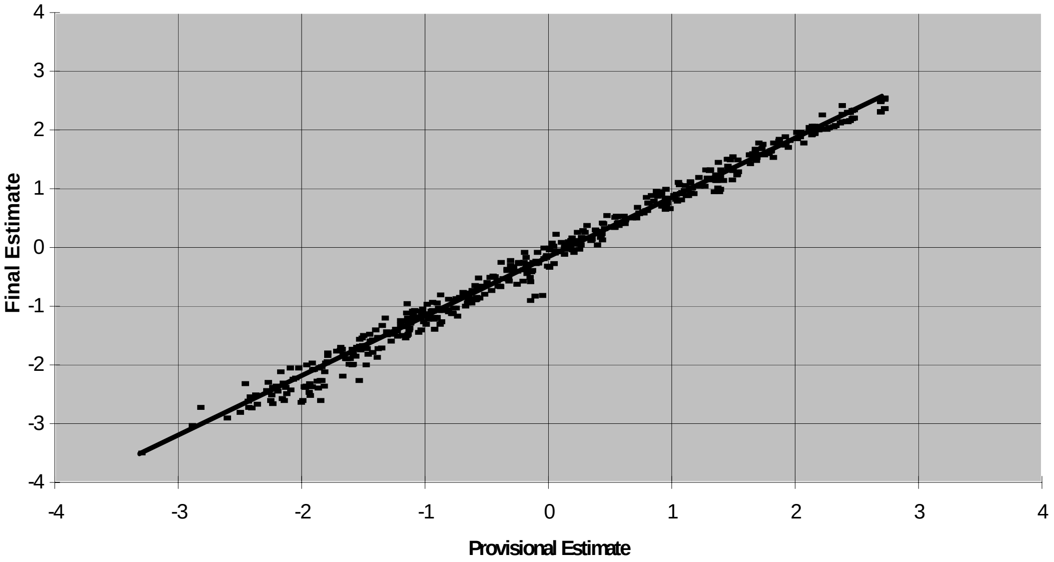


Figure 3-2. Penalty Function for EI Form 2.

Figure 3-2 displays the outcome of this last step. By regressing the final estimate $\hat{\theta}_{15}$ on the provisional estimate $\hat{\theta}_{13}$, we can obtain an expected penalized $\dot{\theta}$ for any provisional $\hat{\theta}_{13}$. The final results of the simulation are slope and intercept parameters for the penalty function

$$\dot{\theta} = A + B \times \hat{\theta}_{13} \,. \tag{3-1}$$

Since this simulation is conditional on (a) number of unfinished items, (b) test, and (c) test form, separate (*A*, *B*) parameters must be obtained from each of the

$$(15 \times 6 \times 2) + (10 \times 3 \times 2) = 240$$

simulations. To apply this penalty, these three pieces of information are used to identify the appropriate (*A*, *B*) parameters that are applied to the provisional BME estimate to compute the final penalized value.

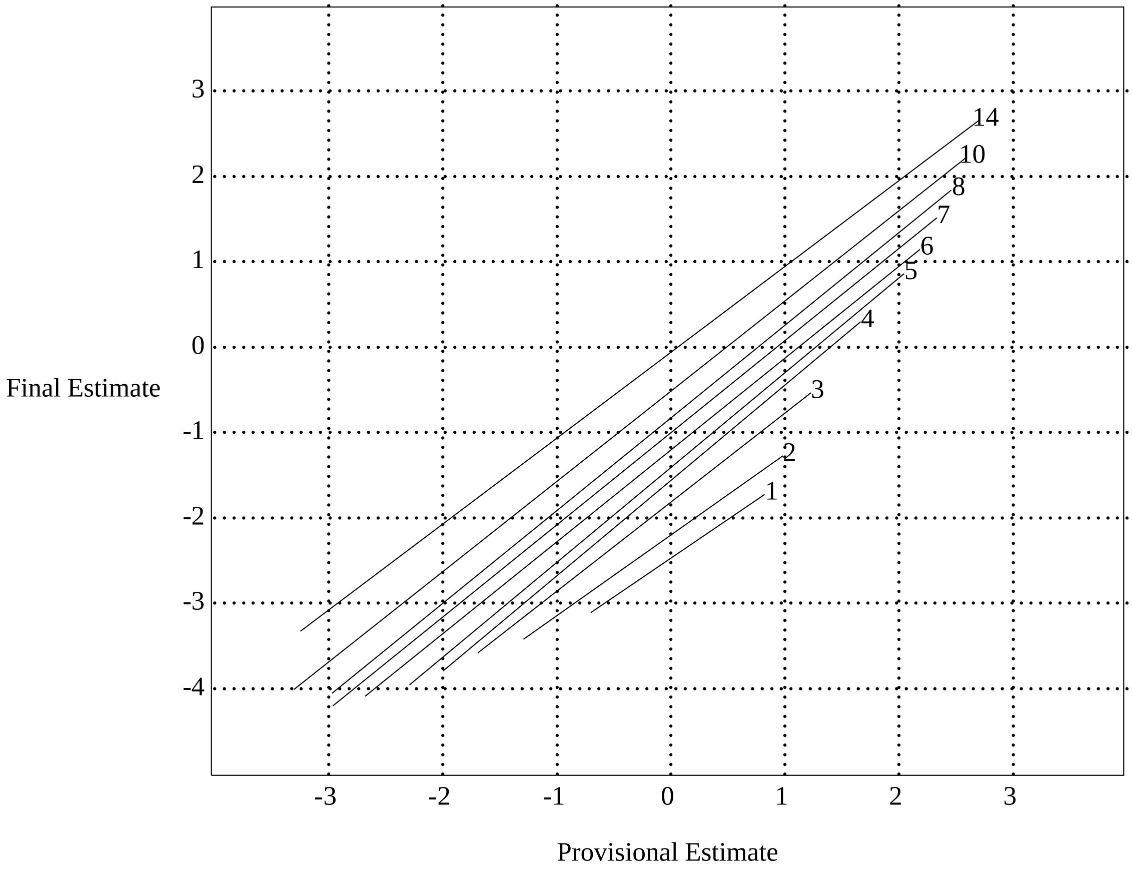


Figure 3-3. Selected Penalty Functions (by number of completed items) for EI Form 2.

Figure 3-3 displays selected functions for different numbers of completed items for EI Form 2. Note how these functions satisfy all the requirements stated earlier:

- The size of the penalty is positively related to the number of unfinished items.

- Applicants who have answered the same number of items and have the same provisional ability estimate will receive the same penalty.

- The procedure eliminates coachable test-taking strategies. There is no advantage for low-ability examinees to leave items unanswered, and applicants should be indifferent about guessing at random on remaining items or not answering them at all.

One undesirable consequence of the penalty procedure is a degradation in the precision of the final ability estimate. The penalty may not in general be correlated with the applicant's ability level. This degradation is expected to be small, however, mainly due to the infrequent application of this procedure. The time limits for each power test allow almost all examinees to finish. Table 3-2 provides the completion rates for those participating in the CAT-ASVAB Score Equating Verification (SEV) study (Chapter 9 in this technical bulletin). As indicated by the distribution of unfinished items (Table 3-2), the penalty procedure was applied to a small number of applicants, and among those receiving a penalty, almost all received a mild value.

Table 3-2. Frequency of Incomplete Adaptive Power Tests (*N* = 6,859)

Test	Number of Unfinished Items										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	≥ 10
General Science (GS)	6,762	52	18	13	3	4	2	1		2	2
Arithmetic Reasoning (AR)	6,788	47	14	5	1	3	1				
Word Knowledge (WK)	6,820	18	6	4	4	3	2		1		1
Paragraph Comprehension (PC)	6,807	36	10	6							
Auto Information (AI)	6,820	28	9	2							
Shop Information (SI)	6,779	52	20	5	2	1					
Mathematics Knowledge (MK)	6,797	29	10	9	8	3	1	1	1		
Mechanical Comprehension (MC)	6,843	12	1	1	2						
Electronics Information (EI)	6,833	16	7		1	1			1		

8. Number-Correct Transformation

For each power test, the BME (or penalized BME if incomplete) is converted to an equated number correct score. Procedures used to obtain the equating transformations for converting scores are described in Chapter 9 in this technical bulletin. Composite scores used for selection and classification are calculated from these number-right equivalents using the same formulas applied to the P&P-ASVAB reference form (8A).

Seeded Items

One advantage of computer-based testing is the ability to intersperse new, experimental test items among operational items to obtain calibration data. This is referred to as “seeding” items. Data collected on seeded items can be used to estimate IRT item parameters. This approach eliminates the need for special data collection efforts for the purpose of item tryout and calibration.

In CAT-ASVAB, each power test includes one seeded item. An examinee's response to this item is not used to estimate the examinee's provisional or final score. The seeded item is administered as the second, third, or fourth item in a test, with the position being randomly determined by the computer software at the time of testing. This approach, using only one seeded item per power test and administering it early in the test, was taken so that it would not be apparent to the examinee that the item is experimental. As a result, the examinee should answer the item with the same level of motivation as other items in the sequence. In full-scale implementation of CAT-ASVAB, one interspersed item per test will produce calibration data on enough new items to satisfy new form-development requirements.

Speeded Test Administration

Note: The two speeded tests, Numerical Operations (NO) and Coding Speed (CS), are no longer part of CAT-ASVAB. Effective January 2002 the Military Services stopped using NO and CS scores in their composites. Nonetheless, information is provided here for those who are interested in how the speeded tests were administered and scored.

The two speeded tests, Numerical Operations (NO) and Coding Speed (CS), are administered in a linear conventional format. For examinees receiving the same form, all receive the same items in the same sequence.

The speeded tests are scored using a rate score. In computerized measures of speeded abilities, rate scores have several advantages over number-right scores. First, rate scores do not produce distributions with ceiling effects that are often observed for speeded tests scored by number right. This is an especially important consideration when converting highly speeded tests from P&P medium to computer. The P&P time limit imposed on the computerized version will produce higher number-right scores, possibly leading to ceiling effect. This result can often be traced to speed of answer entry: entering an answer on a keyboard is faster than filling in a bubble on an answer sheet. Number-right scoring on a computerized measure of speeded abilities would require careful consideration of time-limit specification with special attention given to the shape of the score distribution. P&P-ASVAB time limits applied to the computerized versions of NO and CS produced unacceptably large ceiling effects. Additionally, rate scores have higher reliability estimates than number-correct scores (computed in an artificially imposed time interval).

For CAT-ASVAB running on the Hewlett Packard Integral Personal Computer (HP-IPC), the rate score was defined as

$$\hat{R} = \frac{P_g}{T_g} \times C, \tag{3-2}$$

where

$$T_g = \left(\prod_{i=1}^{n} T_i \right)^{\frac{1}{n}} \tag{3-3}$$

is the geometric mean of screen times T_i, and P_g is the proportion of correct responses corrected for guessing, which is

$$P_g = 1.25P - .25 \quad \textit{(for CS)} \tag{3-4}$$

$$P_g = 1.33P - .33 \quad \textit{(for NO)} \tag{3-5}$$

where *P* is the proportion of correct responses among attempted items. If the proportion in the numerator of Equation 3-2 were not corrected for guessing, an applicant could receive a very high score by pressing any key quickly without reading the items. Such an examinee would receive a low proportion correct, but a high rate score, because of the fast responding. Correcting the score for chance guessing eliminates the advantage associated with fast random responding. The constant *C* in Equation 3-2 is a scaling factor that allows the rate score $\hat{R}$ to be interpreted as the number of correct responses per minute. For NO, *C* = 60, and for CS, *C* = 420.

It is important to note one problem with the geometric rate score that arises when an examinee guesses at random on a portion of the items. If an examinee answers a portion of the test correctly and then responds at random to the remaining items very rapidly, the rate score (based on the geometric mean of response latencies) can be very large. An examinee could use this fact to game the test and artificially inflate his or her score. However, a rate score computed from the arithmetic mean of the response times does not suffer from this potential strategy. For this reason, in a later version of CAT-ASVAB (the version to be used in nationwide implementation) the geometric mean in Equation 3-3 was replaced by the arithmetic mean. The geometric mean was originally selected for

CAT-ASVAB because results of an early analysis (Wolfe, 1985) showed that in comparison with the arithmetic mean, the geometric mean possessed slightly higher estimates of reliability and slightly higher correlations with the pre-enlistment ASVAB speeded tests. However, in a similar analysis conducted on larger samples with more recent data (Chapter 7 in this technical bulletin), no significant difference in precision or validity was found between rate scores based on the arithmetic and geometric means.

For the speeded tests, response choices and latencies for screens interrupted by a “help” call are not included in the rate score. Time spent on a question interrupted by a help call may be atypical of the examinee’s response latency to other items. Although the examinee is returned to the same item after a help call, he or she has unrecorded time for thinking about the interrupted item. This may make the performance on the item systematically better than other items in the test.

Rate scores for each speeded test are converted to an equated number-correct score (see Chapter 9 in this technical bulletin). As with the adaptive power tests, composite scores used for selection and classification are calculated from these number-right equivalents using the same formulas applied to the P&P-ASVAB reference form (8A).

Administrative Requirements

Changing and Confirming an Answer

When the examinee selects an answer to a power test question, the selected alternative is highlighted on the screen. If the examinee wants to change an answer, he or she can press another answer key, and that response is highlighted in place of the first answer. When the examinee's choice is final, pressing the "Enter" key initiates scoring of the response using the answer that is currently highlighted, followed by presentation of the next item. Therefore, once the Enter key is pressed, the examinee cannot change the answer to that item. This procedure parallels, as closely as possible, the paper-and-pencil procedure of allowing the examinee to change the answer before moving on to the next question. Changing an answer once the Enter key is pressed and the next item is selected is not allowed because of the adaptive nature of the test.

However, on the speeded tests, the examinee's first answer initiates scoring the response; there is no opportunity to change an answer. Allowing examinees to change answers on speeded tests would be problematic for several reasons. If examinees were allowed to change responses to speeded tests, a choice between two (undesirable) options must be made on how to measure item latency, since item latencies are used in scoring these tests. One measure of latency might be from screen presentation to response entry, ignoring time to confirmation. This, however, could lead to a strategy where examinees press the answer key as quickly as possible, then take longer to confirm the accuracy of their answer. Another measure of latency might be from screen presentation to pressing of the Enter key or confirmation key. This approach, however, may add error to the measurement of ability, as speed in finding and pressing the Enter key could add an additional component to what the test measures.

Omitted Responses

In CAT-ASVAB, examinees are not permitted to omit items. The branching feature of adaptive testing requires a response from each examinee on each item as it is selected. Permitting examinees to omit items during the test is likely to lead to less than optimal item selection and scoring and may lead to various compromise strategies. While it would be possible to allow omitted responses on the speeded tests, since they are administered in a conventional manner, there is no psychometric or examinee advantage for doing so.

Screen Time Limits

In addition to test time limits, each item screen has a time limit. The purpose is to identify an examinee that is having a problem taking the test but is reluctant or unable to call for assistance. Two objectives were used to set the screen time limits. First, very few examinees should exceed the time limit. Second, the ratio screen and test time limits should not be unacceptably large. That is, it is important to ensure that if the examinee needs help, that not too much of the test time has expired before help is called. Screen time limits differ among the nine adaptive power tests and are displayed in Table 3-3. These screen time limits were first used in the CAT-ASVAB pretest (Vicino & Moreno, 1997), resulting in very few examinees exceeding the limit.

Table 3-3. Test Screen Time Limits (seconds)

GS	AR	WK	PC	NO	CS	AI	SI	MK	MC	EI
120	380	100	390	30	120	120	110	220	240	120

Help Calls

A machine-initiated “help” call is generated by the CAT-ASVAB system if an examinee times out on a screen or presses three invalid keys in a row. An examinee-initiated help call is generated when an examinee presses the Help key. Help calls stop all test timing and cause the system to bring up a series of help screens.

After a machine-initiated or examinee-initiated help call has been handled, all tests return to the screen containing the interrupted item, and the examinee is able to respond to the item. For speeded test scoring, the examinee’s response to the item on the interrupted screen is not counted toward the score. Interrupting a speeded test distracts the examinee and adds error to the latency measure. Since speeded tests use item latency in obtaining the test score, these latencies should be as accurate as possible. On adaptive power tests, the item is scored and is used for computing the examinee's provisional and final scores. Power tests do not use latencies in scoring the test, and test time limits are liberal. Therefore, any distraction caused by an interruption should have a minimal effect on the accuracy of the examinee’s score.

Display Format and Speed

The format of power test items displayed by the computer is as close as possible to the format used in the paper-and-pencil item calibration booklets. This was done to minimize any effects of format differences on item functioning. Speeded test items are presented in a format similar to P&P-ASVAB speeded test items so that the tests will be comparable across media. For NO, one item is presented per screen. For CS, seven items are presented per screen.

For the power tests, a line at the bottom right-hand corner of the screen displays the "number of items" and "time" remaining on the test. The time shown is rounded to the nearest minute, until the last minute when the display shows the remaining time in seconds. This procedure provides standardization of test administration, ensuring that all examinees have the means of pacing themselves during the test. This procedure, however, is not used for the speeded tests. Since these tests are scored with a rate score, pacing against the test time limit is not advantageous; the optimal strategy is to work as quickly and accurately as possible. Having a "clock" on the screen during the speeded tests would be disadvantageous to any examinee who looked at it, since time spent examining the clock would be better spent answering items.

For all tests, the delay between screens is no more than one second. In addition, the entire item is displayed at once and does not "scroll" onto the screen. These conventions were adopted since long delays in presenting items, variability in the rate of presentation of items, and occasional partial displays of items would probably contribute to additional unwanted variability of examinee performance—that is, error variance. Also, test-taking attitude might be adversely affected.

For a newer implementation of CAT-ASVAB presented on PC-based hardware (rather than HP-IPC), it was necessary to insert a delay between screens. The PC computers that were being used in nationwide implementation of CAT-ASVAB were much faster than the HP-based systems. With these fast machines, concerns about delays in item presentation disappeared, but a new concern appeared: items being presented too quickly. For this reason, the new system has a software-controlled constant delay of .5 second between screens.

Summary

CAT-ASVAB procedures described in this chapter have, nearly without exception, proven to be efficient and reliable, and therefore have been implemented in the operational version of CAT-ASVAB administered in locations throughout the United States. The empirical consequences of these psychometric procedures, and the relationship of the resulting CAT scores to the P&P-ASVAB, are documented in several other chapters in this technical bulletin. This information includes an evaluation of alternative forms reliability and construct validity (Chapter 7), an evaluation of predictive validity (Chapter 8), the equating of CAT-ASVAB to P&P-ASVAB (Chapter 9), and the consequence of calibration medium on CAT-ASVAB scores (Chapter 6). The favorable outcomes of these studies provide the best evidence to date of the soundness of these choices.

References

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

McBride, J. R., Wetzel, C. D., & Hetter, R. D. (1997). Preliminary psychometric research for CAT-ASVAB: Selecting an adaptive testing strategy. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 83-96). Washington, DC: American Psychological Association.

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

Segall, D. O. (1988). *A procedure for scoring incomplete adaptive tests in high stakes testing*. Unpublished manuscript. San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O., Moreno, K. E., Kieckhaefer, 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.

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

Wolfe, J. H. (1985). Speeded tests: Can computers improve measurement*? Proceedings of the Annual Conference of the Military Testing Association,* 27, 1, San Diego, CA: Military Testing Association, 49-54. (NTIS No. AD-A172 850)

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. 97-102). Washington, DC: American Psychological Association.

Chapter 4

ITEM EXPOSURE CONTROL IN CAT-ASVAB

Conventional paper-and-pencil (P&P) testing programs attempt to control the exposure of test questions by developing parallel forms. Test forms are usually administered at the same time to large groups of individuals and then discarded after a number of years. Computerized adaptive tests (CATs) require substantially larger item pools, and the cost of developing and discarding parallel forms becomes prohibitive. However, computer-based testing systems can control when and how often items are administered, and the development of procedures for controlling the exposure of test questions has become an important issue in adaptive testing research.

CATs achieve maximum precision when each item administered is the most informative for the current estimate of the examinee's ability level. For any ability estimate, only one item satisfies this requirement; therefore, when ability estimates are the same for different examinees, the item administered must also be the same. In the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), examinees begin the test under the assumption that they have equal abilities. Under a maximum-information selection rule, the most informative item would be the same for every examinee, the second item would be one of two choices (one after a correct answer, another after an incorrect one), and so on. As a consequence, the item sequence in this case is predictable and the initial items are used more frequently—thus becoming overexposed.

Early CAT-ASVAB research with the Apple III microcomputers used a procedure aimed at reducing sequence predictability and the exposure of initial items (McBride & Martin, 1983). In this procedure, called the 5-4-3-2-1, the first item is randomly selected from the best (most informative) five items in the pool, the second item is selected from the best four, the third item is selected from the best three, and the fourth item from the best two. The fifth and subsequent items are administered as selected. The ability estimate is updated after each item. While this strategy reduces the predictability of item sequences, its net effect is substantial use and overexposure of a pool's most informative items.

To reduce the amount of item exposure and satisfy the security requirements of the operational CAT-ASVAB, a probabilistic algorithm was developed by Sympson and Hetter (1985). The algorithm was specifically designed to (a) reduce predictability of adaptive-item sequences and overexposure of the most informative items, and (b) control overall item use in such a way that the probability of an item being administered (and, thereby "exposed") to any examinee can be approximated to a pre-specified maximum value. The algorithm controls item selection during adaptive testing through previously computed parameters (K_i) associated with each item.

Computation of the K_i Parameters

To calculate the K_i, simulated adaptive tests are administered to a large group of simulated examinees ("simulees") whose "true" abilities are randomly sampled from an ability distribution representative of the real examinee population. Test administrations are repeated until certain values (to be defined below) converge to a pre-specified expected exposure rate.

For the CAT-ASVAB, 1,900 “true” abilities were drawn from a normal distribution of ability, $N(0,1)$. To simulate examinee responses, a pseudo-random number was drawn from a uniform distribution in the interval (0,1). If the random number was less than the three-parameter logistic model (3PL) probability of a correct response, the item was scored correct; otherwise it was scored incorrect. The CAT-ASVAB item parameters and the “true” abilities were used to compute the 3PL probabilities. The actual steps in the computations are described below.

Steps in the Sympson-Hetter Procedure

Steps one to three are performed once for each test. Steps four through eight are iterated until a criterion is met.

1. Specify the maximum expected item-exposure rate r for the test. In the CAT-ASVAB battery, the rates were set to match those of the P&P-ASVAB, which comprises six forms. Four of the tests in the ASVAB battery are used to compute the Armed Forces Qualification Test (AFQT) composite score, which is used to determine enlistment eligibility. The AFQT tests in the six P&P forms are different; but each non-AFQT test is used in two forms. This results in exposure rates r = 1/6 for AFQT tests, and r = 1/3 for non-AFQT tests. The CAT-ASVAB has two forms, and to approximate the same values for them, expected exposure rates were set to r = 1/3 for AFQT tests (1/6 over two forms) and r = 2/3 for non-AFQT tests (1/3 over two forms).

2. Construct an information table (infotable) using the available item pool. An infotable consists of lists of items by ability level. Within each list, all the items in the pool are arranged in descending order of the values of their information functions (Birnbaum, 1968, Section 17.7) computed at that ability level. In the CAT-ASVAB, infotables comprise 37 levels equally spaced along the (-2.25, +2.25) ability interval.

3. Generate the first set of K_i values. If there are *i* items in the item pool, generate an *i*-long vector containing the value 1.0 in each element. Denote the ith element of this vector as the K_i associated with item *I*.

4. Administer adaptive tests to a random sample of simulees. For each item, identify the most informative item i available at the infotable ability level (θ) nearest the examinee's current ability estimate ($\hat{\theta}$); then generate a pseudo-random-number x from the uniform distribution (0,1). Administer item i if x is less than, or equal to, the corresponding Ki. Whether or not item i is administered, exclude it from further administration for the remainder of this examinee's test. Note that for the first simulation, all the Ki's are equal to 1.0, and every item is administered if selected.

5. Keep track of the number of times each item in the pool is selected (NS) and the number of times that it is administered (NA) in the total simulee sample . When the complete sample has been tested, compute *P(S),* the probability that an item is selected, and *P(A),* the probability that an item is administered given that it has been selected, for each item:

$$P(S) = \text{NS}/\text{NE} \tag{4-1}$$

$$P(A) = \text{NA}/\text{NE} \tag{4-2}$$

where NE = total number of examinees.

6. Using the value of *r* set in Step 1, and the *P(S)* values computed above, compute new K_i as follows:

$$\text{If } P(S) > r, \text{ then new } K_i = r/P(S) \tag{4-3}$$

$$\text{If } P(S) \leq r, \text{ then new } K_i = 1.0 \tag{4-4}$$

7. For adaptive tests of length *n*, ensure that there are at least *n* items in the item pool that have new K_i = 1.0. Items with K_i = 1.0 are always administered when selected, since the random number is always less than or equal to 1. If there are fewer than *n* items with new K_i = 1.0, set the *n* largest K_i equal to 1.0. This guarantees that all examinees will get a complete test of length *n* before exhausting the item pool.

8. Given the new K_i, go back to Step 4. Using the same examinees, repeat Steps 4, 5, 6, and 7 until the maximum value of *P(A)* that is obtained in Step 5 (maximum across all the items in the test) approaches a limit slightly above *r* and then oscillates in successive simulations.

The K_i obtained from the final round of computer simulations are the exposure-control parameters to be used in real testing.

Use of the K_i During Testing

The process works as follows: (a) select the most informative item for the current ability estimate; (b) generate a pseudo-random number *x* from a uniform (0,1) distribution; and (c) if *x* is less than, or equal to, the item’s K_i, administer the item; if *x* is greater than the K_i, do not administer the item but identify the next most-informative item and repeat (a), (b), and (c). Selected but not-administered items are set aside and excluded from further use for the current examinee; items are always selected from a set of items that have been neither administered nor set-aside. Note that for every examinee, the set of available items at the beginning of a test is the complete item pool.

Simulation Results

For the CAT-ASVAB tests, the maximum *P(A)* values obtained in Step 5 approached the *r* values after five or six iterations. Table 4-1 shows *P(A)* results for two AFQT tests: Paragraph Comprehension and Arithmetic Reasoning. For both tests, the expected exposure rate *r* had been set equal to 1/3.

Table 4-1. Maximum Usage Proportion *P* (A) by Test and Simulation Number

Simulation Number	Paragraph Comprehension Test	Arithmetic Reasoning Test
1	1.000	1.000
2	0.540	0.562
3	0.412	0.397
4	0.361	0.367
5	0.364	0.357
6	0.352	0.354
7	0.359	0.345
8	0.349	0.358
9	0.357	0.352
10	0.357	0.365

Precision

When the exposure-control algorithm is used, optimum precision is not achieved since the best item (most informative) is not always administered. To evaluate the precision of the CAT-ASVAB tests, score information functions were approximated from simulated adaptive test sessions conducted with and without exposure control. The sessions were repeated independently for 500 examinees at each of 31 different theta levels equally spaced along the (–3, +3) interval. These theta levels are assumed to be true abilities for the simulations. Infotables and simulated responses were as in the K_i simulations above. Score information was approximated using Equation 2-4 in Chapter 2 of this technical report.

Figures 4-1 and 4-2 present score information curves for Arithmetic Reasoning and Paragraph Comprehension, respectively. The loss of precision due to the use of exposure control is very small and uniform across the theta range in Arithmetic Reasoning and more noticeable in the average ability region for Paragraph Comprehension. There are no losses, or some gains, at the extremes of the ability distribution. Results for the remaining tests were similar.

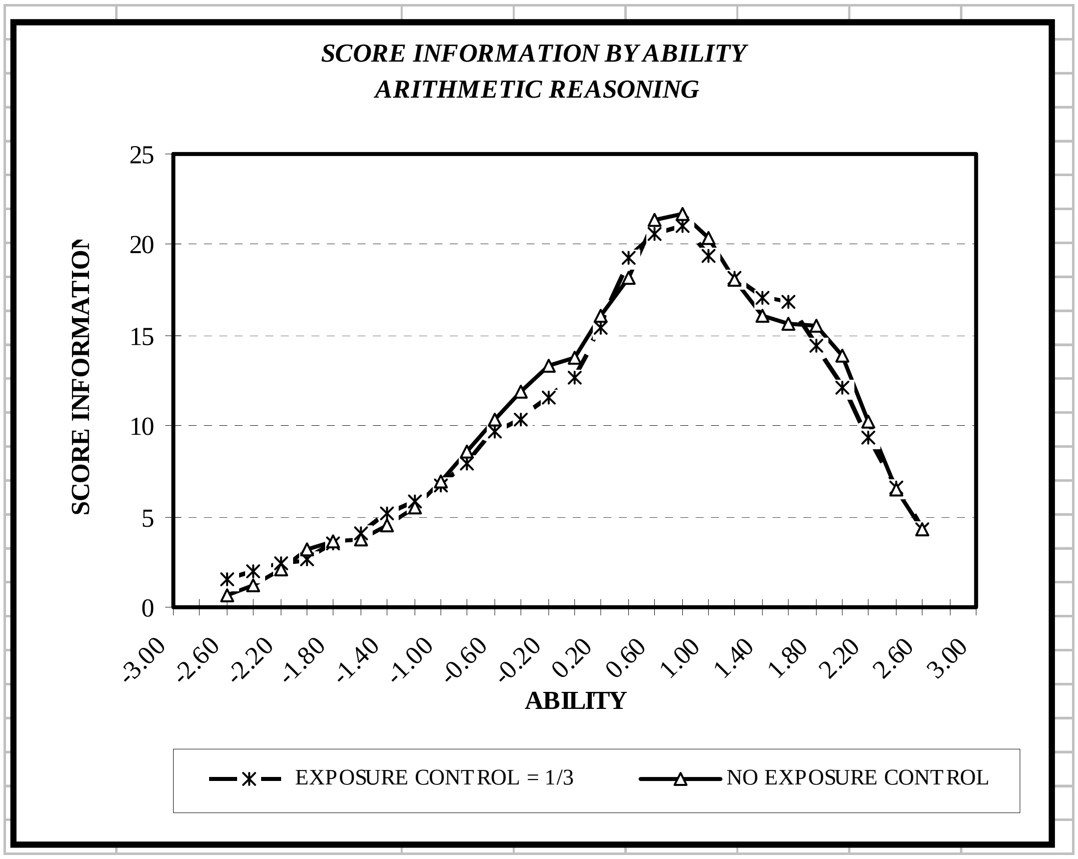


Figure 4-1. Score information by ability: Arithmetic Reasoning Test

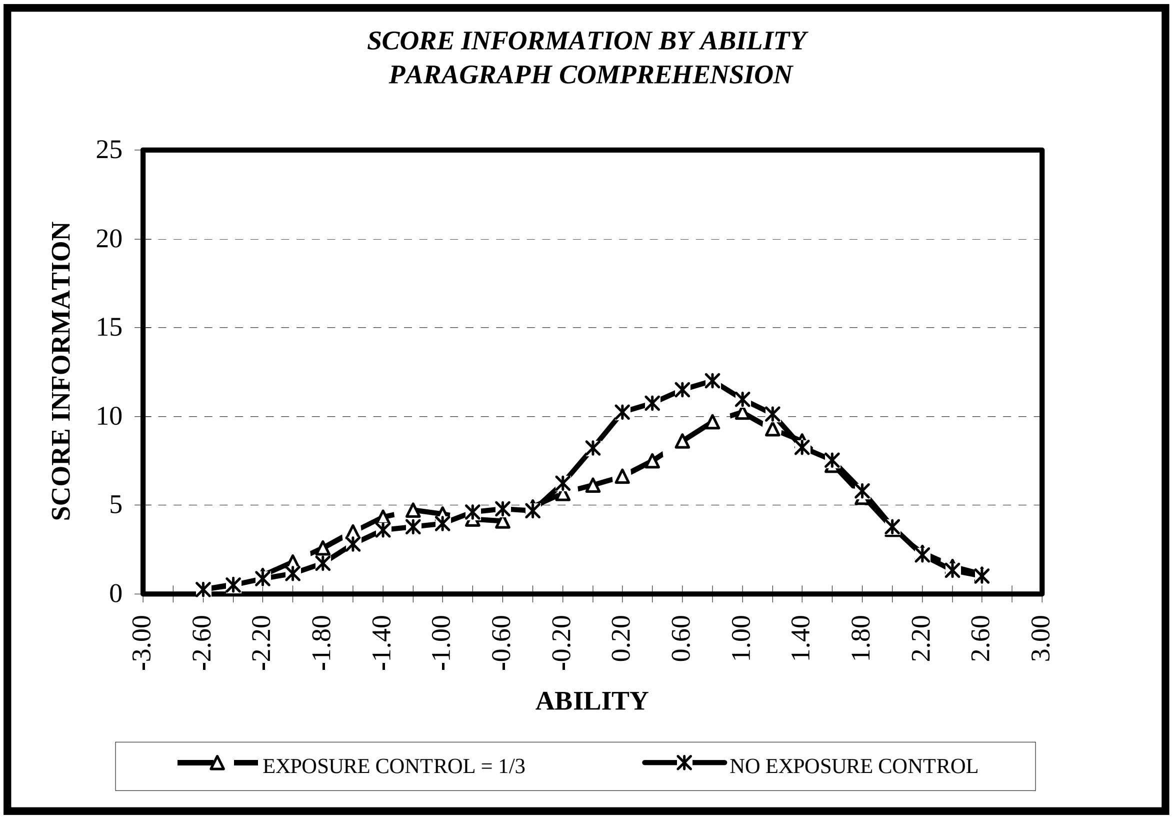


Figure 4-2. Score information by ability: Paragraph Comprehension Test

Summary

These results indicate that the use of exposure-control parameters does not significantly affect the precision of the CAT-ASVAB tests but will reduce the exposure of their best items. Future work should evaluate actual item use from the CAT-ASVAB operational administration data.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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*, 223-235. New York, NY: Academic Press.

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.

Chapter 5

ACAP HARDWARE SELECTION, SOFTWARE DEVELOPMENT, AND ACCEPTANCE TESTING

This chapter discusses the development and acceptance testing of a computer network system to support the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) program from 1984 to 1994. During that time, the program was devoted to realizing the goals of the Accelerated CAT-ASVAB Project (ACAP).

Since 1979, under the CAT-ASVAB program that has been described in the earlier chapters, the Joint Services had been developing a computer system to support the implementation of the CAT strategy at testing sites of the United States Military Entrance Processing Command (USMEPCOM). In 1984, a full-scale development (FSD) contracting effort was initiated with the expectation of using extensive contractor support to design and manufacture a unique computer system that could be used at USMEPCOM. In 1985, the FSD effort was terminated and the ACAP was initiated, primarily because the contracting effort was consuming too many resources to commence, let alone complete, the desired system. In addition, the recent advent of powerful microcomputer systems on the commercial market encouraged program managers to pursue the use of off-the-shelf microcomputers in contrast to developing a system unique to the project.

The implementation concerns for the ACAP system focused primarily on the psychometric requirements of the CAT-ASVAB system—specifically, the evaluation of CAT-generated aptitude scores and the equating of these scores to the paper-and-pencil ASVAB (P&P-ASVAB) aptitude scores. To meet this requirement, the Joint Services decided that all the computer support components should be in place so that the psychometric research could be conducted without confounding by factors other than those affecting operational use of such a system. Therefore, the ACAP was required to develop a computer system capable of supporting all of the functional specifications of CAT-ASVAB in a time frame consistent with continued support of the program.

In brief, ACAP was tasked to develop a CAT-ASVAB computer system to refine the operational requirements for the eventual system and to conduct the psychometric research efforts for equating CAT scores with those of the P&P-ASVAB. To this end, ACAP tried to identify and address these requirements as much as possible in an operational environment. This was accomplished by using commercially available computer hardware in a field test of CAT-ASVAB functions at selected USMEPCOM sites. At those sites, CAT-ASVAB testing must be implemented in accordance with the specifications for the original contracting effort, and in accordance with specifications from new psychometric requirements that arose during the course of ACAP development. The design and development of the computer system to support CAT-ASVAB progressed along two obviously interrelated dimensions: computer hardware and software.

ACAP Hardware Selection

The hardware needed for the CAT-ASVAB system had to be selected before the operating system and programming language could be identified. Specifically, a Local CAT-ASVAB Network (LCN) of interconnected computers was to administer CAT-ASVAB to applicants for enlisted Military Service at any of approximately 64 Military Entrance Processing Stations (MEPSs) or approximately 900 Mobile Examining Team Sites (METSs) within USMEPCOM. In addition, a Data Handling Computer (DHC) at each MEPS handles communication of information between the LCN units and a CAT central research facility. The DHC also stores examinee testing and equipment utilization data for six months, as required.

Original Hardware Specifications and Design

The hardware configuration envisioned by the Joint Services in the original contracting effort involved transportable computer systems at the MEPSs and METSs, based on the concept of a "generic" LCN. A generic LCN consists of six examinee testing (ET) stations monitored (via an electronic network) by a single test administrator (TA) station and peripheral support equipment (e.g., mass storage devices and printers). Under a networked configuration, a single TA station must allow the TA to monitor up to 24 ET stations (i.e., administer the CAT-ASVAB to 24 examinees simultaneously). The CAT-ASVAB portability requirements specify that each generic LCN consist of up to eight components weighing a total of no more than 120 pounds, each component weighing no more than 23 pounds. Environmental requirements for operating temperature, humidity, and altitude are also specified. The TA and

ET stations must be interchangeable so that each TA and ET station can serve as the backup for any other station in the LCN.

The LCN computer hardware specifications have remained relatively unchanged as follows: Each ET station consists of a response device, a screen display, and access to sufficient random access memory (RAM) and/or data storage for administration of any CAT-ASVAB test; the amount of RAM required depends on the specific application software and networking design used. The ET stations are tied to a TA station by networking cables. Each TA station is essentially an ET station with a mass storage device and full-size keyboard. The failure of one station must not affect the performance of any other unit in the LCN. Each TA station has a very portable printer and modem. All components operate on ordinary 110 VAC line current. Battery packs are not used because they add weight and require additional logistic support.

In the METSs, the LCN operational requirements would be as follows: Each LCN administers the CAT-ASVAB to military applicants scheduled for testing at the METS. Initially, an Office of Personnel Management (OPM) examiner would pick up the LCN equipment at a staging area (U.S. MEPCOM, 1983), transport it to the test site, carry it from the vehicle to the test site (sometimes a hotel room), and configure it for testing. When the system is ready for testing (i.e., "booting" and loading of source code/data files are completed), the TA solicits personal data (name, Social Security number [SSN], etc.) from each examinee and enters this information into the system at the examiner's TA station. Then the TA instructs each examinee to sit at a specified ET station and start testing without further TA assistance. Examinee item response information is stored on a nonvolatile medium (e.g., micro floppy disk) to allow the test to continue at another ET station in the event the original ET station fails during a testing session. Finally, the TA is expected to monitor the various testing activities at the ET stations (e.g., CAT-ASVAB testing progress status and use of a "Help" function). After all examinees at a METS have completed testing, the TA sends the entire Examinee Data File (consisting of the personal data, item level responses, test scores, and composite scores) to the DHC unit at the associated MEPS, using a modem and dial-up telephone line if available. If this is not possible (e.g., no telephone line at the test site), the examiner transfers the data after the equipment is returned to the staging area. Finally, the TA packs up and returns all equipment to the staging area.

MEPS equipment is stationary but otherwise identical to METS equipment. In contrast to most METSs, each TA at a MEPS testing site must be capable of monitoring 24 ET stations simultaneously. In addition, on start-up, the TA obtains the latest software and testing data from the DHC unit at the MEPS via either a hard-wired connection or a transportable medium. At the end of testing, testing data are sent to the DHC using the same medium. An LCN at the MEPS would not use dial-up telephone lines.

The MEPS site implementation of CAT-ASVAB also includes a DHC unit to collect data daily from each LCN in the associated MEPS administrative segment, including any LCNs at METSs. These data are to be compiled and organized on the DHC for

- Daily transmission of an extract of examinee data collected that day to the USMEPCOM minicomputer located at the MEPSs.

- Periodic transmission of all examinee data to the Defense Manpower Data Center (DMDC).

- Archiving of all examinee and equipment utilization data at the MEPSs for at least six months.

The MEPS DHC also must be capable of receiving (a) new software, (b) test item bank updates, and (c) instructions from DMDC, and telecommunicating this information to field LCN units.

ACAP Hardware Development

The three generic computer system designs being considered for use as the local computer network for the CAT-ASVAB program were discussed by Tiggle and Rafacz (1985). The three designs differed in how they stored and provided access to test items during test administration. Storing test items on removable media (e.g., 3.5-inch micro floppy disks) or a central file server (e.g., a hard disk) had disadvantages with security, media updating, ease of use, maintenance, reliability, and response time.

The design selected emphasizes the use of RAM. Each TA and ET station requires at least 1.5 megabytes (MB) of internal RAM which can accommodate all the software and data needed to administer the CAT-ASVAB tests. In case of LCN failure, each ET station can operate independently of any other station in the network. The ET station needs one micro floppy disk drive and an electroluminescent, or LCD technology, display screen. In addition, the TA station can perform the functions of an "electronic" file server. The TA station should have a large amount of total

RAM available to provide great flexibility in the total number of alternate forms available during any one test session.

This design offers many advantages, including a large degree of flexibility with respect to design options. The ET stations can operate as stand-alone devices (i.e., without the use of the TA station). This being the case, it would be virtually impossible for an examinee's test session to fail to be completed; each ET station would be a backup station for every other station in the LCN. This design is very reliable because it minimizes use of mechanical devices. Finally, the design provides a very high level of security because volatile RAM is erased when the power to the computer is turned off.

LCN monitoring and the system response-time requirements are not functionally related. The computer hardware can be configured so that the data storage requirements (for any one CAT-ASVAB form) reside at the ET station. Therefore, the response-time display of test items can be independent of the LCN. The item-display process takes place at RAM speed, resulting in a maximum response time on the order of one second, which is well within CAT-ASVAB specifications.

The hardware procurement for ACAP was negotiated by the Navy Supply Center, San Diego, CA, using a brand name or equivalent procurement strategy. This resulted in the selection of the Hewlett Packard Integral Personal Computer (HP-IPC) to meet the specifications. Each ET station consists of the following components in a single compact and transportable (25-pound) package:

- One 8 MHz 68000 CPU with 1.5 MB of internal RAM with an internal data transfer rate (RAM to RAM) of 175 KB/second.

- One read-only memory (ROM) chip with 256 KB of available memory containing a kernel of the UNIX operating system.

- One microfloppy disk drive (710 KB capacity) with data transfer rate (disk to RAM) of 9.42 KB/second.

- One adjustable electroluminescent display with a resolution of 512 (horizontal) by 255 (vertical) pixels (screen size 9 inches measured diagonally; 8 inches wide by 4 inches high).

- One custom-built examinee input device (essentially a modification of the standard HP-IPC keyboard).

- One Hewlett Packard Interface Loop (HP-IL) networking card.

- One integrated ink-jet printer for use when the ET station must serve as a backup to the TA station.

Each TA station is configured identically to the ET station but includes 2.5 − 4.5 MB of internal RAM and a full-size ASCII keyboard.

In summary, each generic LCN (i.e., six ET stations tied to a single TA station) consists of seven transportable components weighing a total of approximately 175 pounds. Using the HP-IL networking card and special network driver software achieves a network data transfer rate of approximately 9KB per second.

The data handing computer (DHC) system, also based on the HP-IPC, consists of the following components:

- One ET station with a full-size keyboard.
- Two 55-MB hard disk drives (primary and backup data archive units).
- One cartridge tape drive unit; periodically, a cartridge tape of examinee testing data is to be sent to NPRDC.
- Telecommunications hardware to communicate with the MEPS minicomputer.

ACAP Software Development

ACAP documentation specified "C" as the programming language for software development because it was native to the UNIX operating system on the selected hardware and had the following characteristics that greatly aided software development, performance, and testing: (a) support of structured programming, (b) portability, (c) execution speed, (d) concise definitions and fast access to data structures, and (e) real-time system programming. The following paragraphs briefly describe the ACAP software development effort.

Technically, the approach to the software development efforts proceeded along traditional lines; that is, a top-down structured design approach was used, consistent with current military standards for software development (e.g., DOD-STD-2167A). The functional requirements for each of the three software packages—TA station, ET station, and DHC —were identified and developed to assist in developing a macro-level design for each package; that is, how is the software going to work from the standpoint of the user/operator?

These requirements also served as the basis for developing detailed computer programming logic to support the main functions within the macro-level design. A thorough study of this logic permitted the identification of the primitive routines and procedures that were necessary (e.g., a routine was required to confirm the correct insertion of a disk into the disk drive and to solicit and confirm the entry of ET station identification numbers). Then, using the primitive routines, main stream (logic) drivers were developed to link the primitives into a working system that mirrors the functional requirements of the macro-level design. The software was then tested, errors were identified and corrected, and re-testing continued until all portions of the software worked together as required. Occasionally, the software design had to be modified as the impact of the interaction among various routines became more complicated and/or specifications were more clearly defined.

TA Station Software

To design the software for the TA station, the functions to be supported by the TA station were compiled. The following outline describes generic TA station functions:

1. The TA must prepare and communicate all software and data necessary for CAT-ASVAB test administration to ET stations in the LCN.

2. The TA must be able to identify examinees by means of a unique identifier (e.g., SSN) and to record (in a retrievable file) other examinee personal data. In addition, it should be easy for the TA to add or modify any of the personal data.

3. The software for the TA station must randomly assign (transparent to the TA) an examinee taking CAT-ASVAB to one of the two CAT-ASVAB forms used. This assignment is subject to the condition that examinees who have previously been administered a CAT-ASVAB form must be re-tested on the alternate CAT-ASVAB form. In addition, the software must maintain an accounting of examinee assignments and be prepared to develop new assignments if any station in the LCN fails.

4. During examinee testing (in the networking mode of operation), the TA station must be able to receive a status report on the progress of examinees upon demand.

5. The TA station must be able to move the completed testing data recorded from an ET for additional processing and, at that time, produce appropriate hard copy of testing results.

6. The TA station must be able to store the testing data for all examinees who have gone through the TA station collection process in a nonvolatile medium (i.e., a Data Disk) for later communication to the parent MEPS.

7. Finally, it must be almost impossible for an examinee's testing session not to be completed. If an examinee's assigned ET station fails, that examinee must be reassigned to another available station and continue testing at the beginning of the first uncompleted CAT-ASVAB test. Likewise, if the TA station fails, the LCN fails, or electrical power is interrupted, the TA must be able to recover and continue the testing session promptly.

In actual use, simply installing a system disk (called a TA disk) and turning on the power to the TA station begins boot-up operations to prepare the LCN for subsequent processing. At this point the TA would normally select the networking mode of operation for the current testing session. The standalone mode is a failure recovery procedure in the event the TA station or the network supporting the LCN failed. After performing several network diagnostic tests, the TA transmits testing data to the ET stations in the LCN. Then the program provides instructions for loading the data from three system disks which contain test administration software, item level data files (encoded), and supporting data (seeded test items, information tables, and item exposure control values). After these data and software are loaded into RAM of the TA station, the system disks are secured.

The TA station randomly identifies a CAT-ASVAB test form with each ET station so that approximately 50 percent of the ET stations receive each of the two CAT-ASVAB forms. The TA station then proceeds to broadcast the test administration software and data files (one at a time, alternately) to the ET stations requiring a given form, then to the remaining stations. Therefore, while one set of stations (identified with one of the two forms) is receiving one file of test items, the remaining stations are storing the test items received into RAM.

At this point the TA identifies the current testing session in terms of the date and approximate starting time for the session, and the Main Menu is displayed. The Main Menu displays the primary functions performed by the TA during a testing session, as explained below.

- PROCESS is a means for the TA to identify examinees to be tested in terms of their name, SSN, and test-type information. The PROCESS function also includes creating a new list of examinees for testing, editing current examinee information, adding (or deleting) an examinee for testing, and providing a screen and/or printed list of examinees for testing.

- The ASSIGN option randomly directs (unassigned) examinees to unassigned ET stations in the network; equivalently, it randomly assigns each examinee to one of the two CAT-ASVAB test item-bank forms. The examinee assignments are recorded on the TA disk at the TA station, printed at the TA station, and then broadcast to the ET stations in the LCN. Unassigned stations may serve as failure recovery stations. At this point, the TA would direct the examinees to sit at the seats corresponding to their assigned ET station, whereupon they receive computer-controlled general instructions that start CAT-ASVAB test administration.

- During the testing session, the TA can use the STATUS option for a screen report on the progress of examinees during testing.

This report includes the examinee's name, SSN, total time accumulated since the CAT-ASVAB began, the test being administered, the accumulated time on that test, and the expected completion time for the entire battery of CAT-ASVAB tests. The examinee's recruiter uses the expected completion time to assist in scheduling.

- The SUBMIT option in the Main Menu enables the TA to enter into a menu-driven dialogue with the TA station that records various personal information from the examinee's USMEPCOM Form 714-A. This information includes Service and component for which the examinee is being processed, gender, education level and degree code, and race/population group.

- At the end of examinee test administration, the TA uses the COLLECT option to retrieve (one at a time, or automatically upon test completion) the examinee's testing data from the assigned ET station. The TA station printer then produces a score report that includes equated number-right scores (interchangeable with the P&P-ASVAB scores) and an AFQT percentile score.

- By selecting the RECORD option, the TA can record (COLLECTed) examinee testing data on a set of microfloppy disks (identified as MASTER and BACKUP Data Disks) for subsequent transfer. The MASTER Data Disk is sent to the parent MEPS for processing, while the BACKUP Data Disk remains secured at the testing site and is sent to the MEPS if needed.

As briefly mentioned above, the software in the ACAP system includes the capability of supporting various failure recovery operations. The interested reader is referred to Rafacz (1995) for additional information.

ET Station Software

The design of the software for the ET station was based on the psychometric requirements for CAT, supplemented by specifications associated with the computer administration of any test, improved psychometric procedures, and requirements unique for military testing. During testing, the ET stations are only required to communicate with the TA station at the end of administration of each item (and before the next item is displayed) to provide status information to the TA station.

In addition to the purely psychometric functions supporting the use of the CAT technology, the software design considers the functions supporting computer operations at the ET station. During examinee test administration, two operations are of concern: (a) failure recovery at the ET station, and (b) examinee implicit and explicit requests for help.

The ET station software design, with respect to all functions supported, is discussed below.

1. Placing an ET disk in the disk drive of the ET station initiates the following boot-up operations: (a) performing hardware verification procedures (screen, disk drive, and keyboard), (b) soliciting the mode of operation for the computer (networking or standalone), (c) requesting the ET station computer identifi-

cation number, and (d) verifying that the ET station computer clock has been set to the correct date and time.

Normally the TA selects the networking mode of operation. If the standalone mode is selected, broadcasting of software and data files is not required. In that case, the ET station reads the necessary testing data and software directly from the ACAP system disks. In addition, ET station assignments, dictated by the TA station and test type (initial or retest), are entered manually by the TA at each ET station. Finally, examinee testing information recorded on the ET disk is collected manually by moving the ET disk to the TA station at the conclusion of examinee testing.

2. Now the ET station is ready to receive test item data files and software from the TA station. The first file is the actual test administration software which, once received, terminates the boot-up program and then monitors receipt of the following data files (from the TA station) to support examinee test administration: (a) power and speeded test item text, graphic, and item parameter files; (b) information table files; and (c) exposure-control parameters for power test items. Each power test item file is stored in the ET station RAM, which is designed to support subsequent random retrieval (according to the information table associated with each power test).

3. After an ET station has received all of the required data files, it is ready to receive the examinee assignment list from the TA station. Once this list is received, the ET station prepares to administer the test to the assigned examinee. This requires confirming that the correct form of test items has been loaded

for the assigned examinee. If not, the ET station requests the ACAP system disks and the correct testing data files are loaded into RAM; this incorrect form loading rarely happens.

4. Now that the ET station is ready to administer the CAT-ASVAB to the assigned examinee, the TA must give the examinee verbal instructions and direct him or her to the assigned ET station. The TA verifies the displayed SSN with the examinee and modifies it if necessary. The examinee presses the Enter key on the keyboard of the ET station when requested to begin CAT-ASVAB administration, in accordance with the interactive dialogues specified by Rafacz and Moreno (1987). The dialogue for the remainder of examinee test administration is between the ET station (software) and the examinee; neither the TA nor the TA station is involved.

5. Initially, the computer screen presents the examinee with information on how to use the ET station keyboard. The examinee learns how to use all of the keys labeled ENTER, A, B, C, D, E, and HELP.

6. Next, the examinee is trained on how to answer the power test items. (Training on how to respond to the speeded test items is given just before these tests are administered.) The examinee can ask to repeat the training on how to use the keyboard and answer test items. If a second request occurs, the ET station halts the interactive dialogue with the examinee so that the TA can be called to enter a pass code for the interactive dialogue to continue. The ET station software describes the current situation and then requests that the TA monitor the examinee's progress briefly before continuing with normal duties.

7. At this point, four power tests—General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), and Paragraph Comprehension (PC)—are administered. For each test, the examinee is initially presented with a practice item, is given an indication that his or her answer is correct or incorrect, and is then given the opportunity to ask to repeat the practice item. The second request initiates a call to the TA, who must enter a pass code to repeat the practice item. Finally, the examinee is ready to be administered the actual test items.

   As the power test items are displayed, the examinee answers the test item by pressing the key corresponding to the alternative selected and then confirms the answer by pressing the Enter key. Any other answer can be selected before Enter is pressed. Selection of a valid response alternative highlights only that alternative on the screen until another alternative is selected. Pressing an invalid key results in an error message being briefly displayed. As each item is displayed on the computer screen, the lower right corner of the screen presents the number of the item being administered, relative to the total number of items, and the number of minutes remaining in the test.

   While the examinee studies the test item, his or her performance is recorded by the software monitoring the keyboard. Overall, if the examinee does not confirm a valid response within the maximum item time limit, the test is halted and a TA implicit Help call is initiated. In addition, if the examinee fails to complete the specified number of test items in the allotted maximum time limit for the *entire test*, the test is automatically terminated (without a TA call) and the examinee continues with the next CAT-ASVAB test. If the examinee presses an invalid key, an error message is briefly displayed. Three invalid key-presses result in an implicit Help call. Pressing the Help key initiates the explicit Help call sequence. For the power tests, a valid key response (A, B, C, D, or E) must be followed by the confirmation key (Enter) to generate the display of the next item.

8. The test continues until the number of items administered (including one seeded item in each power test) equals the required test length or the maximum test time limit has been reached. As soon as the examinee completes the test, certain examinee test administration information is recorded in the ET station RAM and on the ET disk. For each item administered, this information includes the item identification code, the examinee-selected response alternative, the time required to select (but not confirm) the response, the new estimate of ability based on the selected response, and any implicit or explicit Help calls. In addition, the Bayesian modal estimate for the test is recorded, as is information on the examinee's performance on the practice screens for the test. This information is also recorded on the ET disk (a non-volatile medium) as a backup if the ET station fails during testing.

9. The Numerical Operations (NO) and Coding Speed (CS) speeded tests are administered after the first four power tests. As with the power tests, practice test items are administered first. The examinee can repeat the practice items up to three times before a TA call is initiated. Examinee test administration of the speeded tests differs from the power tests. The

speeded test items are administered in the sequence in which they appear in the item file, without using any adaptive testing strategy. In addition, the examinee does not confirm an answer by pressing the Enter key; rather, the ET station selects the first valid key-press (A, B, C, D, or E) as the examinee's answer. The display format of the CS test items is also different in that seven items are displayed on the same computer screen, whereas NO and the power tests display only one item per screen. Rate scores are recorded as the examinee's final speeded test score (see Chapter 2 in this technical bulletin). In all other respects, speeded test administration (including the availability of implicit and explicit Help calls and the recording of examinee performance information) is identical to that of the power tests.

10. Once the speeded tests are completed, the examinee is administered the remaining five power tests: Auto Information (AI), Shop Information (SI), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). The procedure for administering these tests is identical to that for the first four power tests. Once the EI test is completed, the examinee's testing performance is stored in the ET station RAM and on the ET disk into a single file identified by examinee SSN. The TA station collected this SSN file for subsequent compilation onto a Data Disk. The ET station instructs the examinee to return to the TA station for further instructions, and the examinee is then excused. The ET station is now available for testing some other examinee, perhaps one whose assigned station might have failed during the testing session.

During examinee test administration, normal administration activities can be interrupted to accommodate situations involving an examinee's need for assistance. These situations are either implicit Help requests where the software of the ET station infers that the examinee needs assistance or explicit Help requests where the examinee presses the red Help key on the keyboard. Rafacz (1995) discusses in some detail the implementation of Help calls in the ET station software.

Data Handling Computer (DHC) Software

Software development was less critical for the DHC than for the ET and TA stations because the DHC serves primarily as a manager of examinee testing data *after* test administration. The DHC has two primary functions:

- **Data compilation.** The DHC compiles and organizes examinee testing data recorded on the data disks from the testing sites. Data recorded on a data disk must be removed and stored on a non-volatile medium for subsequent communication to users of the CAT-ASVAB system. Appropriate backup mechanisms must be in place before data are purged from a data disk; once purged of its data, the data disk is returned to a testing site for reuse.

- **Data distribution.** The DHC must be able to communicate the examinee testing data to users of the system. Specifically, an extract of each examinee's testing record must be communicated to the USMEPCOM (System 80) minicomputer at the parent MEPS. In addition, all of the examinee testing data must be sent to DMDC for software quality assurance processing and communicating the data to other users of the CAT-ASVAB system.

DHC software must also ensure that the DHC collects each examinee's testing data only once and distributes each compiled data set to each user only once. An override mechanism must be available to send the information again if the original information is lost in transit. Finally, it must be possible for the DHC to recover from a hardware failure. Details concerning the functions and software development issues for the DHC may be found in Folchi (1986) and Rafacz (1995).

Item Pool Automation

In addition to the development of the TA, ET, and DHC software, a requirement of ACAP was to automate the item pools for each of the two forms of the CAT-ASVAB. The automation phase involved preparing the individual components (text, graphics, and item parameters) of candidate test items for storage and administration on the ACAP microcomputer system.

Power Test Items

The ACAP power test items consisted of two components for items with text only and three components for items with graphics. The first two components, the item text files and the item parameter files, existed on magnetic media. The third component, the graphics, existed only as black-and-white line drawings in the experimental booklets used in calibrating the source item bank, the Omnibus Item Pool (Prestwood, Vale, Massey, & Welsh, 1985).

The graphics were captured from the experimental booklets and processed before text and parameters were merged. The ACAP Image Capturing System (Bodzin, 1986) was used. It consisted of an IBM PC-Compatible computer, the Datacopy 700 Optical Scanner, the *Word Image Processing System* (WIPS) (Datacopy Corporation, 1985a), and the HP-IPC. The process also required the program, *boxit16*, which calculates the optimal size for the display of each image on the HP-IPC screen. During the process of scaling an image to the optimal size for the HP-IPC screen, information was lost, reducing the quality of the image. The image was restored to the original quality of the drawing in the test booklet using the *WIPS Graphic Editor* (Datacopy Corporation, 1985b).

After the graphic images were captured and edited, they were transferred to the HP-IPC. Additional processing was necessary before the images could be used with the ACAP test administration program. Special-purpose programs were written to display the images, verify the integrity of the file transfer, define the optimal image size for the HP-IPC screen, and rewrite the file header. Any image editing necessary was performed using *yage*, the graphics editor written for the HP-IPC.

The item text and parameter files were transferred to the HP-IPC and reformatted before being merged with the graphics portion of the items. Reformatting included reducing the size of the files and inserting specific characters recognized by the test administration software. Finally, the item text file, item parameter file, and images were merged in the Item Image Editor using a program called *edit*, written specially for this purpose. To conserve storage space, the graphic components were compressed as the items were stored.

Speeded Item Files

The speeded items were prepared by the Armstrong Laboratory and delivered on IBM-formatted 5.25-inch diskettes. Speeded items, which consist of item text only, had to be modified to be compatible with the ACAP test administration software. These modifications were made using the Unix editor, *vi*.

System Documentation

Documentation requirements that apply to ACAP primarily deal with the design, development, use, and maintenance of the software supporting the ACAP network. For each of the three software systems (TA and ET stations, and DHC), user/operator manuals, programmer's reference manuals, and system test plans were developed for each of the three phases of the ACAP.

To support the use of the ACAP network at selected MEPSs in an operational mode (and provide examinee scores of record), the user of the system, USMEPCOM, had declared its requirements for system documentation, apart from the original Stage 2 RFP. These requirements use DoD-STD-7935A Automated Data Systems [ADS] Documentation as the specification source document. In summary, the following documentation has been completed for each phase of the ACAP in accordance with the standard.

- An ACAP system, including Functional Description, System/Subsystem Specification, Data Requirements, and Data Element Dictionary (four documents)

- A Programmer’s Maintenance Manual and a System Test Plan for the TA station, ET station, and DHC software systems (six documents)

- A User's Manual for all of the ACAP software systems (one document)

- An Operations Manual for the TA station and an Operations Manual for the DHC (two documents)

System Testing Procedures

The approach used to test the software was important to the design and development of the ACAP system. Several things could be done during design and development to avoid (or at least minimize) the generation of software errors. Choice of the programming language was an important decision. The selection of “C” as the programming language for ACAP was based upon its support of structured programming—including concise definitions, fast access to data structures, and a repertoire of debugging aids. These are the characteristics of a language that minimize the chances of errors being created in the software under development.

In addition, appropriate programming standards and practices must be used as the software is designed and developed. For example, the software was designed as modular units with minimal interaction among the units. The modules were executed by a main "driver" program that controls the sequence of executions and verifies the results produced. Above all, the use of "long logic jumps” should be avoided. Appropriate software development standards were used for the specific application area; in the ACAP, as much as possible, DoD-STD-2167A was used.

Once the ACAP software was developed, it was necessary to test the software, locate errors, make necessary corrections, and retest the software until no errors were found. However, there were so many logic flow paths that it was physically impossible to test even a small proportion of such paths in a reasonable period of time. To address this concern, the Stage 2 RFP required the development of built-in test (BIT) software for use within the CAT-ASVAB system.

The BIT procedures that were used for the ET station (the most logically complex package) included adding software with the capability of reading examinee responses directly from a separate (scenario) file in contrast to the keyboard. This "scenario" file also included predetermined response latencies for test items, as well as various testing times for the tests. By using the scenario files, many different logic flow paths and testing configurations were evaluated, yet no (real) examinee was involved in actual test administration.

Once a scenario was completed, the system tester surveyed the output data to confirm that the information recorded matched that specified in the scenario. For the most part, any differences were attributed to software errors, which were then quickly located and corrected. By using such BIT techniques, it was possible within ACAP to minimize the time required to test a logic path within the software. Because more logic paths were tested, uncertainty as to errors that still might be "hidden" in the software was reduced.

Documents describe in detail the testing procedures for evaluating software performance and the checklists to be completed by system testers to record the testing activities.

Software Acceptance Testing

In addition to system testing, software acceptance testing was conducted. While system testing is generally conducted by software designers and developers, software acceptance testing is conducted by an independent group knowledgeable in how the system should function.

Software acceptance testing was a critical element from the beginning of the development of the CAT-ASVAB system. The concerns for quality were twofold and equally relevant. One involved how a user would interact with the system—where a user might be a test administrator or an applicant taking the test—and the other involved the accuracy of the test scores. In a computerized-adaptive test, ensuring accuracy is a complex and difficult process. It means checking for things such as clear and consistent item-screen displays, precise timing (of instruction sequences, Help calls, response times, time limits), the integrity of parameter files throughout test administration, the selection of the proper item in the adaptive sequence, and the calculation and recording of the final scores.

There were three different kinds of checks: configuration management, psychometric performance, and software performance. Some of the checks are simple but tedious, some are manual and extremely detailed, and some are complex and computerized.

Many of the CAT-ASVAB software acceptance procedures were instituted from the start; others were developed as we learned from experience in using the system and from feedback from trainers, examinees, and test administrators. All the checks were performed every time the software changed and required substantial amounts of time from numerous members of the project staff. The rigorous execution of these checks contributed significantly to the consistently high quality of performance of the CAT-ASVAB system.

Configuration Management

The ACAP uses three distinct hardware and software systems: the ET station, the TA station, and the DHC. As the first step in configuration management, each system's components were identified: computers, memory boards, interface boards, and hard disk size and type. Commercial software and versions used in each system were documented, and copies of the programs were archived. The commercial software included the operating system, compilers, various libraries, and numerous utilities.

For each system, every component or module of any software specifically developed for CAT-ASVAB was identified and listed. Included were source code and executables for all programs, subroutines, and procedures; parameter files; and compilation files (such as Unix "make" files). Source code and executables for programs specifically developed for CAT to support software development were also included.

The next step was recompilation of all the software. A computer with a hard disk (called the ATG system, for Acceptance Testing Group) was set aside to be used solely for recompilation and was restarted with all the commercial system and utilities software used by the CAT-ASVAB. The following steps were completed for every recompilation:

1. The software development team delivered diskettes containing source and executable programs to the ATG. Next, all the source and executable CAT-ASVAB files from the prior version were erased from the hard disk.

2. The new source files were loaded from the diskettes and compiled. Executables were created and compared (bit by bit) to those delivered by the development team. If there were no differences, the programs became the "acceptance testing" version of the software. If differences were found, the documented results were provided to the software developers and the diskettes returned.

After corrections were made by the software development team, Steps 1 and 2 were repeated. This process ensured that the correct version of the software was used in subsequent checks. Software specifically developed for the CAT-ASVAB was tested after every change that required recompilation, regardless of the magnitude of the change. The complete system was tested whenever any file in any of the three components changed.

Software Performance

Once the executable programs were accepted after recompilation, members of the ATG took simulated tests, following prescribed scenarios. The tests covered a wide variety of conditions, some designed to check system specifications and others to replicate situations that occur in the field during operational testing. They included manual tests of menu screens, such as TA options during an examinee's Help call, and an examinee's option to repeat a practice problem. They also included checks of item and test elapsed times, performance of failure/recovery procedures, screen sequences, and others. In these checks, a test is taken and all responses are given following a prescribed scenario. For test and item times, a stop-watch is used and the values recorded. The stop-watch value is then checked against the value in the output file. Failures are simulated by turning off the computer, performing the prescribed recovery, examining the output file, and processing it through the quality- control programs.

Speeded Tests. Since these tests are not adaptive, and the item sequence is known, the displayed item screens are checked manually against printed copy. The response times are checked with a stop-watch.

Power Tests. A computer program developed in-house reads the following values from the results of a CAT test (let this be Test 1): The seed used by the pseudo-random number generator, the unique identification number (UID) of all the items administered, and the examinee's responses to the items. Using the UIDs, the program reads the text of the corresponding items from the original archived text files and prints the items (with the corresponding responses) in the form of a "booklet." The items appear in the same sequence as they were administered in the original CAT test.

The booklet is then used to take a second test (Test 2) on an HP-IPC. Test 2 is administered with the operational software, except for the random-generator seed which is forced to be the same value as in Test 1. Using the same seed generates the same random number, which will lead to selection of the same first item. The reviewer compares the item on the screen to the one printed in the "booklet" and then gives the answer printed in the booklet. When this is done for every item, all subsequent items are the same as in the original Test 1.

Psychometric Performance

Examples of psychometric performance are checks to ensure that the computer file that contains the examinee's answers matches what happened during test administration, that the proper questions are selected during adaptive testing, that the time limits are correctly enforced by the software and hardware (for both power and speeded tests and individual items), that the correct keys are used to score the items, and that the items displayed on the screen are the same as those recorded on the output file. Some of the checks were automated; others had to be performed manually. The main procedures are described below.

Quality Control Program 1. This program checks (a) structure and format by screen type, (b) the ranges for all the variables, (c) test time-outs against allotted times, and (d) the sum of elapsed item times for all the tests. It also computes the raw and standard test scores for all of the power and speeded tests, the AFQT, and the Service composite scores, and compares them against the recorded values.

Quality Control Program 2. This program checks adaptive item selection and scoring in power tests and scoring in the speeded tests. The software reads the output of a CAT-ASVAB test and simulates a second test (a replication) using the examinee's responses and the seed for the pseudo-random number generator from the first one. To ensure independence of results, the program runs on a computer system different from the operational HP-IPC; information tables, item parameters, keys, and exposure-control parameters are read from the original archived files, not from the operational diskettes. The program simulates an adaptive test and compares the results, at every step, with the original results. Discrepancies are identified and printed, including those in items selected and their order, and in all the ability estimates: the intermediate Owen's Bayesian and the final Bayesian mode. Optionally, random numbers, exposure-control parameters, and information table indices for every item are also printed. All CAT-ASVAB test protocols—operational, research, and simulated—are processed through these two programs.

ACAP System Summary

To summarize the ACAP system development and acceptance testing efforts: the ACAP computer network can be used as the delivery vehicle for CAT-ASVAB as specified by the Joint Services in the Stage 2 RFP. For all critical functions, the ACAP system provides a capability meeting, if not exceeding, functional requirements specified in the Stage 2 RFP.

The Stage 2 RFP documented CAT-ASVAB system performance requirements over nine evaluation factors:

1. performance,
2. suitability,
3. reliability,
4. maintainability,

5. ease of use,
6. security,
7. affordability,
8. expandability/flexibility, and
9. psychometric acceptability.

Rafacz (1995) describes in some detail the extent to which the ACAP computer network system met the requirements of each factor to support the Score Equating Development (SED) and Score Equating Verification (SEV) phases of the ACAP. The Operational Test and Evaluation (OT&E) functions of expanded examinee score reporting, and the installation of the Enhanced Computer Administered Tests (ECAT) tests, demonstrate the capability of the ACAP system to meet the psychometric criteria for acceptability. Installing the variable-start mechanism, as well as other OT&E enhancements that involve the operator interface, further improve the image of the system in terms of suitability and ease of use.

Finally, it should be observed that the computer software developed to support CAT-ASVAB functions on the HP-IPC has proven to be based on a very flexible and powerful design. Using a large RAM-based design for the ET station has made overall software design and structure less complicated. The net effect was to make it easier for system developers to isolate critical coding segments and minimize the ripple effects due to software errors associated with related functions. For example, the software routines needed to support recovery of the ET station in a failure situation are not dependent on the software of any other station in the testing room. Furthermore, the multi-tasking feature of the UNIX operating system was useful during software development because the system permitted the execution of multiple tasks: text editing, compiling, and executing tasks could proceed concurrently on the same development system. In addition, the ease with which TAs used the system in the field during OT&E implementation (Chapter 10 in this technical bulletin) clearly indicates a system that can effectively serve as the delivery vehicle for CAT-ASVAB.

References

Bodzin, L.J. (1986). *An image capturing and editing system for the HP-Integral computer*. Unpublished paper. San Diego, CA: Naval Ocean Systems Command.

Datacopy Corporation. (1985a). *Datacopy Model 700 User's Guide (Version 1.4).*

Datacopy Corporation. (1985b). *Datacopy WIPS Editor User's Guide (Version 1.0).*

Folchi, J.S. (1986). Communication of computerized adaptive testing results in support of ACAP. *Proceedings of the Annual Conference of the Military Testing Association*, *28,* 618-623. New London, CT: U.S. Coast Guard Academy. (NTIS No. AD-A226 551)

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). Brooks AFB, TX: Air Force Human Resources Laboratory. (AD-A160 608)

Rafacz, B. A. (1995). *Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB): Computer system development* (NPRDC-TN-95-8). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. AD-299 806)

Rafacz, B.A., & Moreno, K.E. (1987). *Interactive screen dialogues for the examinee testing (ET) station.* San Diego, CA: Navy Personnel Research and Development Center.

Tiggle, R.B., & Rafacz, B.A. (1985). Evaluation of three local CAT-ASVAB network designs. *Proceedings of the Annual Conference of the Military Testing Association*, 23-28. San Diego, CA*:* Navy Personnel Research and Development Center. (NTIS No. AD-A172 850)

USMEPCOM. (1983). *U.S. MEPCOM Mobile Examining Team Site requirements for computerized adaptive testing*. Director of Testing, USMEPCOM Letter Report dated December 22, 1983. Vol. I, II, and III.

Chapter 6

EVALUATING ITEM CALIBRATION MEDIUM IN COMPUTERIZED-ADAPTIVE TESTING

Computerized-adaptive testing (CAT) provides efficient assessment of psychological constructs (see Weiss, 1983). When combined with item response theory (IRT), CAT uses item parameter estimates to select the most informative item for administration at each step in assessing an examinee's abilities. In addition, these item parameters are used to update both point and interval estimates of each examinee's score.

A practical concern in the initial development of CAT is whether items must be calibrated from data collected in a computerized administration or whether equally accurate results could be obtained by calibrating the items from data collected in a paper-and-pencil (P&P) administration. For example, in the development of the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), item parameter estimates were available only from a P&P administration of the items (Prestwood, Vale, Massey, & Welsh, 1985) because computers were not available at the testing sites. This made it important to assess whether scores obtained on the CAT-ASVAB using the P&P-based item calibration had the same precision and interpretation as scores obtained from a computer-based calibration of the items.

Previous Research

Generally, research comparing the effects of computer-based and P&P-based administration of cognitive tests has dealt primarily with the medium of administration (MOA) of the actual test rather than the MOA used for calibrating items. Although the investigators did not always explicitly address CAT, the work provided results that were suggestive of the potential importance of three MOA effects.

Two studies by Moreno and her colleagues examined the effect of MOA on the construct assessed by the tests. Observed-score factor analytic and correlational studies (Moreno, Wetzel, McBride, & Weiss, 1984; Moreno, Segall, & Kieckhaefer, 1985) suggested that the factor pattern of a cognitive battery has the same hyperplane pattern whether the tests are administered by conventional P&P or adaptively by computer. A meta-analytic study by Mead and Drasgow (1993) obtained correlations close to 1.00 between computerized and P&P versions of the same power tests when the correlations were corrected for attenuation, whether the computerized tests were adaptive or non-adaptive. The findings of Mead and Drasgow imply that the disattenuated correlations among tests of different traits are essentially the same whether the traits are measured using the same MOA or a different MOA. However, this implication had yet to be tested empirically.

Researchers also have examined MOA effect on test precision. Green, Bock, Linn, Lord, and Reckase (1984) suggested that nonsystematic MOA effects could degrade CAT precision if the tests were administered and scored using P&P-based item calibrations. They noted that such effects could arise when some items were affected (e.g., in difficulty) by MOA and other items were not. Divgi (1986) and Divgi and Stoloff (1986) found that item response functions (IRFs) estimated from items administered adaptively by computer differed from IRFs obtained from a conventional P&P administration of the same items. However, these differences were not systematically related to the content of the items and, when applied to the scoring of adaptively administered items, produced only slight effects on final test scores. Moreno and Segall (Chapter 7 of this technical bulletin) showed that even if nonsystematic effects of calibration error resulted from using a P&P-based calibration in an adaptive test, the adaptive test still could have greater reliability than a longer, conventional P&P test.

Although these results were reassuring about the relative precision of CAT and P&P tests, what remained to be demonstrated was whether the medium used to obtain item parameters affects CAT precision. Specifically, the issue was whether or not non-adaptive, computer-administered items produce a calibration that results in CAT scores with greater reliability than scores produced from a P&P-based calibration.

Previous work investigated MOA effect on the score scale of the tests. Green, Bock, Linn, Lord, and Reckase (1984) suggested that MOA could also have a systematic effect on the score scale—for example, by making the items more difficult or easier to a similar extent. Empirical results reported by Spray, Ackerman, Reckase, and Carlson (1989) and Mead and Drasgow (1993) indicated that computer-administered items can result in slightly lower mean test scores than P&P-administered items. Spray et al. investigated whether effects were general to all items or specific to certain items. They found no MOA effect for most of their items, which made their results inconclusive. An important issue that remained to be investigated was whether MOA effects on the score scale of a test are systematic—that is, removable by a transformation (e.g., linear) of the score scale—or nonsystematic—that is, altering the reliability of scores of some items but not others.

Study Purpose

This study compared effects on CAT-ASVAB scores using a P&P calibration versus a computer-based calibration. The two primary effects investigated were (a) the construct being assessed, and (b) the reliability of the test scores. The specific question was the extent to which adaptive scores obtained with computer-administered items and a P&P calibration corresponded to adaptive scores obtained with the same computer-administered items (and responses) and a computer calibration. A secondary inquiry concerned the influence of calibration medium on the score scale: the extent to which IRT difficulty parameters obtained with a P&P calibration corresponded to those obtained with a calibration of the same items from a non-adaptive computer administration.

Method

At each testing session, examinees were randomly assigned to one of three groups. Fixed blocks of power test items were administered by computer to one group of examinees (Group 1) and by P&P to a second group (Group 2). Those data were used to obtain computer-based and P&P-based three-parameter logistic (3PL) model calibrations of the items. Then each calibration was used to estimate IRT adaptive scores (θs) for a third group of examinees who were administered the items by computer (Group 3). The effects of the calibration MOA (CMOA) on the construct being assessed and on the reliability of the test scores were assessed by comparative analyses of the θs using the alternative calibrations. CMOA effects on the score scale were assessed by comparing IRT difficulty parameters from computer-based and P&P-based calibrations.

Examinees

Examinees were 2,955 Navy recruits stationed at the Recruit Training Center in San Diego, CA: 989 in Group 1, 978 in Group 2, and 988 in Group 3. A simulation study by Hulin, Drasgow, and Parsons (1983, pp. 101-110) indicated that larger samples produce little improvement in the precision of IRFs and test scores, given the 40 items used in these calibrations. ASVAB scores were obtained from file data for nearly all examinees and were used to assess whether the groups were comparable in ability level.

Calibration Tests

Items were taken from item pools developed for the CAT-ASVAB by Prestwood, Vale, Massey, and Welsh, 1985. Forty items from each of four content areas—General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), and Shop Information (SI)—were used (160 items total). Although only 4 of the 11 CAT-ASVAB tests were included in this study, MOA tests were administered in the same order as in the CAT-ASVAB. The three groups received exactly the same instructions, the same practice problems, the same items, in the same order, and with the same time limits. The items were conventionally administered in order of ascending difficulty, using the 3PL model difficulties obtained by Prestwood et al.

The P&P test employed a booklet and an optically scannable answer sheet; the booklet format was the same as that used in the original P&P calibration by Prestwood et al. (1985). The computer-administered format was the same as that used in CAT-ASVAB (one item per screen, no return to previous items, no omits allowed). Practice problems and instructions were printed on the booklet and read aloud by the proctor for the P&P group (Group 2) and presented on the screen, with the option-to-repeat, for the computer groups (Groups 1 and 3). Tests were timed; however, time limits were liberal. Test order and time limits were GS, 19 minutes; AR, 36 minutes; WK, 16 minutes; and SI, 17 minutes.

Item Calibrations

IRT parameter estimates based on the 3PL model (Birnbaum, 1968) were obtained in separate calibrations for computer Group 1 (calibration C1) and for P&P Group 2 (calibration C2). The response data sets on which the calibrations were based were labeled U1 and U2, respectively. The calibrations were performed with LOGIST 6 (Wingersky, Barton, & Lord, 1982), a computer program that uses a joint maximum-likelihood approach. Response data set U3 from Group 3 (the second computer group) was not used in the calibrations. The design with the corresponding notations is shown in Table 6-1.

Table 6-1. Calibration Design

Group	Medium	Data Set/ Item Responses	Item Parameters/ Calibrations
1	Computer	U1	C1
2	P&P	U2	C2
3	Computer	U3	—

Scores

For each examinee in Group 3, two $\hat{\theta}$s were computed for each test (see Table 6-2). All $\hat{\theta}$s were based on the U3 responses. $\hat{\theta}$s for variables X_{GSC}, X_{ARC}, X_{WKC}, and X_{SIC} (C is computer CMOA) were calculated using the computer-based item parameters (C1). Scores for variables X_{GSP}, X_{ARP}, X_{WKP}, and X_{SIP} (P is P&P CMOA) were calculated using the P&P-based item parameters (C2). All $\hat{\theta}$s were based on simulated CATs, computed as described below, using only 10 of the 40 responses from a given examinee.

Adaptive Scores. To compare the adaptive $\hat{\theta}$s, 10-item adaptive tests were simulated using actual examinee responses. As in CAT-ASVAB, a normal (0,1) prior distribution of $\hat{\theta}$ was assumed. Owen's (1975) Bayesian scoring was used to update $\hat{\theta}$, and a Bayesian modal estimate was computed at the end of the test to obtain the final $\hat{\theta}$. Items were adaptively selected from information tables on the basis of maximum information. An information table consists of lists of items by θ level; within each list, all items in the pool (40) were arranged in descending order of the values of their information functions computed at that θ level. The information tables used in this study were computed for 37 θ levels equally spaced along the (–2.25 to 2.25) interval.

ASVAB Scores. The Armed Forces Qualification Test (AFQT) scores were obtained from the enlistment records of most examinees. These scores, which all the Military Services use to determine eligibility for enlistment, were used to assess the equivalency of the three groups.

Covariance Structure Analysis

The equality of $\hat{\theta}$s calculated from P&P and computer-estimated item parameters was investigated using covariance structure analysis based on the eight variables defined in Table 6-2.

Table 6-2. Variable Definitions

Variable	Content Area	Responses	Item Parameter Calibration Medium
X_{GSC}	GS	U3/Group 3	Computer
X_{ARC}	AR	U3/Group 3	Computer
X_{WKC}	WK	U3/Group 3	Computer
X_{SIC}	SI	U3/Group 3	Computer
X_{GSP}	GS	U3/Group 3	P&P
X_{ARP}	AR	U3/Group 3	P&P
X_{WKP}	WK	U3/Group 3	P&P
X_{SIP}	SI	U3/Group 3	P&P

The formal model was defined as follows. Let a random observation *i* from Group 3 be denoted as Y_{ti}, where *t* denotes one of four adaptive tests (GS, AR, WK, or SI). In the adaptive test, item selection and scoring were assumed to be based on item parameters representative of a population of item parameters, where the population consists of parameters obtained from each of a large number of CMOAs. A large number of hypothetical media of administration can be defined from various combinations of item display format (defined, in turn, by the choice of font, color, and display medium) and response format (defined, in turn, by the choice of format of the answer sheet or automated input device). The random observation is assumed to be on a standardized score scale with a mean of 0 and a variance of 1. The 1 x 4 vector of observations, $Y_i = \{Y_{ti}\}$, is assumed to be a multivariate normal random variable with a 4 × 4 correlation matrix, Φ. A standardized random observation based on the use of item parameters from a specific CMOA is denoted W_{tmi} and is assumed to have a linear regression on Y_{ti},

$$W_{tmi} = P_{tm}Y_{ti} + e_{tmi} \,. \tag{6-1}$$

The ρ_{tmi} are errors assumed to have a multivariate normal distribution and to be independent of each other and of the Y_{ti}. They are interpreted as errors in test scores due to nonsystematic departure of item parameters from the population-representative item parameters used to obtain Y_{ti}. These errors are a combination of various CMOA effects not definable by a linear transformation of the score scale, such as sampling variation of the parameter estimates and variation due to the interaction of specific item contents and the CMOA. Note that, because the W_{tmi} and Y_{ti} are both standardized variables, the regression coefficient, ρ_{tm}, is the correlation between these variables, and the error variance is $1 - \rho^2_{tm}$. Also, note that the equivalence of ρ_{tm} across CMOA for each test can be taken as an indicator of similar amounts of nonsystematic calibration error across CMOA.

From these definitions of W_{tmi} and Y_{ti}, it follows that the observed score on test *t* in medium *m* can be written as

$$X_{tmi} = \sigma_{tm} W_{tmi} + \mu_{tm} \,, \tag{6-2}$$

where σ_{tmi} and μ_{tmi} are the observed scale standard deviation and location (mean) parameters, respectively. If the CMOA has no linear effect on the score scale for test *t***,** then σ_{tmi} and μ_{tmi} are the same for all m (i.e., for all CMOA).

The covariance matrix Σ among the eight variables can be modeled in terms of several parameter matrices:

$$\Sigma = \Lambda\left(R^{1/2} J\Phi J' R^{1/2} - R + I\right)\Lambda \,, \tag{6-3}$$

where **Λ** and **R** are 8 x 8 diagonal matrices with elements

$$\Lambda = diag\{\sigma_{GSC,}\sigma_{ARC,}\sigma_{WKC,}\sigma_{SIC,}\sigma_{GSP,}\sigma_{ARP,}\sigma_{WKP,}\sigma_{SIP}\}$$

and

$$R = diag\{\rho_{GSC,}\rho_{ARC,}\rho_{WKC,}\rho_{SIC,}\rho_{gsp}, \rho_{ARP,}\rho_{WKP,}\rho_{SIP}\} \,.$$

The Λ matrix contains the standard deviations of the observed variables, and the R matrix contains the reliability parameters.

These reliability parameters measure only one source of error variance: the random error variance in test scores arising from sampling errors in item parameters. These reliability parameters do not measure error in the traditional sense, which measures the error in test scores associated with the sampling of items from an infinite pool of items.

The matrix J is 8 x 4 with

$$J = \begin{bmatrix} I_4 \\ I_4 \end{bmatrix}$$

(6-4)

where $\mathbf{I}_4$ is a 4 × 4 identity matrix. Additionally, let $\mathbf{I}_8$ denote an 8 x 8 identity matrix.

In Equation 6-3, Φ is a 4 × 4 symmetric matrix with diagonal elements equal to 1. The Φ matrix contains the disattenuated correlations among the four tests. Note that in this context, the correlations are corrected for calibration error only. These correlations are not corrected for attenuation due to measurement errors.

From Equation 6-3, the disattenuated correlation matrix among the eight variables is given by

$$J\Phi J' = \begin{bmatrix} \Phi_{cc} & \Phi_{pc} \\ \Phi_{cp} & \Phi_{pp} \end{bmatrix}$$

(6-5)

where the three non-redundant submatrices are constrained by the model to be equivalent: $\Phi_{cc} = \Phi_{pc} = \Phi_{pp}$ (= Φ). From classical test theory, the product $R^{1/2}\ J\Phi J'R^{1/2}$ represents the correlation matrix among observed variables, with the eight reliability parameters along the diagonal. Consequently, the sum $R^{1/2}$ $J\Phi J'$ $R^{1/2} - R + I_8$ represents the correlation matrix among observed variables, with 1s in the diagonal. Finally, by pre- and post-multiplying the observed correlation matrix by Λ (the 8 x 8 diagonal matrix of standard deviations), the observed covariance matrix Σ is obtained.

In addition to estimating the model given by Equation 6-3, an additional model was examined to test the equivalency of the reliability parameters across the CMOA. The constraints imposed by the two models are summarized in Table 6-3. Model 1 imposed constraint A, which equated the disattenuated correlations across the CMOA. Model 2 imposed both constraints A and B, where B constrained the reliability parameters. Consequently, in Model 2, the reliability values for each test were constrained to be equivalent across the two calibration media. Model parameters were estimated by normal-theory maximum-likelihood using the SAS procedure CALIS (SAS Institute, 1990).

Models 1 and 2 represent a hierarchy of nested models. Consequently, the χ^2 difference test can be used to examine the statistical significance of each set of constraints. Significance tests were performed on each set of constraints listed in Table 6-3. For both models, the likelihood ratio χ^2 statistic of overall fit was calculated. To test the equivalency of disattenuated correlations across the CMOA ($\Phi_{cc} = \Phi_{pc} = \Phi_{pp}$), the likelihood χ^2 value for Model 1 was used. To test the equivalency of the reliability parameters, the difference between the χ^2 values of Models 1 and 2 was evaluated. Under the null hypothesis, this difference was distributed as χ^2 with 4 degrees of freedom (*df*).

Table 6-3. Model Constraints

Constraint	Parameters			
A	$\Phi_{cc} = \Phi_{pc} = \Phi_{pp}$			
B	$\rho_{GSC} = \rho_{GSP}$,	$\rho_{ARC} = \rho_{ARP}$,	$\rho_{WKC} = \rho_{WKP}$,	$\rho_{SIC} = \rho_{SIP}$

Results

Group Equivalence

Two examinees in Group 3 had fewer than ten valid responses for WK and SI and were eliminated from all subsequent analyses of these two tests. Thus, the Group 3 sample sizes were 988 for GS and AR and 986 for WK and SI. An analysis of variance indicated a nonsignificant difference among the three group means on AFQT. This result provided some assurance that the three groups were equivalent with respect to ASVAB aptitude.

Difficulty Parameter Comparison

A comparison of the IRT difficulty parameters across the two media for Groups 1 and 2 provided one assessment of the effects of using alternative CMOA on the score scale. Ideally, the parameters from the two media should fall along a diagonal (45°) line. Systematic effects on the score scale would cause the points to fall along a different line (if linearly related), or curve (if non-linearly related). Non-systematic effects would influence the degree of scatter around the line.

Figure 6-1 (a–d) displays the plots of difficulty parameters estimated from the two CMOAs, for each of the four tests. As each plot indicates, the parameters fell along the diagonal with a small degree of scatter. This result is consistent with small or negligible effects of the calibration media on the score scale.

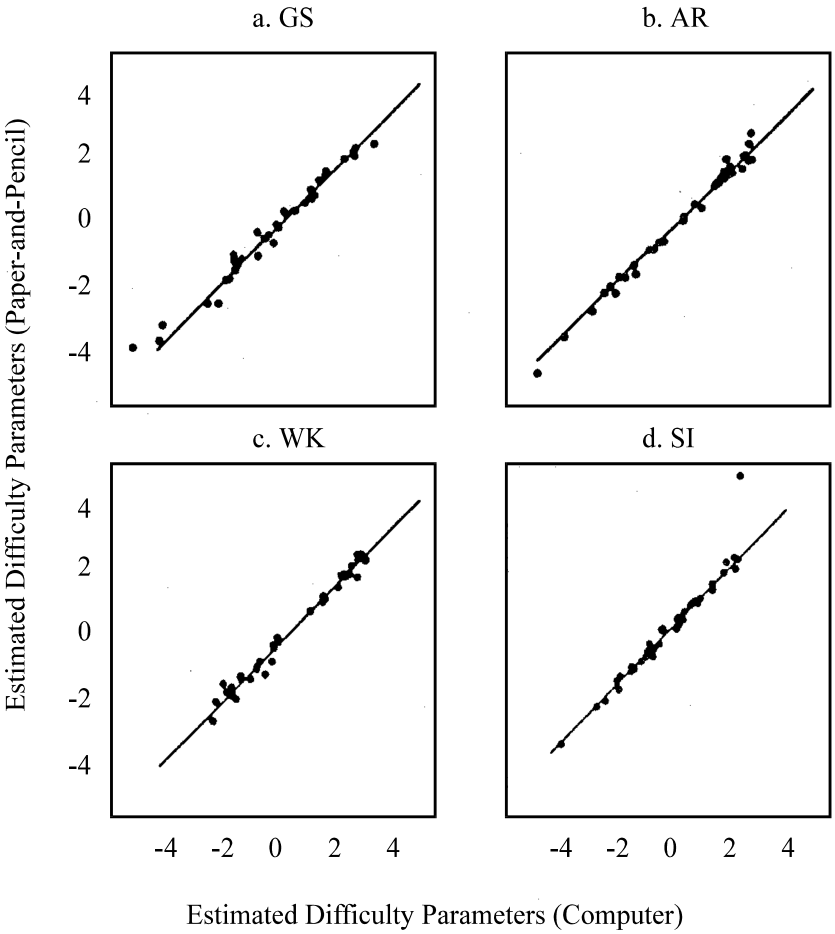


Figure 6-1. Paper-and-Pencil Versus Computer Estimated Difficulty Parameters.

Covariance Structure Analysis Results

The sample correlation matrix among the eight $\hat{\theta}$s for Group 3 is displayed in Table 6-4. Also displayed in the table are the means and standard deviations of these variables.

Table 6-4. Means, Standard Deviations, and Correlations Among Group 3 Scores

	Computer				**P&P**			
Variable	$\mathbf{X_{GSC}}$	$\mathbf{X_{ARC}}$	$\mathbf{X_{WKC}}$	$\mathbf{X_{SIC}}$	$\mathbf{X_{GSP}}$	$\mathbf{X_{ARP}}$	$\mathbf{X_{WKP}}$	$\mathbf{X_{SIP}}$
X_{GSC}								
X_{ARC}	.504							
X_{WKC}	.734	.446						
X_{SIC}	.601	.354	.496					
X_{GSP}	.970	.506	.728	.587				
X_{ARP}	.507	.981	.449	.351	.506			
X_{WKP}	.737	.450	.980	.500	.730	.451		
X_{SIP}	.605	.351	.490	.956	.587	.349	.494	
Mean	.025	–.027	.012	.042	.069	–.068	.034	.012
SD	.857	.927	.877	.866	.863	.947	.853	.896

The estimated parameters of Model 1 are displayed in Tables 6-5 and 6-6. As indicated by the $\hat{\rho}$ columns of Table 6-6, the reliability values for both CMOAs were quite high, approaching 1.0. These results indicate that a very small amount of random error among test scores was attributable to estimation errors among item parameters. The estimated σ values for each CMOA are provided in the last two columns of Table 6-6.

Table 6-5. Model 1: Estimated Disattenuated Correlation Matrix $\hat{\Phi}$

Test	**GS**	**AR**	**WK**	**SI**
GS	1.00			
AR	.52	1.00		
WK	.75	.46	1.00	
SI	.62	.36	.51	1.00

Table 6-6. Model 1: Estimated Reliabilities $\hat{\rho}$ and Standard Deviations $\hat{\sigma}$

	$\hat{\rho}$		$\hat{\sigma}$	
Test	**Computer**	**P&P**	**Computer**	**P&P**
GS	.983	.958	.857	.863
AR	.978	.985	.927	.947
WK	.976	.984	.877	.853
SI	.956	.957	.866	.896

The results of overall fit for Models 1 and 2 are displayed in Table 6-7. As indicated in this table, the likelihood ratio χ^2 value for Model 1 was nonsignificant, which provides support for the equivalency of the disattenuated correlation matrices: $\Phi_{cc} = \Phi_{pc} = \Phi_{pp}$. This result indicates that CMOA did not alter the constructs measured by the four tests.

Table 6-7. Model Evaluation: Tests of Overall Fit

Model	**Constraints**	***df***	χ^2	**p-value**
1	A	14	14.066	.44
2	A, B	18	19.267	.38

The χ^2 test based on differences between Models 1 and 2 indicated no difference between reliability parameters across the two calibration media (χ^2 = 19.267 – 14.066 = 5.201, *df* = 18 – 14 = 4, *p* = .27). This result supports the contention that the reliability of CATs is independent of the medium used to calibrate the item parameters.

Summary

The good fit of Model 1 to the data indicated that, for the four tests, the disattenuated correlations among the scores based on the computer-based calibration, Φ_{cc} did not differ significantly from the disattenuated correlations among the scores based on the P&P-based calibration, $\Phi_{pp,}$ and neither of these sets of correlations differed significantly from the disattenuated cross-correlations of scores based on the two types of calibration, $\Phi_{PC.}$ This is consistent with the lack of within-trait medium-of-administration correlational effects found by Mead and Drasgow (1993). It also extends the conclusions drawn by Mead and Drasgow to the consistency of disattenuated correlations between traits.

The results from the comparison of Models 1 and 2 indicated that, for the four tests, equal amounts of non-systematic error variance ($1- \rho^2_{tm}$) were obtained with the use of the computer-based and P&P-based item calibrations. This is generally consistent with, and extends, the findings of Divgi (1986) and Divgi and Stoloff (1986), in which the computer-based calibration was based primarily on data from adaptively administered items.

The secondary effect under investigation was the influence of calibration medium on the score scale. A comparison of the difficulty parameters across the two media indicated very little or no distortion in the scale. For all four tests, the difficulty parameters tended to fall along a diagonal (45°) line.

An important practical implication of the results of this study is that item parameters calibrated from a P&P administration of items can be used in power CATs of cognitive constructs—such as those found on the CAT-ASVAB—without changing the construct being assessed and without reducing reliability. The descriptive analyses of difficulty parameters suggest little or no effect of calibration medium on the score scale. However, Green, Bock, Linn, Lord, and Reckase (1984) noted that if scale effects do exist, they can be corrected by equating to a reference form that defines the score scale to be used for selection and classification decisions. When this is done, distortions in the mean, variance, and higher moments of the observed scores have no effect on selection and classification decisions.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Divgi, D. R. (1986). *On some issues in the Accelerated CAT-ASVAB Project* (CRM 86-231). Alexandria, VA: Center for Naval Analyses. (NTIS No. AD-A178 558

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves* (Report 86-24). Alexandria, VA: Center for Naval Analyses. (NTIS No. AD-B103 889)

Green, B. F., Bock R. D., Linn, R. L, Lord, F. M., & Reckase, M. D. (1984). A plan for scaling computerized adaptive Armed Services Vocational Aptitude Battery. *Journal of Educational Measurement*, 347-360.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.

Mead, A. D., & Drasgow, F. (1993). Effects of administration medium: A meta-analysis. *Psychological Bulletin, 114, 3*, 449-458.

Moreno, K. E., Segall, D. O., & Kieckhaefer, W. F. (1985). A validity study of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the Annual Conference of the Military Testing Association, 27, 1,*29-33. San Diego, CA: Military Testing Association. (NTIS No. AD-A172 850)

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 ,2,* 155-163.

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). Brooks AFB, TX: Air Force Human Resources Laboratory. (AD-A160 608)

SAS Institute. (1990). *SAS/STAT User's Guide* (4th ed.). Raleigh-Durham, NC: Author.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristic. *Journal of Educational Measurement, 26,* 261-271.

Weiss, D. J. (1983). *Computer-based measurement of intellectual capabilities*. Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota. (NTIS No. AD-A144 065)

Wingersky, S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.

Chapter 7

RELIABILITY AND CONSTRUCT VALIDITY OF CAT-ASVAB

One of the most important steps in evaluating the first operational forms of the computerized adaptive Armed Services Vocational Aptitude Battery (CAT-ASVAB) was to empirically demonstrate that the CAT-ASVAB tests were as reliable as their paper-and-pencil (P&P) counterparts and that they measured the same constructs. While this step is important for any new test form, this was especially true for the first two forms of CAT-ASVAB. First, computerized-adaptive testing (CAT) was a new method of testing, never having been used before in a large-scale testing program. Second, the P&P-ASVAB had a long history of use, demonstrating its predictive validity and, therefore, the importance of measuring a particular set of constructs. Third, CAT-ASVAB and P&P-ASVAB would be in use at the same time, and scores from the two versions must be interchangeable.

Data collection for this study was conducted in 1988-89, after development of the CAT-ASVAB item pools, initial evaluation of these pools, and development of the Hewlitt-Packard (HP)-based CAT-ASVAB system. Data analyses were completed early in 1990, and results of the study played a significant role in the decision to use CAT-ASVAB operationally.

Earlier Research

Earlier studies showed that CAT results in more reliable scores than conventional P&P testing methods. Kingsbury and Weiss (1981) found that the alternate-form reliability for a computerized adaptive word knowledge test was higher than that of a corresponding conventional test administered by computer. McBride and Martin (1983) found that adaptive verbal and arithmetic reasoning tests were more reliable than corresponding conventional tests administered by computer.

Previous studies have also shown that the adaptive testing methodology can be used to measure constructs traditionally assessed by conventional, paper-and-pencil tests. A comparison of the relationship between three CAT-ASVAB and corresponding P&P-ASVAB tests showed that the patterns of factor loadings for the two versions were very similar (Moreno, Wetzel, McBride, & Weiss, 1984). A validity study comparing an experimental version of CAT-ASVAB to the P&P-ASVAB found the same result (Moreno, Segall, & Kieckhaefer, 1985). In a meta-analysis of such studies, Mead and Drasgow (1993) found that medium of administration—computer versus paper-and-pencil—has little effect on power tests. Results for speeded tests were mixed.

These studies, as a whole, provided valuable information on the reliability and validity of CAT instruments. However, until the study described in this chapter was conducted, only a limited number of content areas had been examined in other research studies. In addition, the reliability and construct validity of a test is dependent on the quality of the item pools and the item selection and scoring procedures. The study described in this chapter provided information on the reliability and validity of tests in the first two operational forms of CAT-ASVAB—01C and 02C.

Method

Design

This study used an equivalent-groups design, with examinees randomly assigned to one of two groups. Group 1 was administered Form 01C of the CAT-ASVAB in the first testing session, followed by Form 02C of the CAT-ASVAB in the second session. Group 2 was administered Form 9B of the P&P-ASVAB, followed by Form 10B of the P&P-ASVAB. There was an interval of five weeks between the first test and the second test. This interval was constant for all examinees. A five-week interval was chosen because applicants taking the ASVAB must wait 30 days before retesting.

Examinees

Two thousand ninety male Navy recruits stationed at the Recruit Training Center in San Diego, CA, served as examinees in this study: 1,057 in the CAT-ASVAB group and 1,033 in the P&P-ASVAB group. A substantial percentage of the subjects did not have complete data because they did not return for the second of the two tests. After examinees with incomplete data were eliminated, the sample sizes were 744 for CAT-ASVAB and 726 for P&P-ASVAB.

Test Instruments

P&P-ASVAB. The P&P-ASVAB consists of ten tests: eight power tests and two speeded tests. (Note: The two speeded tests are no longer part of the ASVAB, effective January 2002.) Each test consists of items with difficulty levels that span the range of abilities found in the military applicant population. Most tests, however, are peaked at the middle of the ability distribution. There are six forms of the P&P-ASVAB in operational use at any given time. All operational forms have been equated to a common P&P-ASVAB reference form (8A).

CAT-ASVAB. CAT-ASVAB forms 01C and 02C were used in this study. These are the two forms that were developed for initial operational implementation of CAT-ASVAB. Item pool development is described in Chapter 2 of this technical bulletin. The psychometric procedures used to administer the tests were identical to those used operationally, and are described in Chapter 3. The computer system used to administer the tests was the HP-IPC, described in Chapter 5.

Procedures

All examinees had taken an operational P&P-ASVAB to qualify for entrance into the Navy. As part of the present study, they took either a non-operational CAT-ASVAB or a non-operational P&P-ASVAB, with the scores used for experimental purposes only. Upon arrival at the test site, examinees were given general instructions explaining the experimental testing and signed a privacy act statement allowing use of the data for research purposes. Then they were seated in the appropriate room (CAT-ASVAB or P&P-ASVAB), based on a random-assignment list. CAT-ASVAB was administered following procedures developed for operational implementation; P&P-ASVAB was administered following procedures outlined in the ASVAB Test Administrator Manual. At the conclusion of testing, the Test Administrators (TAs) collected additional data from the examinee's personnel records, including population group, ethnic group, date of birth, education, operational ASVAB test form, operational ASVAB test scores, and date of enlistment.

Scores

All analyses for both the CAT-ASVAB and the P&P-ASVAB tests were based on standard scores. ASVAB standard scores are scaled to have a mean of 50 and a standard deviation of 10 in the 1980 youth population (DoD, 1982). Since CAT-ASVAB is equated to P&P-ASVAB Form 8A, standard scores for the CAT-ASVAB tests were obtained by converting the final theta estimate to the equated raw score and then using P&P-ASVAB Form 8A conversion tables to obtain standard scores.

Data Editing

A data editing procedure which compared non-operational scores to operational scores was used to eliminate "unmotivated" examinees (Segall, 1996). After editing, the sample size was 701 for the CAT-ASVAB group and 687 for the P&P-ASVAB group. One limitation of the structural modeling procedure, CALIS (SAS Institute, 1990), is that samples used in multi-group analyses must be of equal size; to satisfy this requirement, 14 examinees were selected at random and deleted from the CAT group. The final sample size in both groups was 687.

Data Analyses

Evaluation of equivalent groups. To assure the equivalency of the two samples, demographic variables were checked by (a) comparing the two groups on race and years of education, and (b) comparing the distribution of operational test scores by the two groups.

No significant differences between the CAT and P&P groups were found on race or years of education. For both variables, an χ^2 test for the differences between distributions indicated no significant difference. For each test of the operational ASVAB, a Kolmogorov-Smirnov [K-S] test was conducted to evaluate the difference between the score distributions for the two groups. There were no significant differences among the ten tests examined.

Correlational analyses. To compare alternate form reliabilities, Pearson product-moment correlations were computed between alternate forms of both batteries: CAT-ASVAB and P&P-ASVAB. Fisher's ***z*** transformation was used to evaluate the difference between CAT-ASVAB and P&P-ASVAB reliabilities for each content area. Cross-medium Pearson product-moment correlations were computed between examinee performance on CAT-ASVAB tests and operational P&P-ASVAB tests and compared to correlations between non-operational and operational P&P-ASVAB tests.

Structural Analysis. If CAT-ASVAB and P&P-ASVAB are to be used interchangeably, it is essential for the two versions of the battery to measure the same traits. This issue was investigated using structural modeling. The analysis described below was performed separately for each of the ten content areas contained within the ASVAB. To begin, we defined six variables that represent standardized test scores on different versions of the ASVAB. The notational convention is provided in Table 7-1. All six variables were assumed to represent a single content area (e.g., General Science).

Table 7-1. Variable Definitions for the Structural Analysis

Variable	Medium	Form	Group
C_1	CAT	1	CAT
C_2	CAT	2	CAT
X_o^c	P&P	Operational	CAT
X_1	P&P	9B	P&P
X_2	P&P	10B	P&P
X_o^p	P&P	Operational	P&P

Further, let $\mathbf{\Sigma}_c$ represent the 3 × 3 covariance matrix of C_1, C_2, X_o^c (for the CAT group) and $\mathbf{\Sigma}_p$ represent the 3 × 3 covariance matrix of X_1, X_2, X_o^p (for the P&P group). Each covariance matrix can be expressed in terms of several parameter matrices:

$$\sum\nolimits_c = \Lambda_c \left(R_c \Phi_c R_c - R_c^2 + I \right) \Lambda_c \tag{7-1}$$

and

$$\sum_p \;=\; \Lambda_p \left(R_p \;\; \Phi_p \;\; R_p - R_p^2 + I \right) \Lambda_P \, .$$

(7-2)

The model given by Equation 7-1 refers to the covariance matrix among three tests measuring a common content area (two CAT forms and one P&P form) for the CAT group. The model given by Equation 7-2 refers to the covariance matrix among three tests measuring the same content area (three P&P forms) for the P&P group. In Equation 7-1 the parameter matrices for the CAT group take the following form:

$$\Lambda_c = \begin{pmatrix} \sigma(C_1) & 0 & 0 \\ 0 & \sigma(C_2) & 0 \\ 0 & 0 & \sigma(X_o^c) \end{pmatrix},$$

$$R_c = \begin{pmatrix} \sqrt{\rho(C_1)} & 0 & 0 \\ 0 & \sqrt{\rho(C_2)} & 0 \\ 0 & 0 & \sqrt{\rho(X_o)} \end{pmatrix},$$

and

$$\Phi_c = \begin{pmatrix} 1 & 1 & \rho(C, X_o) \\ 1 & 1 & \rho(C, C_o) \\ \rho(C, X_o) & \rho(C, X_o) & 1 \end{pmatrix},$$

(7-3)

where $\sigma(C_1)$, $\sigma(C_2)$, and $\sigma(X_o^c)$ denote the standard deviations of C_1, C_2, and X_o^c, respectively, and $\rho(C_1)$, $\rho(C_2)$, and $\rho(X_o)$ denote the reliabilities of C_1, C_2 , and X_o^c. In Equation 7-3, we assume that $\rho(C_1, X_o) = \rho(C_2, X_o)$ [= $\rho(C, X_o)$], where $\rho(Y, Z)$ denotes the correlation between Y and Z, corrected for attenuation.

In the model given in Equation 7-1, the $\mathbf{\Phi}_c$ matrix represents the disattenuated correlation matrix among C_1, C_2, and X_o^c. From classical test theory, we see that the product $\mathbf{R}_c\,\mathbf{\Phi}_c\,\mathbf{R}_c$ provides the correlation matrix of observed variables, with the diagonal elements equal to the test reliabilities. Consequently, the sum $\mathbf{R}_c\,\mathbf{\Phi}_c\,\mathbf{R}_c - \mathbf{R}_c^2 + \mathbf{I}$ provides the correlation matrix among the observed C_1, C_2, and X_o^c, with ones in the diagonal. Finally, by pre- and post-multiplying this correlation matrix by $\mathbf{\Lambda}_c$ (which contains the standard deviations), we obtain $\mathbf{\Sigma}_c$, the covariance matrix among the observed C_1, C_2, and X_o^c.

The parameter matrices for the P&P group model, given by Equation 7-2, take on a similar form:

$$\Lambda_p = \begin{pmatrix} \sigma(X_1) & 0 & 0 \\ 0 & \sigma(X_2) & 0 \\ 0 & 0 & \sigma(X_o^p) \end{pmatrix}, \tag{7-4}$$

$$R_p = \begin{pmatrix} \sqrt{\rho(X_1)} & 0 & 0 \\ 0 & \sqrt{\rho(X_2)} & 0 \\ 0 & 0 & \sqrt{\rho(X_o)} \end{pmatrix}, \tag{7-5}$$

and

$$\Phi_p = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \tag{7-6}$$

where $\sigma(X_1)$, $\sigma(X_2)$, and $\sigma(X_o^p)$ denote the standard deviations of X_1, X_2, and X_o^p, and $\rho(X_1)$ and $\rho(X_2)$ denote the reliabilities of X_1 and X_2.

Several constraints imposed by the model should be noted. First, the reliability of X_o^p is assumed to be equivalent to the reliability of X_o^c. That is, the reliability of the operational form is assumed to be equivalent for the CAT and P&P groups. This assumption is imposed by constraining the lower diagonal elements of the $\mathbf{R}_c$ and $\mathbf{R}_p$ matrices to be equal. Second, the disattenuated correlation between the two CAT forms is assumed to be 1. This constraint is imposed by fixing the (2, 1) element (and its transpose) of the $\mathbf{\Phi}_c$ matrix equal to 1. We make an additional assumption, which is consistent with this constraint, that $\rho(C_1, X_\mathrm{o}) = \rho(C_2, X_\mathrm{o})$. That is, we assume that the disattenuated correlation between CAT and P&P is the same for both forms of CAT. This assumption is imposed by constraining the appropriate elements of the $\mathbf{\Phi}_c$ matrix to be equivalent. Third, the disattenuated correlations among the P&P-ASVAB forms (for the P&P group) are assumed to be equal to 1. This constraint is imposed by fixing all elements of the $\mathbf{\Phi}_p$ matrix equal to 1.

The multigroup model given by Equations 7-1 and 7-2 is exactly identified since there are 12 unknown parameters and 12 non-redundant covariance elements among the two 3 × 3 covariance matrices. These 12 parameters were estimated by normal-theory maximum-likelihood using the SAS procedure CALIS (SAS Institute, 1990).

Results and Discussion

Table 7-2 displays the correlations between alternate forms for CAT-ASVAB and P&P-ASVAB. Seven of the ten CAT-ASVAB tests displayed significantly higher alternate form reliabilities than the corresponding P&P-ASVAB tests. The other three tests displayed non-significant differences. Also displayed in Table 7-2 are the correlations between the operational and non-operational forms for the CAT and P&P groups. It is important to note that CAT-ASVAB tests correlated as highly with the operational P&P-ASVAB as did alternate forms of the P&P-ASVAB.

Table 7-2. Alternate Form and Cross-Medium Correlations

	Alternate Form Reliability		**Correlations With Operational P&P-ASVAB**			
Test	**CAT**	**P&P**	**CAT Form 01C**	**CAT- Form 02C**	**P&P Form 9B**	**P&P Form 10B**
General Science	.843**	.735	.83	.82	.79	.73
Arithmetic Reasoning	.826**	.773	.81	.75	.76	.72
Word Knowledge	.832	.811	.83	.81	.81	.78
Paragraph Comprehension	.535	.475	.54	.43	.48	.38
Numerical Operations	.817**	.708	.60	.60	.65	.56
Coding Speed	.770	.747	.57	.54	.65	.62
Auto and Shop Information	.891**	.776	.83	.83	.76	.74
Mathematics Knowledge	.883**	.819	.86	.83	.83	.80
Mechanical Comprehension	.749*	.703	.69	.64	.66	.65
Electronics Information	.727**	.648	.73	.72	.66	.65

* Statistically significant ($p < .05$) ** Statistically significant ($p < .01$)

A separate covariance analysis was performed for each of the ten content areas contained within the ASVAB. Table 7-3 lists the estimated reliabilities for CAT-ASVAB and P&P-ASVAB forms. Table 7-4 provides $\hat{\rho}(C, X_o)$, the maximum likelihood estimate of the disattenuated correlation between CAT and P&P. Table 7-4 also provides SE ($\hat{\rho}$), the asymptotic standard error of $\hat{\rho}$ (C, X_o).

Table 7-3. Test Reliabilities

Test	CAT-ASVAB			P&P-ASVAB		
	$\hat{\rho}(C_1)$	$\hat{\rho}(C_2)$		$\hat{\rho}(X_1)$	$\hat{\rho}(X_2)$	$\hat{\rho}(X_o)$
General Science	.86	.82		.80	.67	.78
Arithmetic Reasoning	.89	.77		.82	.73	.72
Word Knowledge	.86	.81		.84	.79	.78
Paragraph Comprehension	.67	.43		.59	.38	.37
Numerical Operations	.79	.84		.82	.61	.52
Coding Speed	.81	.73		.79	.70	.54
Auto and Shop Information	.89	.89		.80	.76	.74
Mathematics Knowledge	.92	.85		.85	.79	.80
Mechanical Comprehension	.80	.70		.73	.68	.61
Electronics Information	.74	.71		.66	.64	.66

The hypothesis that $\rho(C, X_o) = 1$ was tested for each content area by fixing all elements of Φ_c equal to 1 and re-estimating the remaining model parameters. The χ^2 goodness-of-fit measure provides a test of the null hypothesis that $\rho(C, X_o) = 1$. Under the null hypothesis, this measure is χ^2 -distributed with $df = 1$. The χ^2 and *p*-values for each content area are listed in the last two columns of Table 7-4.

Table 7-4. Disattenuated Correlations Between CAT- and P&P-ASVAB

Test	$\hat{\rho}(C, X_o)$	$\mathrm{SE}(\hat{\rho})$	$\chi^2(\mathrm{df} = 1)$	p
General Science	1.01	.018	.55	.456
Arithmetic Reasoning	1.02	.021	1.13	.287
Word Knowledge	1.02	.017	.80	.370
Paragraph Comprehension	1.11	.082	2.12	.145
Numerical Operations	.94	.044	1.73	.189
Coding Speed	.86	.043	9.12	.002
Auto and Shop Information	1.02	.020	.83	.363
Mathematics Knowledge	1.00	.015	.001	.975
Mechanical Comprehension	.99	.035	.13	.715
Electronics Information	1.05	.031	3.20	.074

The test reliabilities shown in Table 7-3 display the same pattern of differences across media as those shown in Table 7-2. The multigroup model provides a separate reliability estimate for each form, whereas the analysis provided in Table 7-2 provides a single estimate. However, for each content area, the alternate form correlations (Table 7-2) fall at about the midpoint of the two separate reliability estimates given in Table 7-3. For example, the GS (CAT-ASVAB) alternate form correlation of .84 (Table 7-2) falls at the midpoint of the separate Form 01C and 02C reliabilities of .82 and .86 (Table 7-3). A similar pattern is evident for other tests.

From Table 7-4, we observe that the first forms administered (C_1 and X_1) tended to have higher reliabilities than the second forms administered (either C_2 or X_2). That is, for most tests we observe that $\hat{\rho}(C_1) > \hat{\rho}(C_2)$ and $\hat{\rho}(X_1) > \hat{\rho}(X_2)$. This pattern is evident for both CAT-ASVAB and P&P-ASVAB. One possible cause is a difference in precision between the forms. Another possible cause is motivation: examinees tend to be less motivated for the second administration of the battery than for the first. Since the order of form administration was not counterbalanced (CAT Form 01C and P&P Form 9B were always administered first, followed by CAT Form 02C or P&P Form 10B), it is impossible to isolate the cause of the difference. However, since the construction procedures for both CAT-ASVAB and P&P-ASVAB attempted to ensure equal precision among forms, and the simulations results reported in Chapter 2 of this technical bulletin indicate that this goal was achieved, we speculate that the within-medium differences in reliabilities are due to motivational effects. Table 7-4 displays $\hat{\rho}(C, X_o)$, the disattenuated correlations between CAT-ASVAB and the operational P&P-ASVAB. Although the theoretical upper limit of a correlation coefficient is 1.00, no upper bound was placed on the estimates obtained in this analysis. However, those estimates exceeding 1.00 imply that the population disattenuated correlation is equal to or less than 1.

As indicated by the significance tests in Table 7-4, only one test displayed a disattenuated correlation significantly different from 1. This was the non-adaptive speeded test, Coding Speed (CS). This test had an estimated disattenuated correlation of .86 ($\chi^2 = 9.12$, $df = 1$, $p = .002$). We know from examinee feedback that some had difficulty understanding the instructions that are administered by computer. During P&P-ASVAB administration, test administrators often work through several examples to help examinees understand the task. Although several example questions are given on the CAT-ASVAB for CS, some examinees may need more practice. Because of the difficulty in understanding the CAT-ASVAB instructions for CS, the CAT version may have had a higher general ability (“g”) component than its P&P counterpart.

The findings (from Table7-4) indicate that none of the disattenuated correlations between CAT-ASVAB and P&P-ASVAB power tests were significantly different from 1.00. Of course, one reason for this lack of significance may be due to a lack of power to detect small- or moderate-sized differences. However, the standard error of estimate of $\hat{\rho}$, (SE($\hat{\rho}$)), displays a narrow confidence interval around nearly all estimated correlations. Consequently, even if the population $\rho(C, X_o)$ is less than 1 for one or more adaptive tests, it is improbable that it would fall below .97. This is true for nearly all adaptive tests examined.

Summary

Taken together, the estimated test reliabilities and disattenuated cross-medium correlations provide a compelling case for the virtues of CAT. Many concerns about the validity of CAT scores have been cited in the literature. These concerns include the impact of medium of administration (i.e., use of computers to administer tests), adaptive item selection, item-response theory (IRT) techniques used in scoring, and paper-and-pencil calibration of item parameters. The findings of this study indicate that the aggregate effect of these threats to reliability and validity appears to be minimal or non-existent. The results demonstrate that the adaptive tests within CAT-ASVAB measure the same traits measured by the P&P-ASVAB, with equal or greater precision, and with test lengths only half as long as their P&P counterparts.

References

Department of Defense. (1982). *Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery*. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics).

Kingsbury, G. G., & Weiss, D. J. (1981). *A validity comparison of adaptive and conventional strategies for mastery testing* (Report 81-3). Minneapolis, MN: Department of Psychology, University of Minnesota.

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*, 223-235. New York, NY: Academic Press.

Mead, A. D., & Drasgow, F. (1993). Effects of administration medium: A meta-analysis. *Psychological Bulletin, 114, 3,* 449-458.

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 ,2,* 155-163.

Moreno, K. E., Segall, D. O., & Kieckhaefer, W. F. (1985). A validity study of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the Annual Conference of the Military Testing Association, 27, 1,*29-33. San Diego, CA: Military Testing Association. (NTIS No. AD-A172 850)

SAS Institute. (1990). *SAS/STAT User's Guide* (4th ed.). Raleigh-Durham, NC: Author.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Chapter 8

EVALUATING THE PREDICTIVE VALIDITY OF CAT-ASVAB

Although computerized-adaptive testing (CAT) can be expected to improve reliability and measurement precision, the increased reliability does not necessarily translate into substantially greater validity. In fact, there is always a danger when changing item content or format that the new test may be measuring a slightly different ability, which may not relate to, or predict outcomes as well as, the old test. Findings of the earlier validity study of the experimental item pools (Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997), therefore, did not necessarily generalize to the new, operational item pools.

The purpose of the research reported here was to evaluate whether the predictive validities of CAT-ASVAB forms 01C and 02C are as high as the paper-and-pencil (P&P)-ASVAB. A secondary purpose was to verify that the CAT-ASVAB tests are measuring the same abilities as their P&P-ASVAB counterparts. While the construct validity of the operational CAT-ASVAB forms had already been evaluated as part of an alternate forms study (Chapter 7 of this technical bulletin), data collected as part of this study provided an opportunity for a second check.

The research was designed to answer three questions:

1. Whether the means and standard deviations of the pre-enlistment ASVAB scores were the same for the CAT and P&P

groups. This test was done to verify that the groups were equivalent.

2. Whether the correlations between pre-enlistment and post-enlistment ASVAB were the same for CAT and P&P groups. This test was done to verify that the two media of test administration measured the same abilities.
3. Whether the validities of the tests for predicting final school grades (FSGs) were the same for P&P-ASVAB and CAT-ASVAB.

Method

Participants in this study were drawn from Navy recruits at the Navy Recruiting Center at Great Lakes, IL. Subjects were in one of two research projects—the Navy Validity Study of New Predictors (NVSNP) or the Enhanced Computer Administered Test (ECAT) study. Recruits were chosen for participation in the present study if they had been pre-assigned to enter one of a specified list of technical schools following their basic training. They were randomly assigned to either CAT-ASVAB or P&P-ASVAB test groups. Some months later, the school records were obtained to determine the examinees' FSGs and other criteria of school performance. The examinees' pre-enlistment ASVAB scores were also obtained.

For the ASVAB (post-enlistment) testing at Great Lakes, the recruits spent a morning as subjects in the NVSNP or ECAT experiments. In the afternoon, for the present study, they were administered either the CAT-ASVAB or the P&P-ASVAB in separate rooms. A computer program at the test site that used a random number generator made assignments between the two conditions.

Table 8-1 gives sample sizes and school lists for the recruits. The sample sizes are for “school completers” who had FSGs of record. The rows labeled “Others” show examinees who took the post-enlistment test at Great Lakes but who had no FSGs of record. They include recruits who never went to the designated schools or who dropped out before completing training.

Table 8-1. CAT and P&P Sample Sizes

Code	School	CAT	P&P
Navy Validity Study of New Predictors Study			
AD	Aviation Machinist's Mate	49	43
AMS	Aviation Structural Mechanic - Structures	43	46
AO	Aviation Ordnanceman	49	45
BT/MM	Boiler Technician/Machinist Mate	408	401
GMG	Gunner's Mate - Phase I	155	169
HM	Hospitalman	230	255
HT	Hull Technician	152	170
OS	Operations Specialist	457	447
Enhanced Computer Administered Test Study			
AC	Air Traffic Controller	29	21
AE	Aviation Electrician's Mate	80	91
AMS	Aviation Structural Mechanic - Structures	75	61
AO	Aviation Ordnanceman	78	59
AV	Avionics Technician (AT, AQ, AX)	184	179
EM	Electrician's Mate	402	375
EN	Engineman	356	378
ET	Electronics Technician	29	30
FC	Fire Controlman	370	399
GMG	Gunner's Mate - Phase I	221	195
MM	Machinist Mate	368	409
OS	Operations Specialist	367	333
RM	Radioman	18	16
Total	School Completions	4,120	4,122
Others	Others tested	1,550	1,599

Statistical Analyses

The equivalence of means and standard deviations was tested with a t-test for differences in means and the F-test for ratios of variances, respectively. To correct for any differences between the groups, validities and pre-post correlations were corrected for range restriction, based on their correlations with the pre-enlistment ASVAB, using the 1991 Joint-Services recruit population (N = 650,278) as the reference population and all ten ASVAB tests as explicitly selected variables (see Wolfe, Alderton, Larson, Bloxom, & Wise, 1997). Post-enlistment scores were treated as implicitly selected. Corrections were made separately in each sample.

The pre-post uncorrected correlation differences were tested with the Fisher transformation: $Z = \tanh^{-1}(r)$. Let r_1 be the pre-post correlation for the CAT group and r_2 be the pre-post correlation for the P&P group. The following Z is approximately normally (0,1) distributed:

$$Z = \frac{\tanh^{-1}(r_1) - \tanh^{-1}(r_2)}{\sqrt{\dfrac{1}{N_1 - 3} + \dfrac{1}{N_2 - 3}}} \qquad (8\text{-}1)$$

The pre-post corrected correlation differences were tested using a modified version of an asymptotic test developed by Hedges, Becker, and Wolfe (1992), where N-2 replaces N in the original formula to produce better performance in small samples (see Samiuddin, 1970). Let corrected correlations be designated by capital R and uncorrected correlations by lower case r. Let $c = R/r$. The following Z is asymptotically normally (0,1) distributed:

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{\left[c_1\left(1-r_1^2\right)\right]^2}{N_1 - 2} + \frac{\left[c_2\left(1-r_2^2\right)\right]^2}{N_2 - 2}}} \qquad (8\text{-}2)$$

Validities of each test for predicting FSG in each school sample were computed and corrected for range restriction. Differences in validities were tested using the same formulas as above. Because many of the sample sizes were small, it was necessary to combine evidence across samples. For each ASVAB test, a combined Z was computed by the formula

$$Z = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \qquad (8\text{-}3)$$

where i ranges over the k = 21 samples. The combined Z was referred to the normal (0,1) distribution for significance.

The final results were expressed in terms of significance tests for each ASVAB test. No attempt was made to explicitly adjust the significance levels to correct for the multiple significance tests performed in the study. Isolated results that were “significant” at the p < .05 level should generally be disregarded, since one would occur 40 percent of the time in any set of 10 hypothesis tests if they were independent. In the ASVAB they are not independent, of course, but similar considerations apply.

Results

Table 8-2 compares the *pre-enlistment* ASVAB scores for the CAT and P&P groups. There are no significant differences between the CAT and P&P groups in their means on pre-enlistment ASVAB tests. In comparing standard deviations, a "significantly" larger value was found for the CAT Paragraph Comprehension (PC) test, but the result could be spurious since 24 significance tests were performed in this table. The randomization procedure for allocating examinees between conditions should be considered successful.

Table 8-2. Pre-Enlistment ASVAB Comparison for the CAT and P&P Groups

	Mean			**Standard Deviation**		
ASVAB Test	**CAT**	**P&P**	**t Diff.**	**CAT**	**P&P**	**F Diff.**
General Science (GS)	52.99	52.98	0.10	7.26	7.11	1.04
Arithmetic Reasoning (AR)	52.51	52.48	0.19	6.92	6.94	1.01
Word Knowledge (WK)	52.55	52.64	-0.96	5.22	5.25	1.01
Paragraph Comprehension (PC)	52.83	52.94	-1.01	5.78	5.62	1.06*
Numerical Operations (NO)	53.73	53.82	-0.73	6.65	6.56	1.03
Coding Speed (CS)	52.47	52.40	0.57	6.81	6.85	1.01
Auto and Shop Information (AS)	53.98	53.83	0.95	7.96	7.88	1.02
Mathematics Knowledge (MK)	54.26	54.27	-0.10	6.62	6.58	1.01
Mechanical Comprehension (MC)	54.32	54.25	0.44	7.81	7.75	1.02
Electronics Information (EI)	52.59	52.52	0.52	7.80	7.72	1.02
Verbal (VE) = [WK + PC]	52.73	52.83	-1.05	5.00	4.99	1.00
AFQT = [VE + AR + NO/2]	58.39	58.50	-0.35	17.32	17.08	1.03

* $p < .05$
N: CAT = 5,670; P&P = 5,721

Table 8-3 shows the correlations between the CAT-ASVAB tests and the pre-enlistment ASVAB, the correlations between the post-enlistment P&P-ASVAB and the pre-enlistment ASVAB, and their differences. Since examinees were selected on the basis of their pre-enlistment scores, range-corrected results were calculated. Nine of the tests differ significantly in their uncorrected pre-post correlations, but this number shrinks to three in the corrected analysis. NO and CS, the two speeded nonadaptive tests in the CAT-ASVAB, had significantly lower correlations with the corresponding pre-enlistment tests than did the P&P tests, indicating that they measure a different construct or measure the same construct differently. The CAT-ASVAB speeded tests were scored with a rate score—the proportion correct (corrected for guessing) divided by the mean of all screen times—whereas the P&P speeded tests were scored by number of items correct within a given time limit. The latter measure has the disadvantage of having a ceiling, which many examinees attained, of all items correct within the time limit. The computerized version is able to distinguish between fast and very fast examinees, but the shape of the score distribution changed so that it did not correlate with the pre-enlistment test as well as another P&P test can.

Table 8-3. Pre-Post Correlations for Combined Navy and ECAT Samples

	CAT-ASVAB		**P&P-ASVAB**		**Z of Difference**	
Test	**Uncorrected**	**Corrected**	**Uncorrected**	**Corrected**	**Uncorrected**	**Corrected**
GS	.718	.812	.716	.812	0.22	0.00
AR	.752	.843	.719	.821	3.84**	2.26*
WK	.558	.719	.587	.747	-2.30*	-1.73
PC	.424	.634	.383	.597	2.61**	1.54
NO	.591	.696	.643	.734	-4.49**	-2.82**
CS	.603	.692	.665	.733	-5.54**	-3.24**
AS	.808	.842	.784	.835	3.50**	0.97
MK	.743	.834	.734	.839	1.06	-0.52
MC	.651	.745	.626	.733	2.25*	0.93
EI	.623	.712	.634	.729	-0.97	-1.31
VE	.762	.866	.733	.852	3.51**	1.47
AFQT	.830	.915	.810	.907	3.26**	1.17

* $p < .05$ ** $p < .01$

Table 8-4 shows the predictive validity coefficients for both pre-enlistment and post-enlistment ASVAB for predicting final school performance for the CAT and P&P groups. Note that the uncorrected pre-enlistment validities were usually lower than their post-enlistment counterparts, but this was not true for the corrected validities. Among the 48 significance tests presented in this table, two, uncorrected WK and corrected AS, were barely “significant” at the .05 level, a result that could easily occur by chance. The two computerized speeded tests that had significantly lower pre-post correlations in Table 8-3 have validities that were at least as high as the P&P versions.

Table 8-4. CAT Group and P&P Group Predictive Validities for School Final Grades

	Uncorrected			Range-Corrected		
Test	**CAT**	**P&P**	**Z (diff)**	**CAT**	**P&P**	**Z (diff)**
Pre-Enlistment ASVAB						
GS	.232	.249	-1.34	.531	.513	0.07
AR	.330	.319	0.81	.603	.576	0.29
WK	.202	.216	-0.62	.468	.473	-0.28
PC	.204	.222	-1.04	.467	.466	-0.17
NO	.118	.135	-1.04	.351	.348	0.15
CS	.193	.150	1.19	.362	.350	0.44
AS	.192	.215	-0.69	.370	.373	-0.35
MK	.298	.261	1.19	.559	.544	0.46
MC	.263	.289	-0.84	.505	.499	-0.48
EI	.220	.250	-0.94	.457	.457	-0.49
VE	.225	.246	-1.08	.495	.487	-0.28
AFQT	.376	.373	-0.30	.626	.615	-0.04
Post-Enlistment ASVAB						
GS	.244	.231	0.84	.528	.477	0.41
AR	.337	.328	0.25	.580	.556	0.26
WK	.227	.272	-1.98*	.476	.503	-0.87
PC	.260	.243	0.83	.510	.461	0.87
NO	.136	.133	-0.31	.377	.321	0.82
CS	.226	.182	1.32	.395	.320	1.38
AS	.174	.220	-1.38	.310	.428	2.01*
MK	.286	.319	-1.68	.521	.530	-0.79
MC	.273	.286	-0.25	.505	.516	-0.51
EI	.231	.267	-1.90	.453	.492	-0.81
VE	.258	.284	-1.06	.528	.519	-0.05
AFQT	.387	.396	-0.68	.630	.617	-0.73

* $p < .05$ N: CAT = 4,120; P&P = 4,122

Summary

The results of this research show no reason to doubt that CAT-ASVAB is as valid as P&P-ASVAB. The two computerized speeded tests yield measures that are not precisely equivalent to their P&P counterparts, but they may be better in some ways and are no less valid. The results of this study support the findings reported in Chapter 7 of this technical bulletin.

References

Hedges, L. V., Becker, B. J., & Wolfe, J. H. (1992). *Detecting and measuring improvements in validity* (TR-93-2). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. AD-A257 446)

Samiuddin, M. (1970). On a test for an assigned value of correlation in a bivariate normal distribution. *Biometrika, 57*, 461-464.

Segall, D. O., Moreno, K. E., Kieckhaefer, 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.

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.

Chapter 9

EQUATING THE CAT-ASVAB

During an extended operational test and evaluation (OT&E) phase, both the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) and the paper-and-pencil version of the battery (P&P-ASVAB) were used operationally to test applicants for the Military Services (see Chapter 10 of this technical bulletin). At some testing sites, applicants were accessed using scores from the CAT-ASVAB, while at most other sites applicants were enlisted using scores obtained on the P&P-ASVAB. To make comparable enlistment decisions across the adaptive and conventional versions, an equivalence relation (or equating) between CAT-ASVAB and P&P-ASVAB was obtained. The primary objective of this equating was to provide a transformation of CAT-ASVAB scores that preserves the flow rates currently associated with the P&P-ASVAB. In principle, this can be achieved by matching the P&P-ASVAB and equated CAT-ASVAB test and composite distributions. This equating allowed cut scores associated with the existing P&P-ASVAB scale to be applied to the transformed CAT-ASVAB scores without affecting qualification rates.

The equating study was designed to address three concerns. First, the equating transformation applied to CAT-ASVAB scores should preserve flow rates associated with the existing cut scores based on the P&P-ASVAB score scale. Second, the equating transformation should be based on operationally motivated applicants, since the effect of motivation on CAT-ASVAB equating has not been thoroughly studied. Third, subgroup members taking CAT-ASVAB should not be placed at a disadvantage relative to their subgroup counterparts taking the P&P-ASVAB.

The first concern was addressed by using an equipercentile procedure for equating the CAT-ASVAB and the P&P-ASVAB. By definition, this equating procedure identifies the transformation of scale that matches the cumulative distribution functions. Although this procedure was applied at the test level, the distributions of all selector composites were also evaluated to ensure that no significant differences existed across the adaptive and conventional versions.

The concern over motivation was addressed by conducting the CAT-ASVAB equating in two phases: (a) score equating development (SED), and (b) score equating verification (SEV). The purpose of SED was to provide an interim equating of the CAT-ASVAB. During that study, both CAT-ASVAB and P&P-ASVAB were given non-operationally to randomly equivalent groups. The tests were non-operational in the sense that the performance on the tests had no impact on examinees' eligibility for the military—all participants in the study were also administered an operational P&P-ASVAB form that was used for enlistment decisions. This interim equating was used in the second phase (SEV) to select and classify military applicants. During the SEV phase, applicants were administered either an operational CAT-ASVAB or an operational P&P-ASVAB. Both versions used in the SEV study did have an impact on applicants' eligibility for Military Service. This new equating obtained in SEV was based on operationally motivated examinees and was later applied to applicants participating in the OT&E study.

The third concern, regarding subgroup performance, was addressed through a series of analyses conducted on data collected during the score equating study. Analyses examined the performance of blacks and females for randomly equivalent groups assigned to CAT-ASVAB and P&P-ASVAB conditions.

This chapter describes the essential elements of the CAT-ASVAB equating. These include the data collection design, sample characteristics, smoothing and equating procedures, composite equatings, and subgroup performance.

Data Collection Design and Procedures

Data for the SED and SEV equating studies were collected from six geographically dispersed regions within the continental United States: Boston, MA; Richmond, VA; Jackson, MS; Omaha, NE; San Diego, CA; and Seattle, WA. Within each region is a Military Entrance Processing Station (MEPS), and associated with each MEPS are a number (between 3 and 16) of Mobile Examining Team Sites (METSs). Each MEPS and associated METSs were included in the data collection for a two- to three-month period. Within each location, testing continued until a pre-set applicant quota had been satisfied. The quotas were based on the applicant flow through the sites during a two-month period prior to testing. The six regions were selected to provide a representative and diverse sample of military applicants. Taken together, they were expected to provide nationally representative samples with respect to race, gender, and AFQT distributions. Data collection for the SED Study occurred from February 1988 to December 1988, and from September 1990 to April 1992 for the SEV study. (The beginning of the SEV Study in September 1990 was an especially noteworthy date since it marked the first operational use of CAT-ASVAB.)

In both studies (SED and SEV), each applicant was randomly assigned to one of three groups, and each group was assigned a different form of the ASVAB. Examinees in one group were given P&P-ASVAB (Form 15C), while examinees in the other two groups were given either Form 1 or Form 2 of the CAT-ASVAB (denoted as 01C and 02C, respectively). The random assignment involved a two-step process. First, the names of all examinees were entered into the random assignment and selection program. This automated program assigned, at random, two-thirds of the applicants to CAT-ASVAB and the remaining one-third to P&P-ASVAB (15C). The second step in the process involved randomly assigning each CAT-ASVAB examinee to an examinee testing station; each CAT station was randomly assigned either 01C or 02C, thus ensuring random assignment of examinees to CAT-ASVAB forms.

In the SED data collection, after taking either a non-operational CAT-ASVAB form or P&P-ASVAB 15C, each applicant was administered an operational P&P-ASVAB form. This operational form was used for enlistment and classification purposes. The non-operational forms were administered in the morning, and the operational forms were administered in the afternoon of the same day, following a break for lunch.

In the SEV study, all examinees were administered only one form of the ASVAB. All forms were administered under operational conditions, where the results (for both CAT-ASVAB and P&P-ASVAB) were used to compute operational scores of record. In the SEV study, the equating transformation used to compute operational scores of record for the CAT-ASVAB was obtained from the SED equating.

Data Editing and Group Equivalence

A small number of applicants were screened from the SED and SEV data sets using a procedure suggested by Hotelling (1931). This procedure identifies cases that are unlikely, given that the observations are sampled from a multivariate normal distribution. For the SED data, a 10 × 1 vector of difference scores was obtained between the operational and non-operational versions of the ASVAB taken by each examinee (each element of the vector corresponded to one of the 10 content areas). The inverse of the covariance matrix of difference scores was pre- and post-multiplied by the vector of centered difference scores to obtain an index for each examinee. Examinees with a large index value were those with an unlikely score pattern and were, therefore, excluded from the analysis. In a similar manner, the 10 × 1 vector of operational scores for the SEV data (obtained from either CAT-ASVAB or P&P-ASVAB) was used to calculate the covariance matrix, the inverse of which was pre- and post-multiplied by the vector of centered observed scores. Again, examinees with a large index value were those with an unlikely score pattern and were, therefore, excluded from the analysis.

In both data sets (SED and SEV) less than one percent of the sample was deleted. For the SED Study, the final sample sizes were 2,641 (01C); 2,678 (02C); 2,721 (15C). For the SEV Study, the final sample sizes were 3,446 (01C); 3,413 (02C); and 3,520 (15C). The SED sample contained about 18 percent females and 29 percent blacks, with corresponding percentages of 21 and 24 in the SEV sample.

The equating design relies heavily on the assumed equivalence among the three groups: (a) 01C, (b) 02C, and (c) 15C. Consequently, it is useful to examine the equivalence of these groups with respect to available demographic information. The numbers of females, blacks, and whites in each group are approximately equal. Two χ^2 analyses for assessing the equivalence of proportions across the three conditions were performed. The χ^2 significance tests for gender (SED: $\chi^2 = 2.95$, $df = 2$, $p = .23$; SEV: $\chi^2 = .20$, $df = 2$, $p = .90$) and race (SED: $\chi^2 = 2.98$, $df = 4$, $p = .56$; SEV: $\chi^2 = 7.57$, $df = 4$, $p = .11$) were non-significant, supporting the expectation of random equivalency across groups. In addition, the data collection and editing procedures resulted in groups of approximately equal sizes. For both the SED and SEV datasets, the χ^2 test of equivalent proportions of examinees across the three groups was non-significant (SED: $\chi^2 = 1.20$, $df = 2$, $p = .55$; SEV: $\chi^2 = 1.74$, $df = 2$, $p = .42$). These findings are consistent with expectations based on random assignment of applicants.

Smoothing and Equating

The objective of equipercentile equating is to provide a transformation of scale that will match the score distributions of the new and existing forms (Angoff, 1971). This transformation, which is applied to the CAT-ASVAB, allows scores on the two ASVAB versions to be used interchangeably, without disrupting applicant qualification rates.

One method for estimating this transformation involves the use of the two empirical cumulative distribution functions (CDFs). Scores on CAT-ASVAB and P&P-ASVAB could be equated by matching the empirical proportion scoring at or below observed score levels. However, this transformation is subject to random sampling errors contained in the CDFs. The precision of the equating transformation can be improved by smoothing either (a) the equating transformation, or (b) the two empirical distributions that form the equating transformation. For discrete number-right distributions, a number of methods and decision rules exist for specifying the type and amount of smoothing (e.g., Fairbank, 1987; Kolen, 1991).

The precision of any estimated equating transformation can be decomposed into a *bias* component and a v*ariance* component. Smoothing procedures that attempt to eliminate the bias will increase the random variance of the transformation. A high-order polynomial provides one example. The polynomial may track the data closely but may capitalize on chance errors and replicate poorly in a new sample. On the other hand, smoothing procedures that attempt to eliminate the random variance do so at the expense of introducing systematic error, or bias, into the transformation. Linear equating methods often replicate well but display marked departure from the empirical transformation. It should be noted that whatever equating method is used, the choice of method, either implicitly or explicitly, involves a trade-off between random and systematic error.

One primary objective of the CAT-ASVAB equating was to use smoothing procedures that provided an acceptable trade-off between random and systematic error. In this study, smoothing was performed on each distribution (CAT-ASVAB and P&P-ASVAB) separately. These smoothed distributions were used to specify the equipercentile transformation.

Two different smoothing procedures were used. One method, designed for continuous distributions (Kronmal & Tarter, 1968), was used to smooth CAT-ASVAB distributions. Another method, designed for discrete distributions (Segall, 1987), was used to smooth P&P-ASVAB distributions. These procedures are described below.

One additional concern arose over the shape of the equating transformation in the lower score range where data are sparse. Typically, most equating procedures provide a transformation that is either undefined or poorly defined over this lower range. This problem was overcome by fitting logistic tails to the lower portion of the smoothed density functions. These tails achieved two desirable results. First, the distributions were extended to encompass the lower range, thus defining the equating transformation over the entire score scale. Second, by pre-specifying the fit-point of the tail, the distribution (and consequently the equating transformation) above that point was left unaltered by the tail. Consequently, the tail-fitting procedure altered the equating only over a pre-specified lower range; the equating transformation above that range was unaltered. The details of the fitting procedures are described in conjunction with the density estimation procedures below.

Smoothing P&P-ASVAB Distributions

The procedure used to smooth the P&P-ASVAB, developed by Segall (1987), estimates the smoothest density that deviates from the observed density by a specified amount. Roughness is measured by

$$R = \sum_{j=1}^{n-2} \left[\hat{h}_j - 2\hat{h}_{j+1} + \hat{h}_{j+2} \right]^2 , \tag{9-1}$$

where $\hat{h}_j$ is the smoothed density estimate for the bin (or score level) *j*, and *n* is the number of bins. The index *R* can be viewed as a discrete analog to the squared integrated second derivative—an index which has wide application as a measure of roughness for continuous distributions.

The deviation of the estimated density from the empirical density can be measured by

$$X^2 = 2N \sum_{j=1}^{n} \dot{h}_j \ln\left(\dot{h}_j / \hat{h}_j \right) , \tag{9-2}$$

where $\dot{h}_j$ is the empirical sample proportion at score level *j*, and *N* is the sample size. The index X^2 is the likelihood ratio statistic and is asymptotically χ^2 distributed with $df = n - 1$. Notice that if the solution is constrained to have a small X^2, the estimated $\hat{h}_j$ and empirical $\dot{h}_j$ will deviate very little from one another, and the roughness index *R* is likely to be large. On the other hand, if the solution is allowed to have a large value of X^2, the resulting density is likely to have a small value of roughness *R* but possess a large deviation between the estimated $\hat{h}_j$ and the empirical $\dot{h}_j$. In effect, the constraint imposed on X^2 determines the trade-off between smoothness and the degree of deviation between the empirical and estimated densities.

The procedure used here placed the following constraint on X^2:

$$X^2 = df - 2 = n - 3 .$$

(9-3)

The rationale for this constraint can be obtained from the following considerations. Suppose that our smoothed $\hat{h}_j$ was the true density and the observed $\dot{h}_j$ was generated from observations that were sampled from this density. What value of X^2 would we be most likely to observe? The most likely value would be equal to the mode of the χ^2 distribution, which occurs at $n - 3$.

The density estimation procedure then minimizes roughness given by Equation 9-1, subject to the constraint that $X^2 = n - 3$. Several other constraints are imposed on the $\hat{h}_j$ to ensure that the solution defines a density: $\hat{h}_j > 0$ (for j = 1, 2,..., n), and $\sum_{j=1}^{n} \hat{h}_j = 1$. As a consequence of these constraints, the smoothed $\hat{h}_j$ deviates from the observed sample values by an amount to be expected by sampling error, and the resulting solution is the smoothest possible with this degree of deviation. The solution that satisfies the above constraints is obtained using an iterative numerical procedure that solves $n + 2$ simultaneous nonlinear equations.

The logistic CDF

$$F(x) = \frac{1}{1 + \exp\left[-\sigma(x - \mu)\right]}$$

(9-4)

was used to specify density values for the lower tail of the discrete distributions. The function closely approximates the normal CDF and is often used as a substitute since it provides mathematically tractable expressions for both the density and the distribution functions. Although the function is usually used to define a continuous CDF, it is used here to define a discrete density at bin *x* by

$$g\ (x) = F\left(x + \frac{1}{2}\right) - F\left(x - \frac{1}{2}\right) . \tag{9-5}$$

The first step in the tail-fitting process involved finding the largest *x*-value, x_r, from the smoothed solution that contained no more than five percent of the distribution. Once x_r was identified, two constraints were placed on the logistic function

$$g\ (x_r) = F\left(x_r + \frac{1}{2}\right) - F\left(x_r - \frac{1}{2}\right) = \hat{h}_r \tag{9-6}$$

and

$$\sum_{j=1}^{r} g\ (x_j) = F\left(x_r + \frac{1}{2}\right) - F\left(-\frac{1}{2}\right) = \sum_{j=1}^{r} \hat{h}_j \ . \tag{9-7}$$

The first constraint given by Equation 9-6 ensures that there is a smooth fit of the logistic tail to the estimated density defined by $\hat{h}_j$. This is accomplished by constraining the last bin of the tail $g(x_r)$ to equal the estimated value of the smoothed solution at that bin $\hat{h}_r$. The second constraint given by Equation 9-7 ensures that the proportion contained in the logistic tail will equal the proportion contained in the tail of the smoothed solution. It follows from this constraint that together, the logistic tail and the upper portion of the smoothed solution will define a density (i.e., sum to 1). Once the above constraints are imposed, values for μ and σ can be obtained through an iterative numerical procedure.

Smoothed distributions were estimated for each of the ten P&P-ASVAB tests. (Separate estimates were obtained for the SED and SEV data sets.) Figures 9-1 and 9-2 display the smoothed solutions and the fitted tails for two tests (General Science and Arithmetic Reasoning) of the P&P-ASVAB 15C estimated from SEV data. The empirical proportions for each bin are indicated by the height of the bar. The smoothed (or fitted) density values are indicated by the small bullets joined by the dotted lines. The point at which the tail was joined to the smoothed solution $\{x_r, g\ (x_r)\}$ is indicated by an arrow in each figure.

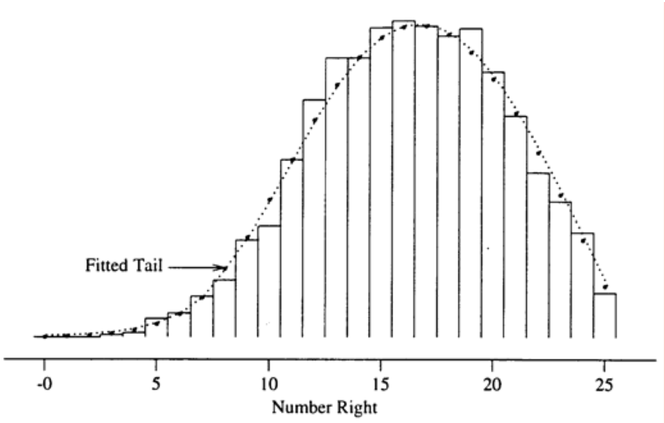


Figure 9-1. Smoothed and Empirical Density Functions for P&P-ASVAB 15C (General Science).

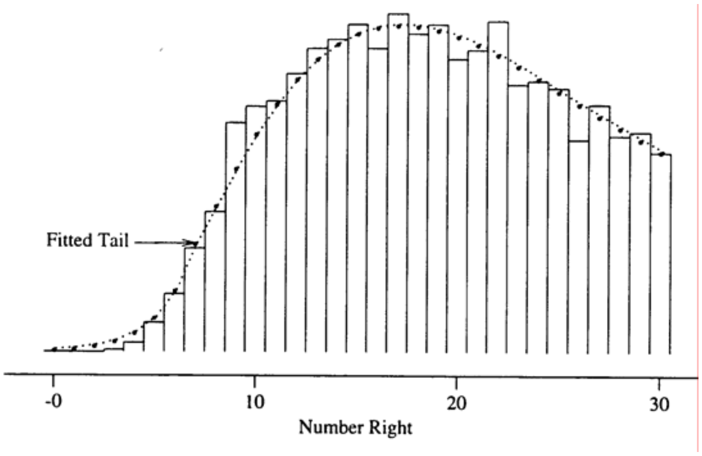


Figure 9-2. Smoothed and Empirical Density Functions for P&P-ASVAB 15C (Arithmetic Reasoning).

Smoothing CAT-ASVAB Distributions

The procedure developed by Kronmal and Tarter (1968) was used to smooth the CAT-ASVAB distributions. This procedure, which was designed for smoothing continuous distributions, provides a Fourier estimate of the density function using trigonometric functions. To obtain a useful density estimate, it is necessary to smooth the series by truncating it at some point. Kronmal and Tarter provide expressions that relate the mean integrated square error (MISE) of the Fourier estimator to the sample Fourier coefficients. The MISE expressions are used to specify a truncation point for the series, making it possible to specify an optimal number of terms in the series.

The distributions of penalized modal estimates (for seven adaptive power tests) and rate scores (for the two speeded tests) were smoothed using the Kronmal and Tarter (1968) method. Details about the item selection and scoring procedures are provided in Chapter 3 of this technical bulletin. Since the CAT-ASVAB measures Automotive Information (AI) and Shop Information (SI) separately, it was necessary to combine the two ability estimates into a single score; this composite measure must be formed because the P&P-ASVAB measures both content areas within a single test (AS). Smoothing was performed on the composite measure.

This composite measure was formed for each examinee using estimated AS parameters from P&P-ASVAB-9A. The AS items were divided into two sets based on their content: (a) auto-information (AI) items, and (b) shop-information (SI) items. AI items were calibrated among CAT-ASVAB AI items, and similarly, SI items were calibrated among CAT-ASVAB SI items (Prestwood, Vale, Massey, & Welsh, 1985). For each applicant, the expected number-right scores were obtained. In each case, the expected number-right scores were computed from the sum of item response functions evaluated at the examinee's estimated ability level. One expected number right score, τ_{AI} , was obtained from the AI-9A item parameters and the examinee's penalized ability estimate $\dot{\theta}_{AI}$. The other expected number-right score, τ_{SI} , was obtained from the SI-9A item parameters and the examinee's penalized ability estimate $\dot{\theta}_{SI}$. A composite measure was formed: $\tau_{AS} = \tau_{AI} + \tau_{SI}$. A smoothed density estimate of this composite measure was obtained in the subsequent equating analyses.

The logistic CDF given by Equation 9-4 was also used here to smooth the lower portion of the Fourier estimate where data are sparse. This tail fitting involved several steps. First, the proportion contained in the tail p_t was specified according to the proportion contained in the tail of the corresponding discrete (P&P-ASVAB) distribution given by Equation 9-6. That is, $p_t = \sum_{i=1}^{r} = \hat{h}_j$. Next, the value of x_c was specified using the inverse Fourier estimate. That is, x_c is the value below which p_t proportion of the distribution falls, according to the Fourier estimator. The values x_c and p_t were used to constrain the CDF, such that $F(x_c) = p_t$. This constraint imposed in this manner ensures the equivalence of the three proportions: (a) the proportion in the continuous logistic tail below x_c , (b) the proportion in the Fourier series tail below x_c , and (c) the proportion in the fitted discrete tail. A second constraint, $\partial F(x_c)/\partial x_c = d_c$ was added to ensure that the density value of the logistic tail at the join-point x_c equals the density of the Fourier estimate d_c at x_c. This constraint provided a continuous transition between the Fourier estimate and the logistic tail. Once the above constraints were imposed, values of μ and σ were obtained using an iterative numerical procedure.

Tail fitting posed a special problem for the CAT-ASVAB AS composite. The AS scores are on the τ metric, due to the transformation used to combine the AI and SI scores. This τ metric is bounded on the upper and lower ends over the interval $\left(\sum_{i=1}^{25} = c_i ,25\right)$. Consequently, scores below $\sum c_i$ are undefined. If the τ scores are smoothed directly, and a tail is fit to this smoothed distribution, much of the logistic tail falls below Σc_i, over a range that is undefined. This problem was circumvented by transforming the AS τ scores using the arcsin transform

$$w = \sin^{-1}\left[\frac{\tau - \sum_i c_i}{25 - \sum_i c_i}\right]^{\frac{1}{2}}, \tag{9-8}$$

and performing the smoothing and fitting to the w-values. This change of metric achieved two desirable results. First, the distribution of the transformed scores w appeared more “normal-like” than did the distribution of τ scores. Second, the transformation helps contain the logistic tail within the defined interval. This becomes evident after transforming the metric of the smoothed *w* distribution back to the original τ-metric using the inverse of Equation 9-8

$$\tau = \sin^2(w)(25 - \sum_i c_i) + \sum_i c_i .$$

(9-9)

Since 01C and 02C were smoothed separately, 20 density estimates were obtained for both the SED and SEV studies. Figures 9-3 and 9-4 display the smooth Fourier estimates and the fitted tails for 2 of the 10 tests of the CAT-ASVAB (01C), using data collected from the SEV study. In Figures 9-3 and 9-4, the empirical histograms for the CAT-ASVAB distributions are indicated by the height of the bar. The smoothed (or fitted) density functions are displayed by the dotted lines. The fitted logistic tail is displayed by the dotted curve to the left of the join-point (indicated by the solid bullet).

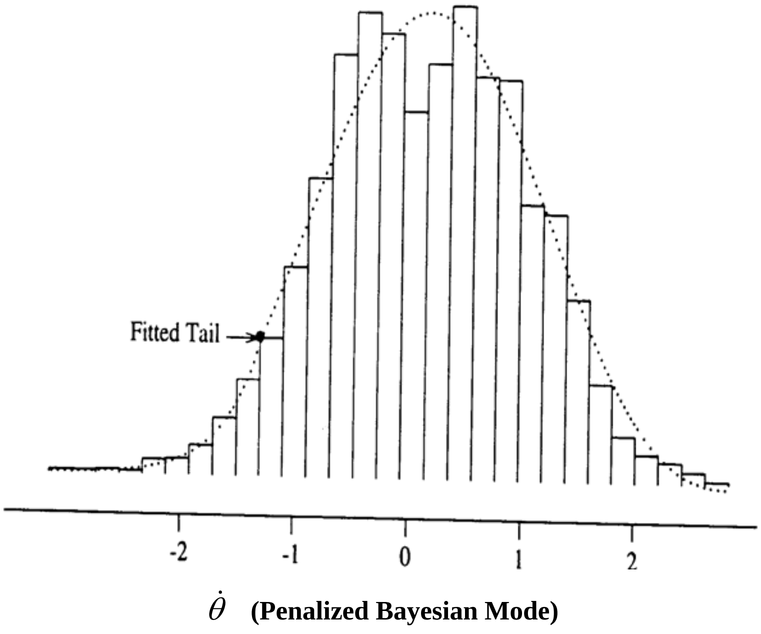


Figure 9-3. Smoothed and Empirical Density Estimates: CAT-ASVAB (Form 01), (General Science).

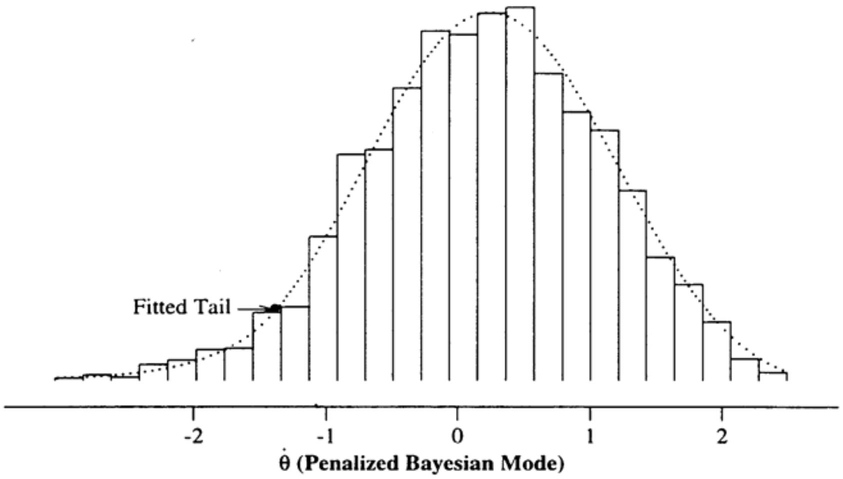


Figure 9-4. Smoothed and Empirical Density Estimates: CAT-ASVAB (Form 01), (Arithmetic Reasoning).

Equating Transformations

The smoothed distributions were used to specify the equipercentile transformation for the CAT-ASVAB. In each study (SED and SEV), there were a total of 20 equatings, one for each content area of each CAT-ASVAB form. For each P&P-ASVAB number-right score, an interval of the continuous CAT-ASVAB scores that contained the same estimated proportion was obtained. A sample conversion table for Paragraph Comprehension (PC), based on SEV data, is provided in Table 9-1. The column labeled $\hat{h}$ displays the smoothed 15C density estimate. The next two columns provide the CAT-ASVAB lower and upper limits (*LL*, *UL*) of score intervals which contain that proportion for the smoothed estimate based on 01C, and the last two columns contain the score interval for 02C.

Table 9-1. Paragraph Comprehension Conversion Table

		01C (Form 1)		**02C (Form 2)**	
		$LL \le \theta < UL$		$LL \le \theta < UL$	
Raw Score *X*	$\hat{h}$	***LL***	***UL***	***LL***	***UL***
0	0.0	–999.000	–3.484	–999.000	–3.497
1	0.1	–3.484	–2.923	–3.497	–2.976
2	0.2	–2.923	–2.483	–2.976	–2.566
3	0.4	–2.483	–2.081	–2.566	–2.192
4	0.9	–2.081	–1.695	–2.192	–1.833
5	1.9	–1.695	–1.316	–1.833	–1.481
6	2.3	–1.316	–1.072	–1.481	–1.207
7	3.2	–1.072	–0.877	–1.207	–0.931
8	5.1	–0.877	–0.667	–0.931	–0.673
9	7.3	–0.667	–0.438	–0.673	–0.449
10	10.0	–0.438	–0.164	–0.449	–0.218
11	13.2	–0.164	0.154	–0.218	0.061
12	16.2	0.154	0.483	0.061	0.447
13	17.0	0.483	0.839	0.447	0.908
14	14.2	0.839	1.321	0.908	1.374
15	8.0	1.321	999.000	1.374	999.000

Figures 9-5 and 9-6 compare the equating functions based on the smoothed densities with functions based on the empirical unsmoothed distributions for 2 of the 20 equatings obtained in the SEV study. The smoothed function is indicated by the bullets joined by solid lines. The dogleg portion of the function obtained from the tail-fitting procedure is indicated by a large bullet. The unsmoothed transformation is indicated by the dotted function. For both the smoothed and unsmoothed transformations, each number right (on the *y*-axis) is plotted against the midpoint of the CAT-ASVAB score interval (on the *x*-axis). The agreement between the smoothed and unsmoothed functions is very high above the dogleg portion. Notice that the tail appears to provide a smooth extrapolation of the equating function over the lower range and does not affect the agreement between the smoothed and empirical functions above the dogleg portion. Also notice that the dogleg provides a monotonic increasing function for mapping CAT-ASVAB scores into number-right scores.

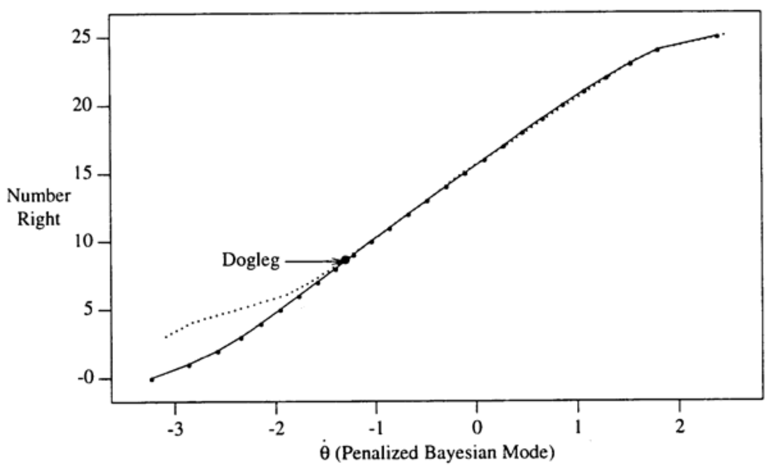


Figure 9-5. Smoothed and Empirical Equating Transformations for General Science (Form 01).

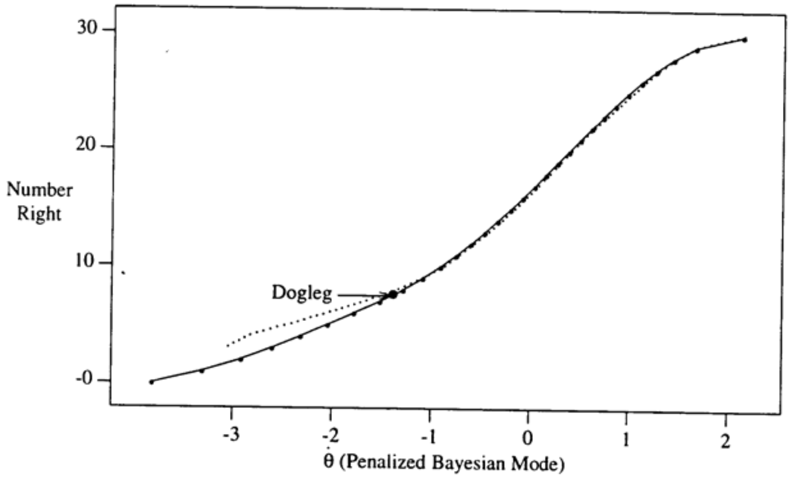


Figure 9-6. Smoothed and Empirical Equating Transformations for Arithmetic Reasoning (Form 01).

Composite Equating

Equating the CAT-ASVAB to the P&P-ASVAB involves matching test distributions using an equipercentile method. This distribution matching provides a transformation of the CAT-ASVAB ability estimates to number-right equivalents. Once this transformation is specified for each test, raw-score equivalents can be computed. These raw-score equivalents provide the basis for computing Service-specific selection composites, as well as the Armed Forces Qualification Test (AFQT) and Verbal (VE) composites.

One concern is that the distributions of CAT-ASVAB composites might differ systematically from P&P-ASVAB composite distributions. Such a difference could be caused by differences in test reliabilities. A more reliable CAT-ASVAB would have higher covariances among tests. Since the variance of a composite is partially affected by the covariance among tests, differences in composite variances could result as a consequence of differences in reliabilities. Higher order moments of the composite distributions could be affected in a similar manner. Thus, it is important to assess the need for equating CAT-ASVAB/P&P-ASVAB composites by examining the similarity of composite distributions.

Sample and Procedures

The sample consisted of 10,379 military applicants tested during the SEV data collection phase. The steps involved in computing composite score distributions differed among the three conditions (01C, 02C, and 15C) and are described below.

Each CAT-ASVAB content area was equated to the P&P-ASVAB using the procedures described in the preceding section. This equating was performed separately for each CAT-ASVAB form. First, CAT-ASVAB ability estimates were transformed to raw score equivalents using the smoothed equating transformations. Next, raw scores (from 15C) and raw score equivalents (from 01C and 02C) were transformed to standard scores using the standardization based on the 1980 reference population. (This standardization is derived from the means and variances of P&P-ASVAB Form 8A administered in the reference population.) Then, sums of test standard scores were computed for the 29 Service composites and for the AFQT. The Verbal (VE) composite was also computed from the sum of Word Knowledge (WK) and Paragraph Comprehension (PC) raw scores. A list of Service composites is provided in Table 9-2. After the sums were obtained, the appropriate scale conversion was applied to place each composite score on the metric used for classification decisions.

Each CAT-ASVAB composite distribution (for 01C and 02C) was compared to the corresponding 15C composite distribution. Two different methods were used to examine the significance of the differences. First, the Kolmogorov-Smirnov (K-S) two-sample test was used to detect overall differences between 01C and 15C, and between 02C and 15C. Since this test is not highly sensitive to differences of a specific nature (e.g., differences in variances), an *F*-ratio test was also used to test the differences between 01C and 15C variances and between 02C and 15C variances. Both significance tests were performed on all 31 composites.

Table 9-2. Significance Tests of Composite Standard Deviations

	Standard Deviation			F-ratio	
Composite	**01C**	**02C**	**15C**	**01C vs. 15C**	**02C vs. 15C**
Army					
GT = AR + VE	16.02	15.97	15.62	1.053	1.045
GM = GS + AS + MK + EI	16.07	15.72	16.38	1.039	1.086
EL = GS + AR + MK + EI	16.59	16.23	16.37	1.026	1.017
CL = AR + MK + VE	15.69	15.74	15.79	1.013	1.006
MM = NO + AS + MC + EI	15.88	15.71	15.97	1.012	1.034
SC = AR + AS + MC + VE	16.60	16.44	16.52	1.010	1.010
CO = AR + CS + AS + MC	16.29	16.02	16.32	1.003	1.037
FA = AR + CS + MK + MC	16.27	16.15	16.12	1.019	1.003
OF = NO + AS + MC + VE	14.97	14.89	15.24	1.036	1.048
ST = GS + MK + MC + VE	16.22	16.12	16.07	1.019	1.006
Navy					
EL = GS + AR + MK + EI	29.31	28.68	28.92	1.027	1.017
E = GS + AR + 2MK	30.15	30.32	30.38	1.016	1.004
CL = NO + CS + VE	17.97	17.94	17.90	1.008	1.004
GT = AR + VE	14.84	14.79	14.45	1.053	1.047
ME = AS + MC + VE	21.48	21.44	21.62	1.013	1.017
EG = AS + MK	12.75	12.89	13.89	1.187*	1.161*
CT = AR + NO + CS + VE	24.67	24.57	24.28	1.033	1.024
HM = GS + MK + VE	21.26	21.26	21.02	1.023	1.023
ST = AR + MC + VE	22.37	22.13	21.66	1.067	1.044
MR = AR + AS + MC	22.84	22.56	22.81	1.002	1.023
BC = CS + MK + VE	18.72	18.69	18.64	1.009	1.005
Air Force					
M = GS + 2AS + MC	25.61	25.22	26.08	1.037	1.069
A = NO + CS + VE	24.41	24.32	24.16	1.021	1.013
G = AR + VE	25.03	24.85	24.58	1.038	1.022
E = GS + AR + MK + EI	24.43	23.97	24.43	1.000	1.038
Marine Corps					
MM = AR + AS + MC + EI	17.30	17.02	17.06	1.028	1.005
CL = CS + MK + VE	14.64	14.62	14.59	1.007	1.005
GT = AR + MC + VE	16.91	16.72	16.37	1.067	1.043
EL = GS + AR + MK + EI	16.59	16.23	16.37	1.026	1.017
All Services					
AFQT = AR + MK + 2VE	23.78	23.79	23.87	1.008	1.006
VE = PC + WK	7.44	7.42	7.21	1.065	1.060

* p <.01

Note: See key of abbreviations in Exhibit 9-1.

Exhibit 9-1: Key Service and DoD composite and Test Abbreviations in Table 9-2

Service Composites				DoD	ASVAB Tests
Army	**Navy**	**Air Force**	**Marine Corps**		
GT = General Technical	EL = Electronics	M = Mechanical	MM = Mechanical Maintenance	AFQT = Armed Forces Qualification Test	AR = Arithmetic Reasoning
GM = General Maintenance	E = Basic Electricity and Electronics	A = Administrative	CL = Clerical		AS = Auto and Shop Information
EL = Electronics	CL = Clerical	G = General	GT = General Technical		CS = Coding Speed
CL = Clerical	GT = General Technical	E = Electronics	EL = Electronics Repair		EI = Electronics Information
MM = Mechanical Maintenance	ME = Mechanical				GS = General Science
SC = Surveillance / Communications	EG = Engineering				MC = Mechanical Comprehension
CO = Combat	CT = Cryptologic Technician				MK = Mathematics Knowledge
FA = Field Artillery	HM = Hospitalman				NO = Numerical Operations
OF = Operations/ Food	ST = Sonar Technician				PC = Paragraph Comprehension
ST = Skilled Technical	MR = Machinery Repairman				WK = Word Knowledge
	BC = Business and Clerical				

Results and Discussion

Of the 62 comparisons examined using the K-S tests, only one was significant at the .01 level. This comparison was between CAT-Form 2 and 15C for the Navy EG composite. Two of the 62 variance comparisons (Table 9-2) were significant at the .01 level. These significant variance differences existed between both CAT-ASVAB forms and 15C for the Navy EG composite.

The results of the K-S and *F*-ratio tests are generally indicative of no differences between CAT-ASVAB and P&P-ASVAB composite score distributions, with the possible exception of the Navy EG composite. It is possible that the significant differences were due to Type I errors that occur when a large number of comparisons are made. In this study, over 124 comparisons were made. Finding at least three significant differences (at the .01 level) is highly probable, even when no true differences exist between the composite distributions.

However, this same Navy composite exhibited significant variance differences (between CAT-ASVAB and P&P-ASVAB) in the SED analysis (Segall, 1989). That is, the results found here were consistent with those found in the SED study. Therefore, it is unlikely that both sets of significant differences were due to Type I errors. Consequently, it is prudent to examine the consequence of not equating this composite, under the assumption that the observed differences are not subject to sampling errors. That is, suppose the observed differences in composite distributions were treated as true differences; what consequence would this difference have on flow rates?

The Navy training schools that select on EG all employ a cut-score of 96. An analysis of the proportion of applicants scoring at or above 96 on each of the CAT-ASVAB forms, and 15C shows that $P(X \geq 96 \mid 01C) = .704$, $P(X \geq 96 \mid 02C) = .709$, and $P(X \geq 96|15C) = .668$. Consequently, if the observed sample differences were treated as true differences, then about four percent additional CAT-ASVAB applicants would qualify for schools using the Navy EG composite. This difference is relatively small.

Subgroup Comparisons

Although equipercentile equating matches CAT-ASVAB and P&P-ASVAB distributions for the total applicant sample, it does not necessarily guarantee a match for distributions of subgroups contained in the sample. This result follows from the fact that the two versions (CAT-ASVAB and P&P-ASVAB) are not parallel. Although we might expect small differences in subgroup performance across the two versions as a result of differences in measurement precision, a multitude of other factors could also cause group differences. It is therefore instructive to examine the performance of subgroups to determine whether any are placed at a substantial disadvantage by CAT-ASVAB. Two subgroups were examined in this analysis: (a) females and (b) blacks.

Test Comparisons

The equating transformation based on the total edited sample (*N* = 10,379) was applied to members of the two subgroups who had taken CAT-ASVAB. For each subgroup, the subgroup's performance on CAT-ASVAB was compared with its performance on the P&P-ASVAB. All ten content areas were examined, as well as the VE and AFQT composites. For each test and composite, three statistics for assessing distributional differences were computed. The K-S test was used to identify overall differences; the *F*-ratio statistic was used to identify differences in variances; and the *t*-test was used to test mean differences. In instances where overall differences are found, the *t*-test can be used to identify which version (CAT-ASVAB or P&P-ASVAB) provides an advantage, on the average, to members of the specified subgroup.

Tables 9-3 and 9-4 provide the results of the significance tests for females and for blacks, respectively. Among the comparisons for females, two tests (PC and AS) displayed significant differences at the .01 alpha level. For both tests, P&P-ASVAB applicants possessed an advantage. Among the comparisons for blacks, two tests

(AS and MK) displayed significant differences. For both tests, CAT-ASVAB applicants displayed a slight advantage.

Only 2 of 24 female and black comparisons showed a significant disadvantage for CAT-ASVAB. Both involved female comparisons. One difference was for PC, and represented about one standard score unit, or about 1/10 of a standard deviation. Since PC is never used in a composite without WK, comparisons involving the VE composite are more relevant than PC alone. The VE composite comparisons were non-significant for females. The other difference was for AS and is discussed below.

Table 9-3. Female Differences Between P&P-ASVAB and CAT-ASVAB Versions in the SEV Study

	K-S		***F* Ratio**		***t* test**				
Test	***Z* Value**	***p***	***F* Value**	***p***	***t***	***p***	$\bar{X}_{cat}$	$\bar{X}_{pp}$	**Advantage**
GS	.426	.993	1.10	.178	.11	.912	48.02	47.98	None
AR	.660	.777	1.03	.662	–1.15	.252	48.56	49.03	None
WK	.502	.963	1.03	.634	.39	.699	51.08	50.95	None
PC	1.776	.004	1.03	.720	–2.82	.005	51.37	52.35	P&P-ASVAB
NO	1.223	.100	1.00	.993	–2.22	.026	54.61	55.34	None
CS	1.082	.192	1.03	.706	–1.98	.047	55.71	56.44	None
AS	3.075	.001*	1.27	.001*	–7.23	.001*	42.05	44.37	P&P-ASVAB
MK	.724	.671	1.00	.958	.58	.560	52.29	52.05	None
MC	.718	.680	1.11	.124	–1.48	.140	45.34	45.89	None
EI	.967	.307	1.01	.832	–1.20	.231	44.75	45.19	None
VE	.548	.925	1.04	.611	–.56	.573	51.21	51.40	None
AFQT	.777	.582	1.05	.511	–.58	.563	50.99	51.62	None

* *p < .001* CAT-ASVAB: *N* = 1,184; P&P-ASVAB: *N* = 620

Table 9-4. Black Differences Between P&P-ASVAB and CAT-ASVAB Versions in the SEV Study

	K-S		*F* Ratio		*t* test				
Test	***Z* Value**	***p***	***F* Value**	***p***	***t***	***p***	$\bar{X}_{cat}$	$\bar{X}_{pp}$	**Advantage**
GS	.790	.561	1.02	.769	–.88	.381	44.78	45.07	None
AR	.364	.999	1.00	.988	–.53	.599	45.22	45.38	None
WK	.762	.607	1.10	.114	–.16	.871	46.76	46.81	None
PC	.778	.580	1.08	.176	–1.05	.292	47.20	47.56	None
NO	.595	.870	1.07	.252	1.24	.217	52.21	51.79	None
CS	.671	.759	1.02	.719	.76	.450	51.30	51.05	None
AS	1.704	.006	1.22	.001	2.43	.015	45.00	44.27	CAT-ASVAB
MK	1.504	.022	1.08	.184	3.00	.003	49.71	48.69	CAT-ASVAB
MC	1.137	.151	1.03	.578	1.23	.217	44.98	44.59	None
EI	.973	.300	1.23	.001	1.36	.174	44.76	44.31	None
VE	.732	.657	1.05	.385	–.54	.590	46.78	46.95	None
AFQT	.834	.490	1.11	.081	.25	.803	38.73	38.52	None

CAT-ASVAB: *N* = 1,649; P&P-ASVAB: *N* = 830.

Supplemental Auto/Shop Analyses

Among the subgroup differences, those found for females on AS are especially noteworthy. Females traditionally score lower than males on AS, resulting in fewer opportunities for women in jobs requiring this knowledge. Lower scores for women on CAT-ASVAB AS have the potential for reducing still further the number of women qualifying for these traditionally male jobs. Although two differences were identified for black applicants across CAT and P&P versions, these differences are potentially beneficial to black applicants taking a CAT. Black applicants taking CAT-ASVAB are likely to have higher qualification rates than blacks taking P&P-ASVAB (although this increase may be small).

Similar female difference on AS were obtained in the SED study (Segall, 1989), with females scoring about 2.7 standard score points higher on AS-P&P than on AS-CAT. Because of these noteworthy female differences on AS, supplemental analyses were performed on data collected during the SED study to investigate potential causes. The plausibility of four different causal factors were examined: group equivalence, precision, dimensionality, and dimensionality/precision interaction.

Group Equivalence. The group equivalence hypothesis asserts that (a) females taking CAT-ASVAB were less able on AS than females taking P&P-ASVAB, and (b) this difference contributed to the observed difference between CAT-ASVAB and P&P-ASVAB scores. Although applicants were randomly assigned to CAT and P&P versions, random assignment does not ensure equivalent groups; highly significant differences can occur by chance.

To test this hypothesis, an analysis of covariance was performed using data from the SED study. The dependent variable was the non-operational score on AS; the independent variable was version (either CAT or P&P); the covariate was the operational AS score. The results are summarized in Table 9-5.

Although females taking CAT-ASVAB scored lower (on their operational AS test) than females taking P&P-ASVAB, this difference is very small and does not account for the relatively large difference in non-operational means on AS. This is apparent from the adjusted means presented in Table 9-5. It is unlikely that the difference in AS means was caused by unequal groups, especially since the finding was replicated in the SEV study.

Table 9-5. Analysis of Covariance of Female Differences on the Auto/Shop Test (SED Study)

		Operational	Non-operational	
Group	***N***	$\overline{X}$	**Un-adjusted** $\overline{X}$	**Adjusted** $\overline{X}$
CAT-ASVAB	873	10.75	$9.64^{(c)}$	$9.66^{(c)}$
P&P-ASVAB	478	10.86	$11.20^{(p)}$	$11.15^{(p)}$

Note. *c* = Non-operational CAT-ASVAB; *p* = Non-operational P&P-ASVAB

Precision. This hypothesis states that the increased precision of CAT-ASVAB will magnify the difference between high and low scoring subgroups in comparison to P&P-ASVAB. The direction of the female performance on CAT-ASVAB AS was consistent with the precision hypothesis. However, the hypothesis does not correctly predict the direction of the difference for black applicants on AS; black applicants as a group scored lower on AS than white applicants. In accordance with the precision hypothesis, we would expect blacks to score significantly lower on CAT than on P&P, but just the reverse was true: blacks scored significantly higher on AS-CAT than on AS-P&P. Although precision most likely contributes to the female differences, some additional factor must be invoked to account for black performance.

Dimensionality. This hypotheses asserts that the difference in female Auto/Shop performance between CAT-ASVAB and P&P-ASVAB is caused by a difference in the test's verbal loading. This hypothesis is based on the following suppositions. First, AS-CAT has a lower verbal loading than AS-P&P (15C). Second, males and females have a large difference in mean AS knowledge, with males scoring higher. Third, males and females differ less in their verbal abilities than in their AS knowledge. If test performance is a composite of verbal and AS dimensions, then the test that gives the lowest relative weight to the verbal dimension will provide the lowest mean test performance for females. (In reasoning through this argument, it is helpful to remember that the equating forces the means and variances on the combined "male + female" group to be equivalent across the CAT and P&P versions.)

To investigate this hypothesis, the relationship between the test's reading grade level (RGL) and mean female performance was examined. Here we are assuming that the RGL for an AS test is an indicator of the magnitude of its verbal loading. In addition to the P&P reference form (15C), three other P&P-ASVAB forms were included in this analysis: 15, 16, and 17. After these forms were equated on the combined male + female sample, significant differences in mean female performance were identified (Monzon, Shamieh, & Segall, 1990). For each of the four P&P-ASVAB forms, the Flesch index was calculated, and mean female performance was computed from a sample of applicants tested during the IOT&E of these forms (Table 9-6).

For the CAT-ASVAB, a complication arises when computing the RGL of an applicant's test. Because of the adaptive nature of the test, different applicants receive different questions, and, consequently, some degree of variation in RGL is likely among applicants taking CAT-ASVAB. Furthermore, the RGL of individual items may be correlated with item difficulty, causing low-ability examinees to receive a lower "RGL" test than high-ability examinees. To address this issue, a separate RGL index was computed for female CAT-ASVAB examinees in the SED study. The exact item text was reconstructed from the examinee protocol, and then the RGL was computed from this item text. These two steps were repeated for examinees in the female sample, and an average RGL was calculated across the 407 female examinees. RGL and mean CAT-ASVAB AS performance are shown in Table 9-6.

Table 9-6. Reading Grade Level Analysis of ASVAB Versions of the Auto/Shop Test

ASVAB Version	Reading Grade Level (RGL)	Auto/Shop Mean (Females)
CAT	7.1	41.54
P&P-16	7.5	42.17
P&P-17	7.6	42.81
P&P-15	7.9	42.57
P&P-8A	8.5	43.36

There is a nearly perfect rank ordering between mean female performance and RGL. These results are consistent with the hypothesis that the difference in female AS performance between CAT-ASVAB and P&P-ASVAB is (at least partially) due to differences in their verbal loadings.

Dimensionality/Precision Interaction. Although the RGL analysis supports the role of dimensionality in explaining differences in female performance across CAT and P&P versions, several questions remain. First, does dimensionality account for the entire difference in female AS means across CAT- and P&P-ASVAB? Second, what role does precision play in accounting for female differences? Third, does dimensionality also account for the difference in the performance of blacks across CAT- and P&P-ASVAB?

To address these issues, a confirmatory factor analysis was performed using data collected in the SED study. This analysis modeled observed means as well as observed covariances among selected tests. The objective was to describe the differences in subgroup performance on AS as a function of (a) the Verbal and AS loadings, (b) precision, and (c) the mean latent ability of each subgroup. For this analysis, eight subgroups were defined by crossing ASVAB version with gender and race (Table 9-7).

Table 9-7. Subgroup Sample Sizes for Structural Equations Model

Group	Version	Gender	Race	N
1	P&P	M	White	1,521
2	P&P	M	Black	534
3	P&P	F	White	311
4	P&P	F	Black	179
5	CAT	M	White	2,981
6	CAT	M	Black	1,128
7	CAT	F	White	546
8	CAT	F	Black	345

The observed means and covariances for two tests were included in the analysis: Auto/Shop (AS) and Paragraph Comprehension (PC). The structural relations between *x* (the observed number-right score) and two latent variables τ_{re} (latent reading proficiency) and τ_{as} (latent AS knowledge) are given by the equations

P&P-ASVAB:

$$x_{pc} = \nu_1 + \lambda_1 \tau_{re} + \delta_1, \tag{9-10}$$

$$x_{as} = \nu_2 + \lambda_2 \tau_{re} + \lambda_3 \tau_{as} + \delta_2, \tag{9-11}$$

CAT-ASVAB:

$$x_{pc} = \nu_3 + \lambda_4 \tau_{re} + \delta_3, \tag{9-12}$$

$$x_{as} = \nu_4 + \lambda_5 \tau_{re} + \lambda_6 \tau_{as} + \delta_4 \ . \tag{9-13}$$

Note that the slopes λ's and intercepts ν's are allowed to vary across CAT and P&P versions for corresponding tests. The covariance matrix of measurement errors for P&P is parameterized by a 2 × 2 matrix $\Theta_1 = \mathrm{E}(\delta\delta')$, where $\delta' = [\delta_1, \delta_2]$. Similarly for CAT, the variance-covariance matrix of measurement errors is denoted by $\Theta_2 = \mathrm{E}(\delta\delta')$, where $\delta' = [\delta_3, \delta_4]$. Table 9-8 provides additional model parameters which include the latent means and covariances among the reading and AS dimensions for each of the four groups defined by race and gender.

Table 9-8. Structural Model Parameter Definitions

	Means		
Group	$\mathrm{E}(\xi_{\mathrm{re}})$	$\mathrm{E}(\xi_{\mathrm{as}})$	$\mathrm{Cov}(\xi_i, \xi_j)$
White Male	κ_1	κ_2	$\mathbf{\Phi}_1$
Black Male	κ_3	κ_4	$\mathbf{\Phi}_2$
White Female	κ_5	κ_6	$\mathbf{\Phi}_3$
Black Female	κ_7	κ_8	$\mathbf{\Phi}_4$

Particular constraints were placed on model parameters across the eight groups defined by version, race, and gender. First, the slopes λ's and intercepts ν's depend only on version and are not influenced by subgroup. Second, means κ's and covariances $\mathbf{\Phi}$'s of the latent variables vary only according to subgroup (defined by race and gender) and are not dependent on version. Finally, variances of measurement errors $\mathbf{\Theta}$ depend only on version and are not dependent on subgroup. These constraints can be summarized by the following equations:

P&P-ASVAB:

White Males: $$\boldsymbol{\Omega}_1 = f\left(\nu_1, \nu_2, \lambda_1, \lambda_2, \lambda_3, \Theta_1; \kappa_1, \kappa_2, \Phi_1\right) \tag{9-14}$$

Black Males: $$\boldsymbol{\Omega}_2 = f\left(\nu_1, \nu_2, \lambda_1, \lambda_2, \lambda_3, \Theta_1; \kappa_3, \kappa_4, \Phi_2\right) \tag{9-15}$$

White Females: $$\boldsymbol{\Omega}_3 = f\left(\nu_1, \nu_2, \lambda_1, \lambda_2, \lambda_3, \Theta_1; \kappa_5, \kappa_6, \Phi_3\right) \tag{9-16}$$

Black Females: $$\boldsymbol{\Omega}_4 = f\left(\nu_1, \nu_2, \lambda_1, \lambda_2, \lambda_3, \Theta_1; \kappa_7, \kappa_8, \Phi_4\right) \tag{9-17}$$

CAT-ASVAB:

White Males: $$\boldsymbol{\Omega}_5 = f\left(\nu_3, \nu_4, \lambda_4, \lambda_5, \lambda_6, \Theta_2; \kappa_1, \kappa_2, \Phi_1\right) \tag{9-18}$$

Black Males: $$\boldsymbol{\Omega}_6 = f\left(\nu_3, \nu_4, \lambda_4, \lambda_5, \lambda_6, \Theta_2; \kappa_3, \kappa_4, \Phi_2\right) \tag{9-19}$$

White Females: $$\boldsymbol{\Omega}_7 = f\left(\nu_3, \nu_4, \lambda_4, \lambda_5, \lambda_6, \Theta_2; \kappa_5, \kappa_6, \Phi_3\right) \tag{9-20}$$

Black Females: $$\boldsymbol{\Omega}_8 = f\left(\nu_3, \nu_4, \lambda_4, \lambda_5, \lambda_6, \Theta_2; \kappa_7, \kappa_8, \Phi_4\right) \tag{9-21}$$

where $\boldsymbol{\Omega}_k$ is the model implied moment matrix for group *k*. The parameters contained in the function $f\left(\;\right)$ illustrate the dependence of each of the eight moment matrices on the model parameters defined above.

Maximum likelihood estimates of the model parameters were obtained using LISREL VI (Joreskog & Sorbom, 1984). To identify the model, several additional constraints were necessary. These constraints fixed the origin and unit for the two latent variables.

First $\kappa_1 = \kappa_2 = 0$ (latent means for white males). Second, the diagonal elements of $\mathbf{\Phi}_1$ were set equal to 1 (i.e., the latent variances for white males were fixed at one). And third, the variances of measurement errors were fixed at values calculated from the alternate forms reliability study (Chapter 7 in this Technical Bulletin):

$$\mathbf{\Theta}_1 = \begin{bmatrix} 3.686 & 0 \\ 0 & 5.372 \end{bmatrix}, \tag{9-22}$$

and

$$\mathbf{\Theta}_2 = \begin{bmatrix} 3.904 & 0 \\ 0 & 2.396 \end{bmatrix}. \tag{9-23}$$

The overall fit of the model implied moment matrices to the observed moment matrices is provided by two fit statistics: χ^2 = 47.07, (*df* = 14), and *GFI* = .996. In general, these values indicate a relatively good fit. Parameter estimates for each equation are

P&P Estimates:

$$x_{\text{pc}} = 11.673 + 1.885\xi_{\text{re}} + \delta_1 \tag{9-24}$$

$$x_{\text{as}} = 16.512 + 4.547\ \xi_{\text{re}} + 3.197\ \xi_{\text{as}} + \delta_2 \tag{9-25}$$

CAT Estimates:

$$x_{\text{pc}} = 11.658 + 1.847\ \xi_{\text{re}} + \delta_3 \tag{9-26}$$

$$x_{\text{as}} = 16.734 + 4.378\ \xi_{\text{re}} + 4.170\ \xi_{\text{as}} + \delta_4\ . \tag{9-27}$$

Notice that, as predicted, the loading for x_{as} on the reading dimension is higher for P&P than for CAT (4.547 vs. 4.378). Also notice that x_{as} has a different loading on the latent AS dimension across CAT and P&P versions, 4.170 (for CAT) vs. 3.197 (for P&P). This last result is most likely due to CAT's greater precision. The estimated latent means κ's for each subgroup on each dimension are provided in Table 9-9. The estimated means κ's, slopes λ's, and intercepts ν's can be used to specify model-implied means for the observed indicator variable x_{as}. For each subgroup, two means can be computed, one for CAT and another for P&P:

P&P-ASVAB

$$\mu_{as}^{k} = \nu_2 + \lambda_2 \kappa_{re}^{k} + \lambda_3 \kappa_{as}^{k}, \tag{9-28}$$

CAT-ASVAB

$$\mu_{as}^{k} = \nu_4 + \lambda_5 \kappa_{re}^{k} + \lambda_6 \kappa_{as}^{k}, \tag{9-29}$$

(for $k \in$ {WM, BM, WF, BF}). A comparison of the model implied means with the observed means across subgroups and versions provides an indication of how well the model predicts differential subgroup performance. Substituting the estimated parameters into the above equations provides us with the results displayed in Table 9-10. The last column lists the difference between the observed and model-implied means shown in the first two columns. The observed differences in subgroup performance can be accurately described by the structural model. That is, differences in mean performance across CAT and P&P versions are consistent with the model predictions which describe a subgroup's performance as a function of (a) the Verbal and AS loadings, (b) precision, and (c) the mean subgroup latent ability.

Table 9-9. Latent Subgroup Means

	Means (κ)	
Subgroup	$E(\xi_{re})$	$E(\xi_{as})$
White Males	(0)	(0)
Black Males	–1.106	.104
White Females	.137	–1.558
Black Females	–.691	–1.392

Note. ( ) indicates fixed value

Table 9-10. Observed and Implied Auto/Shop Means

Subgroup	Observed	Implied	Diff.
P&P-ASVAB			
White Males	16.660	16.512	.148
Black Males	11.307	11.816	–.509
White Females	12.334	12.150	.184
Black Females	9.016	8.920	.096
CAT-ASVAB			
White Males	16.667	16.734	–.067
Black Males	12.516	12.326	.190
White Females	10.752	10.834	–.082
Black Females	7.864	7.907	–.043

Impact Assessment. According to the Dimensionality/Precision Model, AS-CAT provides a measure of AS knowledge that is slightly less contaminated by reading proficiency than AS-P&P. From the standpoint of increased classification efficiency and possibly validity, this makes the use of CAT-ASVAB more desirable. However, one of the goals of the equating was to achieve, to the extent possible, an equating that places no subgroup at a substantial disadvantage. Since during an extended implementation phase, both CAT-ASVAB and P&P-ASVAB will be administered operationally, it is desirable for applicants of various subgroups to be indifferent about which of the two versions they receive. If women score lower on the average on AS-CAT, then they might prefer the P&P-ASVAB.

The general question of impact arose during a consideration of the SEV phase, in which a planned sample of 7,500 applicants was to take an operational version of CAT or P&P. Data for addressing the impact on Navy school-qualification rates were available. The specific question was: Among the 7,500 military applicants to be tested during SEV, how many female Navy recruits would be expected to fail their assigned rating entry requirements as a consequence of lower AS performance on CAT-ASVAB?

Data addressing this question came from three sources. The first source was data collected during the SED equating study. From this sample of about 8,000 applicants, a series of conditional probabilities were computed. The series produced the top portion of the probability tree displayed in Figure 9-7. Examinees in each box in the left column were repeatedly divided into exclusive non-overlapping subgroups. First, the applicant group [Box 0] was divided into those taking CAT [Box 2] and those taking P&P [Box 1]. The applicants taking CAT [Box 2] were divided into Navy applicants [Box 4] and non-Navy applicants [Box 3]. The Navy applicants [Box 4] were divided in female applicants [Box 6] and male applicants [Box 5]. The numbers in each successive group were tallied and used to compute the conditional probabilities reported in Figure 9-7.

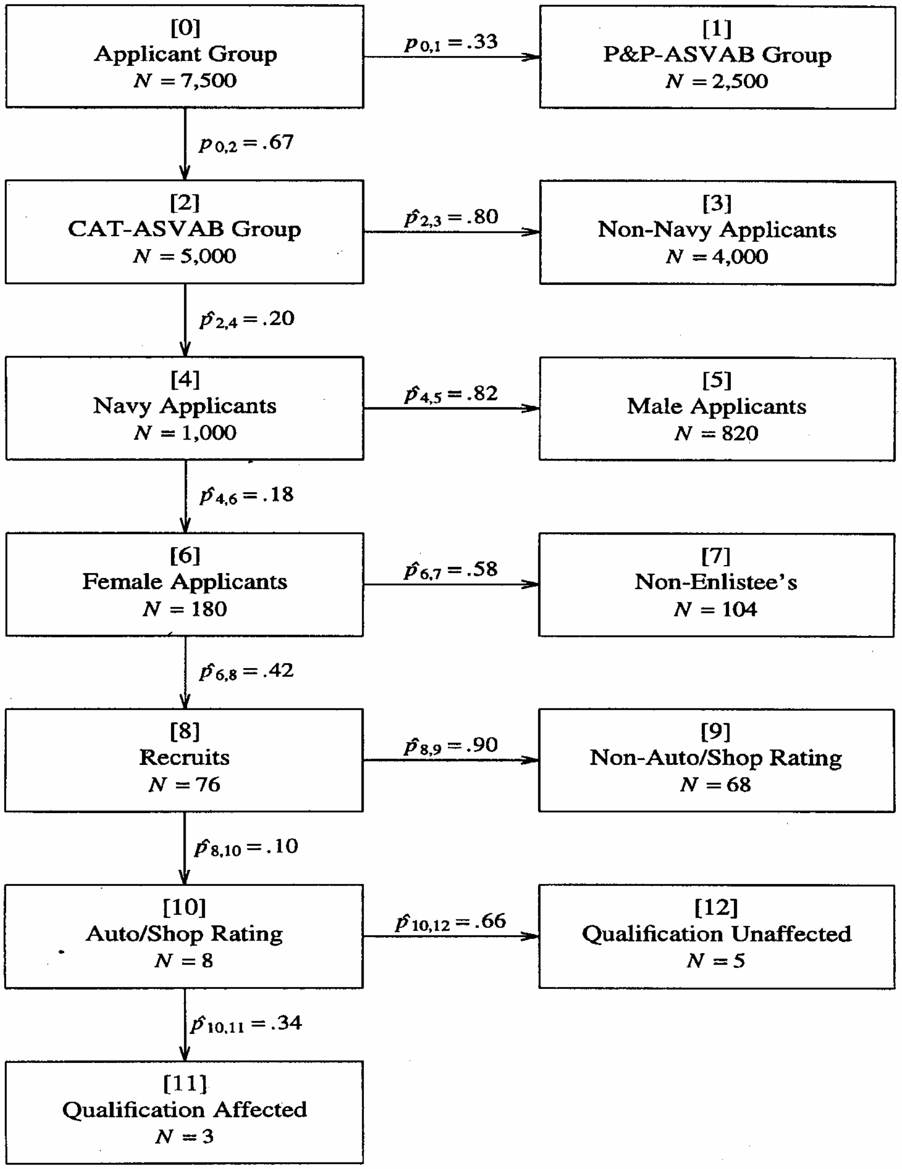


Figure 9-7. Estimated Auto/Shop Effect.

A second sample of about 27,500 examinees was used to determine the probability of a female Navy applicant becoming a female Navy recruit. These data were obtained from the Defense Manpower Data Center using accession data from FY89. As indicated in Figure 9-7, female Navy applicants [Box 6] were divided into recruits [Box 8] and non-enlistees [Box 7], and the resulting frequencies were used to compute the conditional probabilities.

Finally, a third sample of about 10,500 was used to determine the remaining probabilities in Figure 9-7. This sample was obtained from PRIDE (a Navy Recruiting Database) and was based on recruits accessed from June 1989 through May 1990. Female Navy recruits in Box 8 were divided into those who entered a job that used AS in its selector composite [Box 10] and those entering a job that used a selector composite not containing AS [Box 9]. Using the same sample of 10,500, the recruits in Box 10 were divided into two groups on the basis of qualification status change. For each female recruit in Box 10, three standard score points were subtracted from her composite score. This decrement was based on the mean difference between female performance on CAT-ASVAB and P&P-ASVAB in the SED study—about 2.7 standard score points. The reduced composite score was then compared to the cut-score used for the school she had entered. The number of women having their qualification status changed from qualified (before the decrement) to unqualified (after the decrement) was tallied and included in Box 11. The women not having their qualification status altered by the decrement were included in Box 12.

The conditional probabilities obtained from the these frequencies were used to estimate the effect of lower AS-CAT scores for women on their qualification status: among the 7,500 military applicants to be tested during SEV, three female Navy recruits would be expected to fail their assigned rating entry requirements as a consequence of lower AS performance on CAT-ASVAB. This analysis suggests that the impact on qualification rates is very small, both for SEV and for an extended OT&E of CAT-ASVAB.

Summary

The present study addresses three major concerns about equating CAT-ASVAB and P&P-ASVAB versions. First the use of an equipercentile procedure ensures that the transformation applied to CAT-ASVAB scores preserves flow rates into the military and into various occupational specialties. Smoothing procedures were used to increase the precision of the transformation estimates. Although equating was performed at the test level, the equivalence of CAT-ASVAB and P&P-ASVAB composite distributions was verified to ensure that the use of CAT-ASVAB would not disrupt flow rates dependent on the equivalence of these composite distributions.

Second, the equating study was conducted in two phases to ensure that the transformation was based on operationally motivated applicants. The first phase, SED, was used to obtain a preliminary equating based on data collected under non-operationally motivated conditions. The second phase, SEV, was used to obtain an equating transformation based on operationally motivated examinees (whose CAT-ASVAB scores were transformed to the P&P metric using the provisional SED equating). This latter equating was used in the OT&E phase to collect data on concepts of operation (Chapter 10 in this technical bulletin).

The third issue examined by the equating study addressed the concern that subgroup members taking CAT-ASVAB should not be placed at a disadvantage relative to their subgroup counterparts taking the P&P-ASVAB. Results indicate that although it is desirable for exchangeability considerations 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-ASVAB and P&P-ASVAB 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. Although some subgroup differences observed in CAT-ASVAB are statistically significant, their practical significance on qualification rates is small. Once CAT-ASVAB fully replaces the P&P-ASVAB, the exchangeability issue will become less important. The small differences in subgroup performance displayed by CAT-ASVAB may be a positive consequence of greater precision and lower verbal contamination. Ultimately, in large-scale administrations of CAT-ASVAB, we may observe higher classification efficiency and greater predictive validity than is currently displayed by its P&P counterpart.

References

Angoff, W. H. (1971). Norms, scales, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.).* Washington, DC: American Council on Education.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11,* 245-262.

Hotelling, H. (1931). A generalization of student's ratio. *Annual of Mathematical Statistics, 2,* 360-378.

Jöreskog, K. G., & Sorbom, D. (1984). *LISREL VI, Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods.* Mooresville, IN: Scientific Software.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28,* 257-282.

Kronmal, R., & Tarter, M. (1968). The estimation of probability density and cumulatives by Fourier series methods. *Journal of the American Statistical Association, 69,* 925-952.

Monzon, R. I., Shamieh, E. W., & Segall, D. O. (1990). *Subgroup differences in equipercentile equating of the Armed Services Vocational Aptitude Battery: Development of an adaptive item pool*. Unpublished manuscript. San Diego, CA: Navy Personnel Research and Development Center.

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). Brooks AFB, TX: Air Force Human Resources Laboratory. (AD-A160 608)

Segall, D. O. (1987). *A procedure for smoothing discrete distributions*. Unpublished manuscript. San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1989). *Score equating development analyses of the CAT-ASVAB.* Unpublished manuscript, Navy Personnel Research and Development Center.

Chapter 10

CAT-ASVAB OPERATIONAL TEST AND EVALUATION

By Spring of 1990, the technical development and evaluation of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) were nearing completion. Empirical studies had shown that CAT-ASVAB tests measured the same abilities as their paper-and-pencil counterparts (P&P-ASVAB) and were as reliable, and in many cases, more reliable. The Score Equating Development study (SED) eliminated concerns about equating CAT to P&P. The one psychometric study remaining to be conducted was the Score Equating Verification (SEV), which would provide final equating tables for the Accelerated CAT-ASVAB Project (ACAP) system. By 1990, therefore, CAT-ASVAB was psychometrically ready for nationwide implementation. Psychometric readiness, however, was not the only factor influencing a decision on nationwide implementation of CAT-ASVAB. There were two other very important factors to consider: (a) the cost effectiveness of nationwide implementation, and (b) the impact on operational procedures of implementing computer-based testing.

A 1988 cost/benefit analysis had shown that the cost effectiveness of CAT-ASVAB was questionable. (See Wise, Curran, & McBride, 1997, for details.) This study, however, was limited in that it considered using the CAT-ASVAB in very much the same way as the P&P-ASVAB was being used. The study neglected to take into account the flexible nature of a CAT and placed CAT-

ASVAB in the 1980s' group-paced, lock-step processing environments of the Military Entrance Processing Stations (MEPSs) and Mobile Examining Team Sites (METSs). In addition, there had never been an opportunity to collect empirical data on how the CAT-ASVAB would perform in an operational environment. While the SED had been conducted in the MEPSs/METSs environment, it was a non-operational research study that required administration of CAT-ASVAB and P&P-ASVAB to equivalent groups and, therefore, required the typical lock step processing. The SEV, while operational, still required the group-administered, lock step processing.

During the 1989-90 timeframe, there was little Service policymaker support for nationwide implementation of the CAT-ASVAB. This could be contributed in part to the negative findings of the 1988 cost-benefit analysis. In fact, most people in the Joint-Service ASVAB arena felt that the program should be stopped until results from the Enhanced Computer Administered Test (ECAT) study (Wolfe, Alderton, Larson, Bloxom, & Wise, 1997) were available. During the 1990-91 timeframe, however, several events occurred that put the CAT-ASVAB back on track. First and foremost, Captain James Kinney became Director of the Recruiting and Retention Programs Department, the Navy office that managed the CAT-ASVAB program. Coming from a recruiting background, Captain Kinney immediately saw the potential benefits of CAT-ASVAB. He tasked the Navy Personnel Research and Development Center (NPRDC) with developing a plan for limited implementation of CAT and convinced those in his chain of command to support the idea. Second, several of the higher level managers in various recruiting commands visited SEV sites and saw CAT-ASVAB in operation. As did Captain Kinney, they also saw the potential benefits of the CAT-ASVAB and became strong supporters of the program. Third, the Defense Manpower Data Center (DMDC), as lead agency for the ASVAB, was tasked to look at CAT-ASVAB concepts of operation and to conduct a new cost-benefit analysis. Empirical data on an operational CAT-ASVAB system would provide valuable information in conducting their analyses. DMDC, therefore, supported limited implementation of the CAT-ASVAB as a means of collecting the necessary data. These combined events led to development of a plan for the CAT-ASVAB Operational Test and Evaluation (OT&E) and to approval of this plan.

Operational Test and Evaluation Issues

Since data from the OT&E would be used in helping to define CAT-ASVAB concepts of operation and in conducting a new cost-benefit analysis, careful consideration was given to the issues that needed to be addressed. The goal was to collect the most valuable information possible while minimizing the impact on the MEPSs' mission of processing applicants. Following is a list of the questions asked:

- *Flexible start*. Since all test instructions are automated, CAT-ASVAB allows for a "flexible start," where examinees start the test at different times. This flexible-start procedure gives applicants and recruiters more flexibility compared to the conventional group-administered testing procedure, but how does it affect other applicant processing operations, such as applicant check-in and medical examination?

- *Processing of test scores*. Since scores are automatically computed, does CAT-ASVAB save a substantial amount of score processing time? Are procedures for electronically transmitting scores to the main processing computer easy to use and reliable?

- *Equipment needs*. How much equipment is needed at each site, and how are equipment needs affected by the flexible-start procedure?

- *TA training and performance*. How much time should be allowed for Test Administrator (TA) training, and how does the amount of training impact TA performance?

- *User acceptance*. What are the reactions of applicants, recruiters, and MEPS personnel to CAT-ASVAB? Do the flexibility and shorter test times provided by CAT-ASVAB make it easier to schedule applicants for testing, save recruiter and MEPS personnel time, and reduce travel costs?

- *Security issues*. Is the system secure? Extended operational data collection allows the assessment of procedures for identifying potential security problems. It also allows the evaluation of the effectiveness of item exposure control.

- *Administration of experimental tests*. Can experimental tests be easily added to the end of the battery? Since CAT-ASVAB takes less time than the P&P-ASVAB, the Services might be able to add experimental tests to the end of CAT-ASVAB, allowing for pilot testing and data collection to evaluate adverse impact.

- *System performance*. Does the system meet all operational requirements? Is the software easy to use? How does the hardware perform?

Approach

The general approach was to implement CAT-ASVAB at a small number of operational sites, provide some specific guidelines for its use, such as flexible start times, and see what happens. Prior to implementation at a site, program managers met with the MEPS personnel in the selected area to prepare them for this new way of testing. Data collection for this effort began in June 1992 and, for the purposes of this study, ended in February 1993. The CAT-ASVAB OT&E system, however, remained in operational use until 1996, when it was replaced by the "next generation" CAT-ASVAB system.

Test Sites

The military uses two types of sites to administer the ASVAB: Military Entrance Processing Stations (MEPSs) and Mobile Examining Team Sites (METSs). MEPSs are stationary sites where all processing, including aptitude testing and medical examinations, is conducted. There are approximately 65 MEPSs nationwide. At the MEPSs, military personnel administer the ASVAB and conduct test sessions four or five days a week. On the other hand, METSs are usually temporary sites that offer only ASVAB testing. There are approximately 600 METSs nationwide. If an applicant passes the test at a METS, he or she must go to the associated MEPS for all other processing. Office of Personnel Management personnel usually administer the P&P-ASVAB at a METS, and testing schedules vary widely, from four sessions a week to one session a month.

Four MEPSs were selected as CAT-ASVAB OT&E sites: San Diego, CA; Jackson, MS; Baltimore, MD; and Denver, CO. These MEPSs were selected based on location and number of applicants tested. In addition, one METS was selected: Washington, DC. This METS operates under the Baltimore MEPS. It was selected based on the suitability of the facilities for computer-administration and the number of weekly test sessions. At all the OT&E sites, CAT-ASVAB was administered to all military applicants, and the CAT-ASVAB test scores served as the scores of record for these applicants.

A fifth MEPS was added as the study got underway: Los Angeles, CA, MEPS. In May 1992, the Los Angeles MEPS was partially burned during the Los Angeles riots, and it was forced to move to temporary quarters; it lost the capability to score P&P-ASVAB and to provide medical processing. So applicants in the Los Angeles area were bused to San Diego for this part of the processing. The U. S. Military Entrance Processing Command (USMEPCOM), concerned that the San Diego MEPS would be processing over twice their normal load and implementing a new system at the same time, asked the San Diego MEPS managers if they wanted to delay implementation of CAT-ASVAB. San Diego, however, was anxious to begin the implementation, as the MEPS personnel in San Diego viewed CAT-ASVAB as a means to help with their overload. In fact, after the system had been in operational use in San Diego for a few weeks, the San Diego MEPS Testing Officer proposed setting up CAT-ASVAB testing at the temporary Los Angeles site as well. This way, applicants could be tested near home and bused to San Diego for medical processing only if they qualified on the aptitude battery. This approach would save a substantial amount of time and money. In full support of this request, the Navy, as lead Service, sought and received approval to use CAT-ASVAB operationally in Los Angeles on a temporary basis only. CAT-ASVAB was such a benefit to the MEPS, the Commander asked to have Los Angeles included permanently in the OT&E. The Navy, agreeing to pay all costs associated with the addition of this site, sought and received approval from the Manpower Accession Policy Steering Committee (MAP) to continue the OT&E in Los Angeles.

To allow for comparisons between CAT-ASVAB and P&P-ASVAB, five control MEPSs, administering P&P-ASVAB, were selected: Philadelphia, PA; New Orleans, LA; Portland, OR; San Antonio, TX; and Fresno, CA. Several factors were considered in selecting the control sites, including (a) size/throughput, as indicated by the number of examinees tested; (b) demographic characteristics of the examinees, including score levels on the Armed Forces Qualification Test (AFQT), percent completing high school, and gender and race distributions; and (c) geographic size of the region served, as indicated by percent tested in the central MEPS and the number and size of the METSs associated with each MEPS. Statistics from a 13-month period (Oct. 1991 through Oct. 1992) were used in selecting the control MEPSs.

Data Collection Procedures

Data were collected using several instruments: CAT-ASVAB; administration of questionnaires to recruiters, applicants, and MEPS personnel; on-site observation; and interviews.

CAT-ASVAB. In the natural course of administering CAT-ASVAB, data on all interactions between the applicant and the computer system are saved. This includes item-response data, item-response latencies, test times, instruction times, number and type of help calls, and failure/recovery information (if a computer failure occurs). Any unusual events, such as an applicant leaving during testing, are also documented by the TAs.

On-Site Observations. During the first month of testing at each site, NPRDC researchers were on site to observe test administration. After this first month, periodic visits were made to each site. Based on these observations, the reactions of TAs, recruiters, and applicants to CAT-ASVAB were documented.

Interviews. Researchers who were conducting on-site observations also conducted informal, unstructured interviews with MEPS personnel and recruiters. In addition, informal interviews were conducted periodically by telephone.

Questionnaires. Two separate questionnaires were developed, one for recruiters, and one for applicants. The Recruiter questionnaires contained 25 questions, with the majority of the questions focusing on meeting testing goals, factors affecting amount of travel, flexibility of scheduling applicants for testing, and effects of immediate scores. The recruiter questionnaire administered at

CAT-ASVAB sites contained an additional seven questions about their reactions to CAT-ASVAB. Recruiter questionnaires were administered several months after the start of the OT&E to give recruiters using the OT&E sites a chance to evaluate the CAT-ASVAB.

The Applicant questionnaires contained 23 questions designed to measure examinees' general reactions to the test battery, focusing on test length, difficulty, fairness, clarity of instructions, and feelings of fatigue and anxiety. Applicant questionnaires were administered for one-to-two months following the start of the OT&E. Table 10-1 shows the sample sizes.

Table 10-1. Questionnaire Sample Sizes

	Number of Persons		
	OT&E Sites	**Control Sites**	**Total**
Recruiter Questionnaires	167	175	342
Applicant Questionnaires	1,550	1,497	3,047

Results

Flexible-Start Assessment

At the start of the OT&E, all of the CAT-ASVAB MEPSs used a flexible-start option. Each MEPS set an arrival window during which applicants could come in and start the test. For example, at the San Diego MEPS, applicants could arrive and begin the test anytime between the hours of 4:00 p.m. and 6:00 p.m. Recruiters and applicants found that flexible start reduced scheduling problems. MEPS personnel were initially concerned about the flexible-start option because it was so different from the fixed start time for group administration. They found, however, that the procedure worked well. The one disadvantage of using flexible start was that it required two MEPS personnel to be available during the arrival window, one to check applicants in and one to administer the test.

As the OT&E continued, MEPSs that tested in the afternoon or evenings continued to use flexible start. The MEPSs, however, discovered that CAT-ASVAB made the concept of one-day processing very feasible. Some of the MEPSs began conducting early morning sessions so that the applicant could complete processing the same day. In these early morning sessions, the MEPSs tended to minimize flexible start, keeping the arrival window very short so that all applicants would be finished early enough to complete other processing.

Processing of Test Scores

CAT-ASVAB does save TAs a substantial amount of time in processing test scores. When administering the P&P-ASVAB, all answer sheets must be scanned, which is tedious and time-consuming. At the MEPSs, CAT-ASVAB scores were transferred to the main computer by carrying a disk from the testing room to another room, where the data were uploaded in a matter of minutes. Data transfer procedures were very reliable. In the “next generation” system, this process was further simplified by the use of a computer network. Scores are transferred from the testing network to the main computer at the touch of a key.

At the Washington, DC, METS, scores were telecommunicated to the main computer at the Baltimore MEPS. This procedure proved to be less reliable than desired, due to the use of obsolete hardware and software. With the OT&E system, Washington METS personnel had to coordinate the exact time of the transfer with Baltimore MEPS personnel to ensure that the computer receiving the data was in the “host,” or receiving, mode. To complicate the situation, host mode had a time-out feature that automatically took the computer out of this mode after a certain number of minutes. If all data transfer steps were not followed in the exact order at both ends, the transfer failed. This problem, however, would disappear once CAT-ASVAB was transitioned to a new system for METSs use, and an up-dated data communications program could be used.

Equipment Needs

Each of the CAT-ASVAB OT&E sites, with the exception of the Los Angeles MEPS, had enough equipment to test maximum session sizes for that MEPS. However, the use of flexible start and the shorter testing time of the CAT-ASVAB battery reduce equipment requirements. It is estimated that, on the average, a MEPS requires half as many computers as examinees in a maximum session. For example, Los Angeles, one of the largest MEPSs in the country, had 30 computers during the OT&E, with the capability of testing 60 applicants in the same amount of time as a typical P&P-ASVAB test session. In fact, Los Angeles has tested larger numbers than this in an evening session. Equipment needs are less than projected in earlier studies, reducing the cost of implementing CAT-ASVAB nationwide.

Test Administrator Training and Performance

The instruction program that was initially developed to train CAT-ASVAB TAs took about four days of classroom training. At the beginning of the OT&E, it became clear that MEPS personnel could not devote four days exclusively to CAT-ASVAB training. Therefore, for the OT&E effort, the training program was changed to include two days of classroom training and two days of on-the-job training (OJT). This revised training program for TAs has been successful, both at the MEPSs and METSs.

During the classroom part of the training, TAs met all course objectives. The two days of OJT seemed adequate for training TAs to run the system under normal conditions. In addition, observation of performance on the job confirmed this conclusion.

Very few problems were encountered in training. One problem that was noted, however, was that "group-administered" classroom training was not ideal due to the high turnover in TAs and scheduling difficulties. Therefore, a self-administered, computer-based training program using an intelligent tutoring system has been developed.

Another problem that was encountered was TA performance under unusual conditions. Occasionally, a site experienced some type of system failure, and the TA did not know how to recover. While the system was designed to recover from all failures, and procedures for all types of failure/recovery were documented in the User’s Manual, certain types of failures happened so infrequently that TAs needed assistance in the recovery. In these cases, TAs called NPRDC for guidance. This demonstrates the need for some type of “help line,” particularly during nationwide use of the system.

Overall, CAT-ASVAB helped streamline test administration procedures, making it easier for TAs to perform their duties. They no longer needed to read instructions, time tests, or scan answer sheets. Automating these functions also resulted in standardization across all the testing sites.

User Acceptance

Recruiters' Reactions. Based on interview results, recruiters’ reactions were very positive overall. Recruiters were very enthusiastic about the shortened testing time and the immediate scores provided by CAT-ASVAB. Some recruiters felt that because of the standardized testing environment, CAT-ASVAB is a fairer test than the P&P-ASVAB. Some recruiters reported traveling a substantial extra distance so that their applicants could test on CAT-ASVAB rather than P&P-ASVAB. Recruiters, however, expressed some concerns about the differences between CAT-ASVAB and P&P-ASVAB. For example, some feared that CAT-ASVAB might be more difficult than the P&P-ASVAB because it is computer-administered. Other recruiters received reports from high ability examinees that the test was really difficult and, therefore, believed that their applicants would have a better chance qualifying with the P&P-ASVAB. It was also difficult for recruiters to understand how a test with 16 items could provide a number-correct score of 35. It was found that conducting sessions where recruiters could see a demonstration of CAT-ASVAB, learn how the test worked, and have the opportunity to ask questions would address these concerns. As a result of this finding, educational materials on the CAT-ASVAB system were developed prior to nationwide implementation and were distributed to MEPSs and recruiting personnel.

Questionnaire results showed few differences between the reactions of recruiters from the OT&E sites and the control sites. At both types of sites, recruiters felt that the availability of immediate scores and a more flexible testing schedule would greatly increase their productivity. About 65 percent of the recruiters at CAT-ASVAB sites felt that CAT-ASVAB saved them 30 to 90 minutes of time per testing session. About 33 percent felt that applicants were more willing to take the ASVAB when it was CAT-ASVAB, while 11 percent felt it decreased the applicants' willingness. About 16 percent felt that taking CAT-ASVAB instead of the P&P-ASVAB increased the applicants' willingness to enlist, compared to 5 percent who felt it decreased it. About 25 percent of the recruiters were willing to travel at least 30 minutes more so that applicants could take CAT-ASVAB.

Applicants' Reactions. In comparing questionnaire responses from the CAT-ASVAB examinees to the responses from the P&P-ASVAB examinees, the two groups were significantly different on most questions. These differences were small, with both groups giving positive responses about the ASVAB. P&P-ASVAB examinees were slightly more positive than CAT-ASVAB examinees on the following issues: general feelings about the test, feelings of anxiety, test difficulty, and amount of eye strain. CAT-ASVAB examinees were slightly more positive than P&P-ASVAB examinees on the following: general fatigue, test fairness, test length, time pressures during the test, clarity of instructions, convenience of testing schedule, test enjoyability, and the interest level of the test. There were no significant differences between the two groups on distractions from the surrounding environment.

Some of the significant differences in reactions to the tests could be attributed to the adaptive nature of CAT-ASVAB. For example, high-ability examinees are administered more difficult test items than they would typically take on a P&P-ASVAB. This may cause them to be more fatigued at the end of the test and to perceive the test as being very difficult, possibly increasing their anxiety level. On the other hand, because CAT-ASVAB is an adaptive test, and therefore much shorter than the P&P test, examinees were more positive about test length.

Some of the differences in reactions to the test, however, could be attributed to the medium of administration: computer versus paper-and-pencil. Taking the test on the computer causes eye strain slightly more often but is perceived as more enjoyable, more interesting, and having less time pressure. Computer administration also offers flexibility in the testing schedule; examinees are not required to start the test as a group.

Since CAT-ASVAB was administered with a flexible test start time, the finding of no significant difference in terms of environmental distractions was positive. Initially, there was some concern that examinees coming and going during a CAT-ASVAB test session would disturb examinees taking the test. Questionnaire results and on-site observations alleviated this concern. Once the examinee started the test, the focus was on the test, not the surrounding environment. Overall, examinees' reactions to CAT-ASVAB were very positive. In general, we found that most examinees preferred taking CAT-ASVAB to the P&P-ASVAB.

Reactions of MEPS Personnel. Based on interviews and on-site observations, the reactions of MEPS personnel have been overwhelmingly enthusiastic. Initial skepticism on the part of the MEPS commanders at the OT&E sites soon gave way to "couldn't live without it" attitudes. TAs also had a very positive reaction to CAT-ASVAB, preferring to administer it rather than the P&P-ASVAB. TAs felt that CAT-ASVAB allowed them to make much more efficient use of their time. These positive reactions are the reason that the CAT-ASVAB system remained in operational use at the OT&E MEPSs even after data collection for purposes of the OT&E had ended.

Test Security

CAT-ASVAB test items reside on several floppy disks that are never accessible to applicants. In addition, test item files are encrypted. During test administration, the items are loaded into volatile computer memory, disappearing when the computer is turned off. Test compromise from theft of items is much less likely with CAT-ASVAB than P&P-ASVAB. Another security issue does exist, however, and that is security of the computer equipment. MEPSs are very secure, making computer theft unlikely. During the OT&E, no computer equipment was stolen from a MEPS or METS. This may become more of a problem, however, if future use of CAT-ASVAB includes the use of portable notebook computers in the METSs.

Administration of Experimental Tests

To date, one experimental test has been added to the CAT-ASVAB, Assembling Objects (AO), a spatial test. From an implementation standpoint, the addition of this test was "painless." Since it is computer administered, no booklets had to be printed or answer sheets modified. An additional software module was simply added to the CAT-ASVAB test-administration software. In addition, since CAT-ASVAB takes so much less time than the P&P-ASVAB, there were few complaints about the small amount of additional testing time needed to administer the AO test.

System Performance

The OT&E has shown that the CAT-ASVAB system meets all ASVAB testing requirements and that the software is fairly easy to use. It has also helped to identify procedures that could be automated and incorporated into the system to streamline ASVAB testing, (e.g., the automatic generation of forms typically completed by hand). In addition, it has helped to identify CAT-ASVAB procedures that are unnecessary or too time-consuming. Some of the general findings are as follows:

- Random assignment of examinees to machines is not necessary. This procedure requires entering names and social security numbers at the TA station before testing can start, therefore delaying the start of testing. The purpose of this procedure was to ensure that, when session sizes were smaller than the number of computers in the room, the same machines were not used over and over. It is much more efficient, however, to tell the TAs to space the examinees out. Elimination of this procedure will prevent accidentally seating the examinee at a computer designated for another examinee.

- The stand-alone mode of operation takes too long and requires the handling of too many disks. This procedure could not be changed for the HP Integral Personal Computer-based system (HP-IPC), as the system has no hard disk drive and the floppy drive will not read high density disks. In the "next generation" system, however, the stand-alone mode has been streamlined as much as possible.

- The interactive screen dialogues need to be less wordy. If the screens are too wordy, the TAs tend not to read them.

- Procedures in general need to be streamlined. There are too many cases where the TA must remember that a certain procedure must be completed before another, or at a certain point in the session. While, during the course of the OT&E, procedures have been streamlined and automated, due to limitations of the HP-IPC based system and the network for this system, certain desirable changes could not be made. These types of changes, however, are being incorporated into the design of the "next generation" system.

The hardware performed very well during the course of the OT&E. The HP-IPCs that were used in this evaluation were purchased in the 1985 to 1987 timeframe. They were used at the OT&E sites until the end of 1996. By current computer standards, they were, therefore, fairly old. Yet, hardware problems were minimal. The majority of the hardware problems were with the floppy drives and the memory boards. All other computer components performed well above expectation. During the OT&E, non-functioning equipment was shipped to NPRDC for repair, and repairs were performed by NPRDC staff. Since these machines were obsolete, the most challenging part of repairing the equipment was to purchase needed parts within a reasonable timeframe. Another challenge was to keep track of equipment inventory, since there was a lot of movement of equipment between MEPSs and NPRDC. For nationwide implementation, the simplest approach to equipment maintenance is to have an on-site maintenance contract. This approach, however, must be evaluated for cost-effectiveness.

Summary

The OT&E marked the turning point in the CAT-ASVAB program and was the program's biggest achievement. This was true from both a manager's and researcher's perspective. From a manager's perspective, the OT&E demonstrated that CAT-ASVAB meets the needs of recruiters, applicants, MEPS personnel, and USMEPCOM Headquarters. It led to the enthusiastic support of CAT-ASVAB by MEPS and recruiting personnel, which in turn influenced the outcome of the 1993 cost-benefit analysis. Due to the success of the OT&E, in May 1993, the Manpower Accession Policy Steering Committee (MAP) approved implementation of CAT-ASVAB at all MEPSs nationwide. This marked the high point in the CAT-ASVAB program.

From a researcher's perspective, there has been no greater reward than conducting the CAT-ASVAB OT&E. After years of hard work in developing and evaluating the system, we were able to not only see the system in operational use, but to become an integral part of this limited operational implementation. We were able to go out into the operational environment and interact daily with the users of the system: MEPS personnel, applicants, and recruiters. While we expected the system to work well, we did not necessarily expect such a strongly favorable reaction from all the users of the system. For the numerous researchers who have contributed to this program, and in particular, for those researchers working on the program during this effort, the CAT-ASVAB OT&E made those years of hard work all worthwhile.

References

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., Alderton, D. L., Larso, 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.

Chapter 11

DEVELOPMENT OF A SYSTEM FOR NATIONWIDE IMPLEMENTATION

The 1993 approval to implement CAT-ASVAB nationwide started a new phase in the CAT-ASVAB program. One major aspect of this phase of the project was to develop a new CAT-ASVAB system. While the Hewlett Packard–Integral Personal Computer (HP-IPC), used for the Accelerated CAT-ASVAB Project (ACAP), had served its purpose well, by 1993 it was obsolete and no longer manufactured. Developing a new system involved selecting a new computer platform and networking system, designing an input device comparable to the one used in ACAP, and developing new test administration software. This chapter describes all phases of the system development for nationwide implementation of CAT-ASVAB.

Computer Hardware Selection

Computer hardware selection consisted of four steps: (a) developing hardware requirements, (b) conducting a market survey of available systems, (c) evaluating these systems, and (d) developing hardware specifications. In selecting the hardware for nationwide implementation, lessons learned from ACAP were extremely valuable. This was particularly true while conducting the initial step – developing hardware requirements.

Hardware Requirements

The hardware requirements for a new CAT-ASVAB computer system were based on the capabilities of the HP-IPC, with input from the operational CAT-ASVAB MEPS personnel. The new computer system had to meet or exceed system specifications in certain areas. Other requirements, however, were new, having been developed as a result of our experience with the HP-IPC.

Hardware requirements as defined by the ACAP system. The hardware and software system for ACAP was designed, developed, and implemented using the HP-IPC running under a UNIX (System V) operating system. The HP-IPC meets the following requirements:

Portability. The HP-IPC is a portable computer system. It is classified as a transportable suitcase-type portable. It weighs 25.3 pounds, and can be (somewhat) easily assembled and disassembled and moved from one location to the other. It is fully self-contained, with a built-in monitor, floppy disk drive, ink jet printer, and detachable keyboard. It is designed for ease of operation and flexibility.

The 1993 decision to implement CAT-ASVAB nationwide was limited to implementation at Military Entrance Processing Stations (MEPSs). Since MEPSs are permanent sites, they do not require portable systems (i.e., they can use desktop computers). The only sites requiring portable computers are the Mobil Examining Team Sites (METSs), which are typically temporary sites requiring equipment set-up and take-down for each session. However, since implementation at METSs is under consideration, it was necessary to select a computer platform that would meet the needs of both types of sites. To fulfill this requirement, we decided to evaluate desktop computers for MEPSs and portable notebooks for METSs. The advantage in using desktop computers where possible is that they are less costly, easier to maintain, easier to upgrade, and less susceptible to theft. There are some disadvantages to having two types of computers. First, there is a potential for increasing the amount of effort dedicated to software development and software acceptance testing. Second, both types must be equated to the ASVAB reference form, increasing the cost and complexity of score equating.

In evaluating systems for portability the following factors were considered: weight, size, ability to easily assemble, disassemble, and move from one location to the next; and ability to operate as a stand-alone unit. Based on experiences in the field, the new system had to have a substantial size and weight advantage over the HP-IPC system. A portable computer system should be under 10 pounds, and 7 pounds if possible.

Adaptability. The HP-IPC system provides for two additional expansion slots that can be used for additional (random access memory [RAM]) and (input/output) interface capabilities. While only one printer per test site is required, the HP-IPC system comes with a built-in ink jet printer and an IEEE-488 interface, which allows for additional peripherals. The HP-IPC system has a 3.5 inch floppy disk drive. It also has a detachable keyboard, facilitating modifications to the examinee input device.

The new computer system had to be expandable, allowing for specific system growth on the system's main-board. It had to have a minimum of four megabytes of RAM, expandable to 16. A minimum of two I/O interfaces were required, one containing a parallel and serial port for attaching a printer and/or modem, and one for network interfacing. The new system had to be equipped with a 3.5 inch floppy disk drive to allow for flexibility in software design, and had to have the ability to link to a printer or other peripherals as required for operational field use. Ease of keyboard modification or attachable add-on keypads was considered highly desirable.

Performance capabilities. The HP-IPC runs under an eight megahertz (Mhz) processing speed. It is capable of multi-tasking. The new computer system processor speed requirement was based on 1993 industry standards which were faster than 8 Mhz. (The minimum computer processor speed evaluated was 25 Mhz.) While multi-tasking is desirable for software development purposes, it is not necessary for operational examinee test administration or associated system functions needed during test administration.

Monitor. The HP-IPC has a monochrome monitor with a 512 (horizontal) x 255 (vertical) pixels electro-luminescent display. The screen size is 9 inches measured diagonally, 8 inches wide by 4 inches high. The display can be configured for up to 31 lines with up to 85 characters per line, but the ACAP system uses dot matrix dimensions of 5 x 8 dots embedded in a 7 x 11 field. At this resolution, it is possible to display 23 lines with 73 characters per line on the HP-IPC screen.

To display graphics items clearly, the monitor video resolution screen for the new computer system was required to have as a minimum the 1993 industry standard Video Graphics Adapter (VGA). The number of lines per screen and characters per line of the ACAP system was also a minimum requirement so that each item will fit on one screen. The new system did not need to meet other monitor specifications for the HP-IPC, as an equating was conducted prior to implementation. It was required as a minimum that all new computer systems have a built-in external VGA monitor adapter, SVGA being more desirable.

New Requirements. The new system had to meet requirements in addition to those met by the HP-IPC system. One of the biggest problems with the HP-IPC was it did not sell well in the market place, and it was very specialized, making parts costly and hard to obtain. To whatever extent possible, the new system needed to be a commonly used computer system so that replacement parts could be procured near the test sites. This would substantially reduce maintenance costs, would provide for future growth of the system, and would delay system obsolescence. The HP-IPC does not have internal storage capability, limiting system flexibility and expansion capabilities dramatically. The new system had to have internal mass storage capability. This would allow for growth and flexibility in system applications. In addition, a portable system should have upgrade capability similar to that of a desktop computer. A portable system should also have a minimum FCC Class B certification.

Types of Available Systems

An evaluation of the computer systems that were on the market in 1993 took into consideration the various types of microprocessors and the types of portable computers.

Types of microprocessors. There were three predominant microprocessors on the market which fit the personal computer systems profile: (a) Intel (80386/80486/80586) based or compatible, (b) Motorola (68000/680xx) based, and (c) RISC (Reduced-Instruction-Set-Computer) based microprocessors. Intel normally operates under the Disk Operating System (DOS) but does have UNIX and other operating systems capability. Motorola normally operates under a UNIX operating system. RISC runs under a UNIX operating system and is the newest microprocessor on the market.

Types of portable computers. There were two basic categories of portables: those weighing under or over 15 pounds. Styles that fit in the first category are the handheld, the notebook, and the laptop; they usually resemble a clamshell design. These systems are typically referred to as notebooks and portables. Styles that fit in the second category are suitcase and, occasionally, those having the clamshell design. These systems are typically referred to as transportables or luggables.

Transportable computers, similar to the HP-IPC, do not meet minimum size and weight requirements for temporary sites and are too expensive for permanent sites. For these reasons, this category of computers was eliminated from consideration.

Evaluation of Available Systems

A wide variety of desktops (for MEPSs) and notebooks (for METSs) were evaluated as meeting the minimum system requirements. Portable notebook computers, in particular, have grown substantially in performance capability and peripheral expansion capability over the past several years. Previous notebook computer systems seemed to lack the ruggedness needed for operational field use, but technological advancements have established their durability for operational field use. There are certain expansion disadvantages to notebook computers, but performance and physical characteristic advantages outweigh the disadvantages.

The Motorola and RISC-based portable and desktop computers, while meeting minimum specifications, are very limited in type, quantity, and production, and are expensive to purchase, maintain, and upgrade. Systems using the Intel microprocessor, on the other hand, are relatively low cost, widely available, and easy to maintain and upgrade. Based on these findings, IBM-PC/AT (Intel-based compatible) computers were selected as best suited for the new computer platform.

Computer Specifications

Table 11-1 lists the primary computer specifications for the desktop computers and the notebook/laptop computers. These are *not* minimum specifications needed to run CAT-ASVAB software, but specifications that we felt would provide the Government with a reliable, easily maintainable system that has the capability for future expansion. In developing these specifications, we tried to project what would be standard equipment when procuring the systems for implementation. These specifications apply to both the Test Administrator (TA) station and the Examinee Testing (ET) stations.

Keyboard Specifications

The one difference between an ET station and a TA station is the type of keyboard required. Where the TA station requires a full Enhanced AT 101 type keyboard, the ET station requires a modified AT 101 type keyboard. Required modifications include relocating the “A,” “B,” “C,” “D,” and “E” keys, labeling the space bar as “ENTER,” labeling the F1 key as “HELP,” and covering all unused keys. A lot of time and effort went into figuring out how to meet these requirements and still have a durable, easily maintainable keyboard. The ACAP system used a template to cover unused keys and labels to mark the keys needed to take the test. While this method worked reasonably well, over time the templates warped, moved slightly inhibiting key depression, or came unfastened. We experienced some problem with key labels coming off. To avoid these problems, we decided to use blank keycaps on all unused keys. The item response keys (“A,” “B,” “C,” “D,” and “E”) are the original keys moved to the proper location. The “HELP” key (F1) and “ENTER” key (space bar) were labeled using the same process normally used in labeling commercial keyboards. Figure 11-1 shows a picture of the modified ET keyboard.

Table 11-1. CAT-ASVAB Hardware Specifications

<table>
<tr><th></th><th>Desktop</th><th>Notebook</th></tr>
<tr><td>Microcomputer Platform</td><td colspan="2">IBM PC/AT (Intel-Based Compatible)</td></tr>
<tr><td rowspan="2">Microprocessor (CPU)</td><td colspan="2">80486DX (Intel or Intel Compatible) microprocessor
8Kb Internal cache memory</td></tr>
<tr><td>33 MHz or faster</td><td>25 MHz or faster</td></tr>
<tr><td rowspan="2">Mainboard/Motherboard</td><td colspan="2">33 MHz PCI or VLB I/O BUS rated speed
CMOS/ROM BIOS configuration option, during boot-up</td></tr>
<tr><td>Expansion slots, 6 minimum</td><td></td></tr>
<tr><td rowspan="2">RAM</td><td>30 or 72 pin SIMM type modules, with a minimum of 4 MB, expandable to 64 MB</td><td>4 MB, expandable up to 16 MB of RAM</td></tr>
<tr><td colspan="2">70ns or faster RAM</td></tr>
<tr><td rowspan="2">External I/O Bus</td><td colspan="2">One RS-232 Serial I/O port (9-pin)
One Parallel I/O port</td></tr>
<tr><td></td><td>1 external keyboard/keypad port, built-in
1 external mouse port, built-in mouse support must be Microsoft compatible</td></tr>
<tr><td rowspan="3">Display/Video Interface</td><td>Super Video Graphics Array (SVGA) reflective color LCD</td><td>Dual scan color</td></tr>
<tr><td colspan="2">Extended graphics resolution modes, 640 (horizontal) X 480 (vertical) pixels
1MB VRAM</td></tr>
<tr><td>Screen Size, 14” measured diagonally
.28 mm dot-pitch
Non-interlaced and interlaced monitor support
15-pin (DB15) cable, 6 ft.</td><td>Screen Size, 9.5” measured diagonally
Display text up to 80 characters by 25 lines
Viewing angle: greater than “TBS/ TBD” degrees in a horizontal plane
1 external VGA/SVGA port</td></tr>
<tr><td>Floppy Diskette Drive</td><td colspan="2">3.5” 1.44 MB High Density Floppy Disk (HD FDD)</td></tr>
<tr><td rowspan="2">Internal Hard Disk Drive</td><td colspan="2">80MB Internal Hard Disk Drive (80MB measured using no compression software or hardware)</td></tr>
<tr><td>ALL IDE drives must be capable of supporting a second IDE drive from various manufacturers.</td><td></td></tr>
<tr><td>Notebook Size</td><td></td><td>NTE Size (d,w,h) 8.3” x 11” x 1.8”</td></tr>
<tr><td>Notebook Weight</td><td></td><td>NTE 6.3 lbs in weight</td></tr>
</table>

Note. Cells that span both desktop and notebook columns are requirements for both.

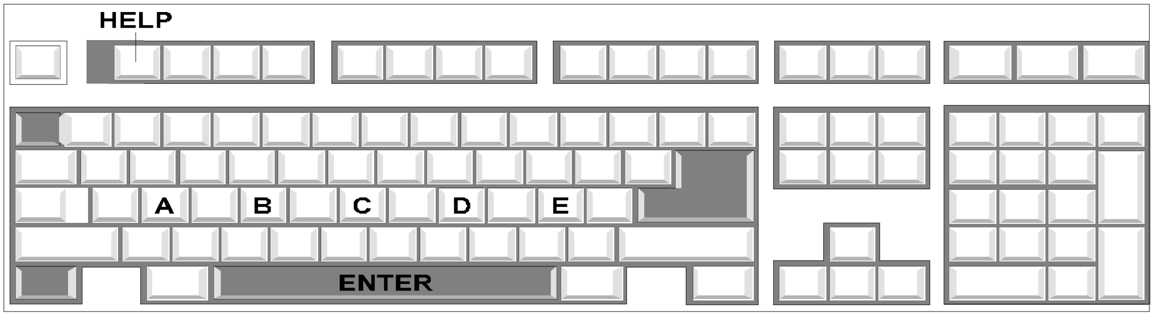


Figure 11-1. Modified ET Keyboard.

Network Selection

Networking of computer systems allows for more efficient administration of CAT-ASVAB, particularly at large sites. Networking helps to eliminate redundancy in procedures, saving a substantial amount of test administrator time when more than ten ET stations are being operated at any one time. For this reason, the HP-IPC CAT system provided the capability of networking, via a local area network (LAN). This is also a requirement of the new desktop computers, but not the portable computers. At this time, notebook computers will not have the capability of networking, as they will be used at the smaller test sites. Networking requires a network interface controller (NIC), cable, and software that runs it. In selecting these components of the network, several options were considered.

Network Hardware

Network interface controller. PC networking hardware consists of using a NIC that provides the physical connection between a computer and the network medium. Several NIC protocols were evaluated.

Arcnet. In 1977, DataPoint Corporation developed Arcnet as a proposed inexpensive solution to connectivity. This protocol allowed up to 255 nodes. Arcnet gives each node a unique ID address in incremental order. It uses a token-passing scheme where a token (sequence of characters) travels to each station according to ascending node addresses. When a PC receives a token, it holds that information and queries other PCs about their ability to accept tokens. When a recipient is available, the system sends the token and continues sending the token to other recipients until the last node receives the token. Because a node may transmit only when it has the token and only after getting an okay from the recipients, Arcnet performance is slow. The data transfer rate is 2 Mbps baseband operation. This may be acceptable if the number of workstations is moderate and their volume of network messages is light. Otherwise, the system will get bogged down by constant group interaction, heavy transmission, or large files. Arcnet's specific hardware and software requirements, along with its proprietary protocol, make it an unpopular network for PCs.

Ethernet. The Xerox Corporation invented this protocol in the early 1970s. It uses a communication technique called Carrier Sense Multiple Access/Collision Detection (CSMA/CD). Workstations with information to send would "listen" for network traffic. If the workstations detect traffic, they pause and listen again until clear. Once there is no traffic, they broadcast the packet (series of bytes) in both directions. The data packets identify the destination workstation by a unique address. Each workstation reads the header of the packet, but only the destination node reads the entire packet. Multiple workstations may transmit simultaneously. When this happens and messages collide, a message goes out to cancel the transmission; the workstation waits a random amount of time and then re-transmits. Ethernet has the advantage of packing the maximum number of messages on the network and producing high-speed performance. This popular protocol (IEEE 802.3) has a data transfer rate of 10 Mbps baseband operation. Because many different platforms support Ethernet, this makes it simple and easy to use Ethernet to link to various computer systems.

Token ring. IBM originally designed this network protocol. It works similarly to Arcnet's token passing scheme, except the tokens travel in one direction on a logical ring and pass through every node to complete the circuit. When a workstation receives the token, it can either transmit a data packet or pass the token to the next station. In this procedure, each node between the originating workstation and the data's destination regenerates the token and all of its data before passing it on. Upon reaching its destination, usually the file server, the receiver reads the data, acknowledges them, and sends the message back into the ring to return to the sender. Again, each workstation along the way reads and re-transmits the token. This scheme creates considerable overhead but assures successful data transmission. Depending on whether twisted-pair or shielded two-pair cabling is used, the data transfer rate is 4 Mbps or 16 Mbps baseband, respectively (IEEE 802.5).

The protocol of choice is Ethernet. We base this on its popularity and the following four factors: (a) it is a low cost network, (b) the protocol is inherently reliable, (c) it is fast, and (d) it has a variety of cabling options. There are many manufacturers of Ethernet NICs that are 100 percent compatible with standards set by the

IEEE 802.3 committee. Eight-bit and sixteen-bit controllers are available for the Industry Standard Adapter (ISA) bus found in desktop PCs. These controllers plug into any open ISA slot and come with connectors for thick-net, thin-net, twisted-pair, or a combination.

Cabling. There are four cabling topologies available for Ethernet: Thin-net (10Base2), thick-net (10Base5), twisted-pair (10BaseT), and fiber optics (10BaseF). Fiber optics is expensive and is only used for long distances. Thick-net is seldom used because its thick cables are bulky and hard to work.

Twisted-pair uses concentrators (hubs) to link the workstations together. This range of ports allows designing networks with simple point-to-point twisted-pair cabling or using structured cabling systems. This gives total flexibility on monitoring and managing the network. Such a setup is easy to configure. If a station fails or the connection between a station and hub fails, all other stations continue to operate. However, if a hub fails, all the stations connected to that hub cease functioning. Twisted-pair cabling provides the capability of running at 100 Mbps.

Thin-net cables are easy to move and connect to workstations. In this type of setup, the trunk segment acts as backbone for all the workstations. Each end of the trunk is a BNC 50-ohm terminator which ends the network signal. Up to five trunks may be connected using a repeater that strengthens network signals. Each trunk supports a maximum of 30 workstations. The nodes connect to the trunk using BNC T-connectors. The biggest advantage of thin-net is that it is low in cost. The disadvantages are that network and station errors are harder to diagnose, if one station goes down there is the potential for all stations to go down, and it can run only at 10 Mbps.

Network Software

There were three options for network software: (a) writing our own network operating system (NOS); (b) selecting a commercial, server-based NOS; or (c) using a peer-to-peer NOS.

Custom developed. Writing our own NOS would be a very large-scale project. First, we would need to select the NIC to use and to develop drivers for that card. Hundreds of NICs are available, and programming drivers are different for each. We would have to solicit technical information from the manufacturer of each NIC we considered. Some NICs come with drivers, but these are usually used for linking with commercial NOS. In the event that a manufacturer discontinued an NIC, developing new drivers would become necessary. Similarly, we would need to provide updates to drivers whenever an NIC changed in revision. Once we completed development of drivers, we would need to write a suite of functions to conform with the IEEE 802.3 Ethernet protocol.

Server-based. The major manufacturers of server-based networks are Novell NetWare and Banyan VINES. With this type of network, each workstation attaches to the server via a protocol driver and workstation shell that loads into memory. The protocol driver creates, maintains, and terminates connections between network devices. The shell intercepts application requests and figures out whether to route them locally either to DOS or to the network file server for processing by the NOS. This creates very little overhead as the workstations interact only with the server.

Configuring a PC for use in a server-based network is quite simple. Drivers come with the NIC, which makes it easy to link with the NOS. Finally, manufacturers supply updates to drivers of each product.

Peer-to-peer. With peer-to-peer networks, only a subset of network commands is available. Major packages are Artisoft's LANtastic and Novell's NetWare Lite. This type of network is also configurable as server-based, although that configuration would involve more overhead. Peer-to-peer networks load seven terminate-and-stay-resident (TSR) drivers into memory. These drivers take over the operating system by assuming that each workstation will communicate with all the others. In the CAT-ASVAB configuration, this is not true. ET stations communicate with the TA station, but not with other ET stations. For peer-to-peer networks, processing appears slower whenever a workstation transmits to the server. Each workstation monitors all input and output. Another shortcoming is their compatibility with networks on other platforms. The main advantage of this type of LAN is the sharing of resources with other nodes without implementing a dedicated server. Many good features exist in peer-to-peer networks which are missing in server-based networks. However, these features are enhancements that the CAT-ASVAB environment does not require.

Other considerations. Each server-based and peer-to-peer system is unique to the manufacturer and is not easily cross-compatible. For instance, LANtastic is not directly compatible with NetWare Lite. To get the NOS from two vendors to talk to each other usually requires purchasing additional software to link the two. Things to consider are compatibility, stability, connectivity options, ease of use, and technical support issues. There are many more Novell CNEs (Certified Network Engineers) than Banyan certified engineers. Most important is to standardize and not consider low-end products. If the manufacturer of a proprietary system goes out of business, support and parts supplies are no longer available (LAN: The Network Solutions Magazine, September 1993). When looking at hardware and software configurations on PCs and other platforms (VAX, Sun, Apple), Novell is used as the measure of network compatibility. Many products carry Novell's stamp of approval indicating "YES NetWare Tested and Approved."

The CAT-ASVAB TA station is required to communicate with the MEPCOM Integrated Resource System (MIRS) system. Initial specifications showed MIRS to be a UNIX workstation running ethernet and Transmission Control Protocol/Internet Protocol (TCP/IP). Novell’s NetWare 3.11, the version on the market at the time of this evaluation, already included the TCP/IP Transport, which is a collection of protocols, application programming interfaces, and tools for managing those protocol. Other NOSs support TCP/IP through add-on packages which increase network traffic and can slow down response times.

Network Selected

After considering CAT-ASVAB's current and future network requirements, the following networking hardware and software were selected: (a) an ethernet NIC; (b) twisted-pair cabling; and (c) Novell NetWare, a server-based NOS. To maintain compatibility across all types of computers, we decided that the file server required by this networking option must meet the same computer specifications as the TA and ET stations. This combination of hardware and software was found to meet all CAT-ASVAB current and projected networking requirements and to be cost-effective.

Software Development

Since the CAT-ASVAB software running on the HP-IPC was in operational use during the time that CAT-ASVAB software was being developed for the IBM-PC compatible, names were assigned to each to avoid confusion. The former is referred to as HP-CAT and the latter as PC-CAT. HP-CAT functional requirements were used as a baseline for the development of PC-CAT, with some exceptions. In particular, “lessons learned” from the CAT-ASVAB Operational Test and Evaluation (OT&E) were used in modifying the functional requirements. Differences between the functionality of HP-CAT and PC-CAT are noted in the paragraphs below.

Minimum System Requirements

Since the computer platform selected for the next generation CAT-ASVAB is an IBM PC/AT compatible, single-user computer, PC-CAT is written for this machine with a minimum configuration of an Intel 80386 CPU, 640 K of conventional memory, and at least three megabytes of extended memory. The speed of the CPU must be at least 25 megahertz. A multi-syncing VGA monitor (interlaced or non-interlaced) with a minimum resolution of 640 x 480 is required. While we had the option of programming the system to run under Windows, we elected to develop a MS-DOS based system. We had learned from the ACAP system that taking the simplest and cleanest approach possible minimizes problems in the field. Windows offered no advantage and requires substantially more resources. PC-CAT requires MS-DOS 5.0 or higher. PC-CAT is fully upwards compatible, but not downwards compatible.

Programming Language

From a technical standpoint, the programming language of choice remained "C." The primary reason for this choice was that HP-CAT had been written in the C language, and many of the psychometric routines for test administration were transportable to the new system (i.e., item selection, test scoring, expected test completion time). About 80 percent of the code, however, was rewritten and designed specifically for the MS-DOS environment. This is a reasonable approach since much of the original OT&E software (dating back to 1986; Folchi, 1986) was designed and written when not all the functions to be supported were known. Over time, as more and more software was added and/or revised to reflect new functional specifications, the required "re-engineering" produced a greater level of convolution in software logic and inefficiency in software that would not have been the case if all of the functions were known at the start. Now that all of the functions are known, and in fact, in the case of the TA station, simplified, the more preferred path, and the one ultimately selected, was to design and write new software relative to the new environment, but taking advantage of that software from the OT&E code that reflected common functions.

A further technical consideration was the choice of a C compiler to support software development and execution. Among those features which characterized HP-CAT was the use of RAM as an electronic storage medium for testing data, particularly the test item files (Rafacz, 1994). This reduced the need to access a mechanical device such as a floppy drive to retrieve test items, thus minimizing wear-and-tear on those devices. Most importantly, however, the storage of test items in volatile RAM provided maximum security for the items because they disappeared once power was removed. Needless to say it was desirable to use the same type of design for PC-CAT, but within an MS-DOS environment. This required using a compiler that included expanded memory capabilities, analogous to that available on the HP-CAT system via the UNIX operating system. The Borland C++ 3.1 compiler provided the necessary capability.

To support software development, a comprehensive collection of functions, referred to as the “In-house Library,” was developed. Most of these functions are written in Intel assembly with some intricate C coding. The In-house Library includes graphics functions and functions to control the use of expanded memory, keyboard interrupts, and high resolution timings. The In-house graphics functions are faster than Borland C compiler routines.

Software Components

There are two major software components in PC-CAT: (a) the Examinee Testing (ET) station software, and (b) the TA station software. Unlike HP-CAT, PC-CAT does not include Data Handling Computer (DHC) software, as these functions will be handled by the MEPS MIRS system. Like HP-CAT, PC-CAT can function in either a networking mode or a stand-alone mode of operation.

ET software. The functionality of the ET software for PC-CAT is almost identical to that of HP-CAT. There are some differences, however. First, with PC-CAT both forms of CAT-ASVAB are loaded into memory, allowing for selection of form at the ET station. In comparison, HP-CAT could store only one form in memory, not because the capability did not exist, but rather because the cost of RAM was too prohibitive. The net result is that PC-CAT enjoys a simplification of some of the software routines concerning the placement of examinees at stations and certain failure recovery situations. Second, because the specification for the random assignment of examinees to testing stations has been removed, test administrators may now seat examinees essentially in a "free-form" format. Test administrators enter the examinee’s social security number at the ET station. In networking mode, the TA station will “get” the examinee identifying information from the file server. This will allow the examinee to start testing immediately, since it is no longer necessary to identify examinees at the TA station prior to examinees commencing testing. Third, all scoring will be done at the ET station. In HP-CAT, the final theta estimate was computed at the ET station, but all subsequent scoring was done by the TA station software. This change allows all psychometric routines to be part of one software component – the ET station software – making software modifications and the associated acceptance testing more straightforward.

There are four main software modules that make up the ET station software: (a) the keyboard familiarization sequence module, (b) the test instruction module, (c) the test administration module, and (d) the “Help” module. The ET station software allows some flexibility in test administration by reading certain information from files. For example, screen.dat is a file of all text dialogs and screens. Therefore, screen text can be changed without changes to the source code. Subtest.cat is a software configuration file for modifying administration of items. This file contains such information as the tests to administer, the order of test administration, the number of items in the test pool, the test length, test time, and the screen time-out limits. Et.cfg is a file that tells the ET station the type of computer (notebook or desktop) that is being used. All item information, such as item text and graphics, exposure control parameters, IRT parameters, and information tables, is external to the source code. Item text, graphics, and item parameters are stored in a database created using “Itemaker,” a program developed specifically for CAT-ASVAB. Exposure control parameters and information tables are stored in ASCII text files.

As with HP-CAT, PC-CAT automatically creates backups of applicant data files. If the system is operating in networking mode, applicant data are stored both on the hard drive of the File Server and the hard drive of the ET Station. If the system is operating in stand-alone mode, applicant data are stored on the hard drive and floppy drive of the ET Station. If the network fails during testing, each ET Station automatically switches to stand-alone mode, using the hard drive as the primary data depository and the floppy drive as the backup.

TA software. Unlike the ET station, the TA station for PC-CAT has been simplified at the functional level. As previously mentioned, the removal of the requirement for the random assignment of examinees to stations simplifies maintaining information on examinees and the availability of stations, as was necessary when designing the OT&E system. In fact, there is now no requirement for the TA station software to be concerned with where examinees are located in the testing room with respect to either test form or station availability. In addition, the immediate availability of either CAT-ASVAB test form at an ET station eliminates operator need to be concerned with where to place examinees when starting tests and, more importantly, in a failure recovery situation. In essence, any available station in the testing room may now be used to start a new examinee for testing, or to continue the testing session of an examinee whose station has failed.

The TA station functional specifications for the new system involve a number of requirements. Upon bootup, the software performs file maintenance activities and requests that the operator confirm the system clock time. The operator then selects the mode of operation for the testing session – network or stand-alone. At a MEPS, the network option will normally be selected; the stand-alone mode will be a failure recovery alternative. At a METS, only the stand-alone mode can be selected as the computers will not be electronically tied together as a "networked" configuration. The operator then enters TA and test session identifying information. Subsequently, the main screen is displayed. This screen allows you to monitor examinee progress and perform all necessary test administrator functions for the session.

About two-thirds of the main screen is used to display the status of examinees in the test session. There are nine data fields:

1. The SSN data field displays applicants' social security numbers. This information is transferred from the ET stations to the TA station. The word "Available" indicates that an applicant is not assigned to the test station. This data field also displays the number of stations and peripherals in the network that are not booted up. These stations and peripheals are referred to as "off-line."

2. The NAME data field displays applicants' last names.

3. The STATION ID data field displays ET station identifying numbers.

4. The FORM/TYPE data field displays the test form and test type assigned to the applicant. "I" indicates an initial test type; "R" indicates a retest; "C" indicates a confirmation test.

5. The TOTAL TIME data field displays the amount of time the applicant has been taking CAT-ASVAB.

6. The SUBTEST data field displays the abbreviated name of the test on which the applicant is currently working. It also indicates when an applicant needs help by displaying the word "HELP."

7. The TEST TIME data field displays the amount of time the applicant has been taking the test.

8. The END TIME data field displays an estimate of when the applicant will complete CAT-ASVAB. The estimate is in hours and minutes, with an error factor that becomes smaller with each test.

9. The STATUS/AFQT data field displays letters representing the testing status of each applicant and applicants' AFQT test scores upon completion of testing. At the start of testing, the field is dash-filled. Each dash represents a single step in the testing progress of the applicant. When an applicant's name has been submitted, the first dash becomes an "S", for already submitted. When the examinee completes testing, the second dash becomes a "C." When an applicant's results are transferred to MIRS, the third dash becomes an "R." When an applicant's unverified score report (described in Rafacz, B. & Hetter, R. D., 1997) has been printed, the fourth dash becomes a "P." If all processing steps are complete, the last dash becomes a "D." If the network detects that the applicant's machine has failed, the last dash becomes an "F." The system automatically performs all of these functions, except SUBMIT.

The arrow keys can be used to move up and down in the list of applicants and to select an applicant for editing of the applicant's identifying information, printing a report, or other available functions.

At the very bottom of the main screen, an electronic banner displays various testing activities. If a computer fails, the banner displays a message telling the test administrator that a station failed, giving the station's identifying information. If an applicant is in "HELP," a message is also displayed. Although this information is contained in the data fields described above, the moving banner is more likely to draw the TA's attention.

Immediately above the banner is a list of all available functions. To select a function, the TA presses the key associated with the function.

1. "N" sorts applicants by name.

2. "S" sorts applicants by SSN.

3. "T" sorts applicants by ET Station ID number.

4. "M" switches between modes of display. There are two modes of display: Session mode and Current mode. Session mode, which is the default mode, displays all applicants who have tested during the session. Current mode displays only those applicants currently testing. The mode that the TA software is in

is displayed at the top, right-hand side of the main screen.

5. “P” provides the test administrator with options to reprint the unverified score reports or the Aptitude Testing Processing List (ATPL), or to print a test session status report. (When an individual applicant completes testing, the unverified score report is automatically printed. When all applicants have completed testing, the TA station automatically prints the ATPL, a standard USMEPCOM form that includes such information as the examinee's last name, SSN, test form administered, Service processing for, sex, AFQT score, and test type.) The reprint option is available in case another copy of these reports is needed.

6. “D” allows the TA to collect applicant test data with a diskette rather than having CAT-ASVAB automatically download the data from the ET Stations to the File Server Station. The only time the TA will use this option is when the network fails during testing.

7. “E” allows the TA to electronically send applicant test data files to MIRS. (MIRS, in turn, sends the data to a central repository at USMEPCOM.) If the connection between the CAT-ASVAB system and MIRS is not functional, the data are automatically written to a floppy disk so they can be “manually” transferred to MIRS.

8. "INS" (the Insert key) allows you to add applicants to the test session at the TA station. This option is functional only in Stand-Alone mode.

9. “ESC” ends the test session.

In summary, the functional capability of the TA station emulates that of the HP-CAT system, but at both a simpler and more encompassing level. The TA station user-interface for PC-CAT is significantly different from that of HP-CAT. HP-CAT required the user to go through a number of menus to perform functions. Until the user became very familiar with the system, he or she could easily “get lost,” not knowing how to get to a certain menu or where to locate certain functions. With PC-CAT, once the user has “logged into” the TA station, everything is on one screen. In addition, all functions that could be automated have been, requiring less computer-user interaction.

Summary

In 1996, USMEPCOM procured the computer hardware for nationwide implementation. When the hardware specifications were written, the procurement was expected to take place in the 1994/95 time frame. While the CAT-ASVAB hardware requirements did not change between 1994 and 1996, what was available on the market did. As a result, the system that was actually procured exceeds some of the specifications. Most notable, the CPU is a Pentium, running at 100 MHZ, with eight megabytes of RAM and a 630 megabyte hard drive. As with the hardware, the networking software has also been upgraded. While initially programmed to run under NetWare 3.1, CAT-ASVAB now runs under NetWare 4.1.

The PC-CAT system is a streamlined, up-to-date version of HP-CAT. This new system is a cost-effective system that allows for ease in operating CAT-ASVAB and in maintaining the CAT-ASVAB software and equipment. There are several main advantages of the PC-CAT system over the HP-CAT system. First, there have been many advances in computer technology since 1985 when the HP-CAT system was selected. Notebook computers are now available that are much smaller, lighter, and more capable than computers available in 1985. Second, prices of computers in general have come down drastically, making both powerful notebooks and desktops available at relatively low cost. Third, the additional computer resources, and a better understanding of the operational requirements, have given designers an opportunity to make the system more efficient.

References

Folchi, J. S. (1986). Communication of computerized adaptive testing results in support of ACAP. *Proceedings of the Annual Conference of the Military Testing Association, 28* 618-623. New London, CT: U.S. Coast Guard Academy. (NTIS No. AD-A226 551)

Rafacz, B. A. (1994). *The design and development of a computer network system to support the CAT-ASVAB program*. San Diego, CA: Navy Personnel Research and Development Center.

Rafacz, B. A., & 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.

Chapter 12

THE PSYCHOMETRIC COMPARABILITY OF COMPUTER HARDWARE

An important issue in the development and maintenance of a computerized adaptive test concerns the comparability of scores obtained from different computer hardware. Previous studies (Divgi & Stoloff, 1986; Spray, Ackerman, Reckase, & Carlson, 1989) have shown that medium of administration (computer versus paper-and-pencil) can affect item functioning. It is conceivable that differences among computer hardware (monitor size and resolution, keyboard layout, physical dimensions, etc.) can also influence item functioning. For example, particular monitor characteristics may influence the clarity and accuracy of graphics items. Variations in clarity and accuracy among monitors may, in turn, affect examinee's performance on particular items. If this effect is sufficiently large, then variation in hardware components can affect three important psychometric properties of the test, including (a) the score scale, (b) precision, and (c) construct validity.

An example of ***score scale*** effects is provided by small low-resolution monitors which might make intricate graphics items difficult to interpret, increasing their difficulty. This effect would lower the mean of the observed scores for this monitor type, and perhaps affect higher order moments of the observed test score distribution as well. If variation among hardware affects the observed score distribution, then separate equatings would be required to place scores obtained from different hardware on a common score scale. The data required to estimate these adjustments however may be costly, since samples of 2,500 examinees may be required for each hardware configuration to perform an adequate equipercentile equating.

A large hardware effect can in addition influence the ***precision*** of the estimated scores. For example, the use of low-resolution monitors may increase the difficulty of particular graphics items, while having no effect on the difficulty of other non-graphics items. This mis-specification of the difficulty parameters of some (but not all) items is likely to introduce both systematic and non-systematic errors in the estimated abilities. If a particular hardware configuration increased the difficulty of some items, we would expect the mean of the estimated abilities to decrease by some amount. If this increase in difficulty is not uniform across items, however, we would expect a random error component to be introduced as well, lowering the precision of the estimated abilities. Poor resolution monitors (for example) may also lower the item's discrimination level, which in turn would affect the precision of the estimated abilities. The introduction of random error is perhaps somewhat more serious than the introduction of systematic error, since no monotonic score scale transformation can equate test reliabilities.

A large hardware effect can also alter the ***construct validity*** of the test or battery. For example, individual differences in visual acuity may affect scores obtained from poor resolution monitors. Those examinees with poor or average eyesight may be at a disadvantage relative to those with above average acuity for answering some graphics items. In this event, the constructs measured by some graphics tests (e.g., Mechanical Comprehension [MC]) may actually be influenced by the accuracy and resolution of the monitor. For low-resolution monitors these tests would measure a combination of visual acuity and mechanical knowledge; for high-quality monitors, these tests would measure only mechanical knowledge. Consequently, it is instructive to examine the affect of hardware characteristics on the constructs measured by the tests. These effects can be examined through an evaluation of construct validity (i.e., test intercorrelations).

There is some evidence to suggest that speeded tests contained in the ASVAB (Coding Speed [CS] and Numerical Operations [NO]) may be especially sensitive to small changes in test presentation format, more so than the adaptive power tests. In paper-and-pencil (P&P) presentation of these tests, the shape of the bubble on the answer sheet has been found to have a significant effect on the moments of number-right scores (Bloxom et al., 1993, Ree & Wegner, 1990). Since speed is a significant component of these tests, larger bubbles require more time to fill and thus produce lower scores on average. In these studies, no answer-sheet effect was found for power tests.

Although previous work on speeded tests (which focused on effects of P&P presentation forms) may not be directly transferable to the study of computer-administered speeded tests, this work suggests that different hardware effects may exist for computer-administered power and speed tests. Characteristics of input devices, for example, which affect the speed of input are likely to affect speed-test scores. It is unclear however that power tests would be similarly affected, since these scores are based primarily on response accuracy and are only indirectly affected by response latency.

The study reported here examines the effects of particular hardware characteristics on psychometric properties of the CAT-ASVAB. The objective of this work is to provide some insight into the exchangeability of different hardware: whether machines of different makes and models can be used interchangeably, and which hardware characteristics must remain constant among testing platforms to ensure adequate precision and score interpretation. The effects of several different hardware characteristics were examined on the score scale, precision, and construct validity of CAT-ASVAB test scores.

Method

A total of 3,062 subjects recruited from the San Diego area participated in the study. Subjects were recruited from local colleges and universities, high schools, trade schools, and employment training programs and were paid $40.00 for approximately 3.5 hours of testing. Subjects consisted of 17–23 year olds responding to advertisements in local, college, and high school newspapers.

Procedures

All subjects were scheduled for a session date and time (either morning or afternoon) prior to the day of testing. For each session, examinees were processed in the order in which they arrived. Upon arrival, TAs inspected photo identification to verify subjects' identities and ages. Each subject was asked to read and sign a consent form which provided (a) background information on the ASVAB, and (b) agreement by the subject to participate in the research study. The consent form also informed subjects that as part of the study, they would take a computerized test which takes approximately three-and-a-half hours to complete, would take the test to the best of their ability, and would receive a check for $40.00 at the conclusion of the test.

After signing the Consent Form, each subject was randomly assigned to one of 28 computers. (This assignment was performed using random assignment sheets which contained a pseudorandom permutation of integers from 1 to 28. The first examinee seated was assigned to the first station listed on the sheet; the second examinee seated was assigned to the second station on the sheet, etc. A different sheet [containing a different random permutation] was used for each test session. This assignment resulted in roughly equal proportions of subjects assigned to each of the 28 computer stations.) As described below, each of the 28 computers belonged to one of 13 experimental conditions.

Experimental Conditions

Thirteen experimental conditions were defined by specific combinations of computer hardware and test presentation format. These are displayed in Table 12-1. Column abbreviations along the top row of the table denote the following:

1. STA Computer station number (from 1–28)

2. CT Computer type

A Panasonic notebook (386 CPU); monochrome LCD
B Dell subnotebook (386 CPU); monochrome LCD
C Texas Instruments (486 CPU); monochrome LCD
D Toshiba (486 CPU); active color matrix display
E Dell desktop (486 CPU); monochrome VGA monitor
F Datel (486 CPU)

3. MNF Manufacturer

Pans Panasonic

Dell Dell Microsystems

TI Texas Instruments

Tosh Toshiba

Datl Datel

4. **Type** Computer Type

D Desktop

N Notebook

S Subnotebook

5. **Monitor** Computer Monitor

Mono Monochrome (VGA)

Color-HC Color (High Contrast—White letters with blue background)

Color-LC Color (Low Contrast—Purple letters with blue background)

6. **COND** Condition (from 1–13) denoting how data from the 28 stations are combined for analyses

7. **Input** Input device

Full Full keyboard where labels “A–B–C–D–E” were placed over the “S–F–H–K–**:**” keys, respectively. The space bar was labeled “ENTER,” and the “F1” key was relabeled “HELP.” All other keys were covered with blank labels.

Pad Key-pad—17 keys (either **G**: Genovation, or **D**: Dell) where labels “HELP–A–B–C–D–E” were placed over the “ - –7–9–5–1–3” keys, respectively.

Tmp Template, where all keys except the “F1,” “spacebar,” and “S–F–H–K– **:**” keys were removed from the full keyboard. A flat piece of plastic with rectangular holes (for the 7 remaining keys) was placed over keyboard. The “F1” and “spacebar” keys were relabeled “HELP” and “ENTER,” respectively. The “S–F–H–K– **:** ” were re-labeled “A–B–C– D–E,” respectively.

8. Order First form administered: Each examinee received both forms of the CAT-ASVAB, with indicated form (C1 or C2) administered first.

Table 12-1. Experimental Conditions

STA[1]	CT[2]	MNF[3]	Type[4]	Monitor[5]	COND[6]	Input[7]	Order[8]
1	A	Pans	N	Mono	1	Pad-G	C1
2							C2
3					2	Full	C1
4							C2
5							C1
6							C2
7	B	Dell	S	Mono	3	Full	C1
8							C2
9					4	Pad-D	C1
10							C2
11	C	TI	N	Mono	5	Pad-G	C1
12							C2
13					6	Tmp	C1
14							C2
15					7	Full	C1
16							C2
17	D	Tosh	N	Color-HC	8	Full	C1
18							C2
19	E	Dell	D	Mono	9	Pad-G	C1
20							C2
21	F	Datl	D	Mono	10	Pad-G	C1
22							C2
23				Color-HC	11	Full	C1
24							C2
25				Color-LC	12	Full	C1
26							C2
27				Color-HC	13	Pad-G	C1
28							C2

Hardware Dimensions

The 13 experimental conditions were constructed to examine five issues related to the effects of particular hardware characteristics on the measurement properties of observed test scores. Using the design outlined above, each of these questions can be addressed by contrasting selected conditions in which all hardware characteristics remained constant, except for the characteristic of interest. A sixth set of conditions was added to address the similarity of scores obtained from different hardware configurations which employ a common input device. The six research questions and associated conditions are provided below.

Input device. Do differences in input devices used by examinees to enter responses affect scores? This can be addressed by a comparison of Conditions 5–6–7, which used the ‘keypad,’ ‘full keyboard,’ and ‘template’ input devices, respectively.

Color scheme. Does the use of different background and foreground colors affect scores? This can be addressed by a comparison of Conditions 11 and 12. Condition 11 presented questions using white letters (foreground) with a blue background (denoted as high-contrast). Condition 12 used purple letters presented on a blue background. In this latter condition (denoted as low-contrast), the contrast between the foreground and background was greatly reduced due to the similarity of colors.

Monitor. Do differences in monitor types (color or monochrome) affect scores? This issue can be examined by contrasting Conditions 10 and 13, which used monochrome and color monitors, respectively.

CPU. Do differences in CPU (make or model) affect scores? This question can be addressed by a comparison of Conditions 9 and 10, which used CPUs from different manufacturers.

Portability. Do differences in portability affect scores? This issue can be addressed by a comparison of Conditions 1–4–9 (Notebook–Subnotebook–Desktop), Conditions 2–3–7 (Notebook–Subnotebook–Notebook), and Conditions 8–11 (Notebook – Desktop). Note that the same input device was used within each of these three subsets.

Input device invariance. Can similar scores be obtained from different hardware configurations using the same input device? This contrast (which contrasts Conditions 1, 4, 5, 9, 10, 13) anticipates that differences (where they exist) might be caused primarily by the input device. This may be especially true for speeded tests. By holding input device constant across different hardware configurations, the remaining differences (if any) can be assessed.

Instruments

All subjects participating in the study were administered both forms (C1 and C2) of the CAT-ASVAB (Segall, Moreno, & Hetter, 1997). Dependent measures consisted of the 22 (11 tests × 2 forms) scores. For the 18 adaptive power tests, these scores were based on Item Response Theory (IRT) ability estimates and were set equal to the mode of the posterior distribution. The four speeded tests were scored using chance corrected rate scores. Scoring details are provided in Segall, Moreno, Bloxom, & Hetter, 1997.

The software that administers the CATASVAB runs under the MS-DOS operating system, requires 4 megabytes of RAM and requires a VGA video card and monitor. The same software was used in all conditions, with only minor modifications required to accommodate differences in input devices.

Analyses and Results

Under the null hypothesis of no hardware effects, the 22 test variables should display equivalent first, second, and cross moments among the 13 experimental conditions. Stated more formally, under the null hypothesis

$$\mu_1 = \mu_2 = \ldots = \mu_{13} \tag{12-1}$$

and

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_{13} \tag{12-2}$$

where μ_k is a 22-element vector containing the test means for the *k*-th condition, and Σ_k is the 22 × 22 covariance matrix for the *k*-th condition. Taken jointly, the parameters $\{\mu_k , \Sigma_k\}$ (for $k = 1, 2, ..., 13$) contain useful information about hardware effects on the score scale, reliability, and construct validity of the battery. This becomes evident by noting that common measures of these properties are functions of these parameters. Score scale effects can be assessed from a comparison of means and variances across conditions; reliability effects can be examined from a comparison of alternate form reliabilities (across conditions); and construct validity effects can be measured from a comparison of test intercorrelations, or from a comparison of disattenuated test intercorrelations. Since all these measures are functions of elements contained in $\{\mu_k , \Sigma_k\}$, the statistical significance of the hardware effects (on the score scale, reliability, and construct validity) can be tested directly from 12-1 and 12-2. That is, if 12-1 and 12-2 hold, then so does the equivalence of score scale, reliability, and construct validity across conditions. This is noteworthy, since standard significance tests exist for testing

12-1 and 12-2. Below, the equivalence of the means and covariance matrices are tested separately. Where differences were found, additional analyses were conducted to help isolate the hardware related cause.

Homogeneity of Covariance Matrices

The likelihood ratio statistic

$$\lambda = \frac{\prod_{k=1}^{13} \left|\hat{\Sigma}_k\right|^{n_k/2}}{\left|\hat{\Sigma}\right|^{N/2}} \qquad (12\text{-}3)$$

was used to test the significance of the difference among the 13 covariance matrices, where $\hat{\Sigma}_k$ is ML estimate of the 22 × 22 covariance matrix for the k-th group, $\hat{\Sigma}$ is the estimated covariance matrix for the total group, n_k is the sample size of the k-th group, and $N = \sum_{k=1}^{13} n_k$ is the total sample size. Under the assumption that the observations were sampled from a normal distribution, $-2 \log \lambda$ is asymptotically chi-square distributed with $df = 3{,}036$. However, in the current application of the test statistic, the asymptotic distribution of λ may not hold since most groups had relatively small sample sizes. For testing the significance of the difference among covariance matrices, the distribution of λ was approximated by a bootstrap method. This was accomplished using the following procedure:

1. Compute the statistic given by Equation (12-3) and denote the statistic value as λ_0.

2. Compute $\mathbf{x}_j$ (j = 1, ..., N), where $\mathbf{x}_j$ is the 22-element vector of difference scores calculated from the difference between the raw observations and the respective group mean vector.

3. Sample N observations ($\mathbf{x}_j$'s) with replacement.

4. Divide the N sampled values into 13 groups of sizes $n_{1,}$, n_2, ..., n_{13}.

5. Compute the 13 covariance matrices from the set of bootstrapped values.

6. Compute the λ statistic given by Equation (12-3) from the bootstrapped covariance matrices.

7. Perform 10,000 replications of Steps 3–6, computing λ_1, λ_2, ..., λ_{10000}.

8. Compute prob($\lambda > \lambda_0$), the proportion of λ values greater than the sample value λ_0. If this proportion is small, we reject the null hypothesis of equivalent covariance matrices.

The bootstrap procedures outlined above resulted in an estimated prob($\lambda > \lambda_0$) = .4785, which leads us to accept the null hypothesis of equivalent covariance matrices. Thus, there appears to be no effect of hardware on the reliability, construct validity, or on the variance of the score scale. Effects of hardware on the score-scale location parameters (means) are examined below.

Homogeneity of Means

To test the equivalence of means across the 13 hardware configurations, separate one-way ANOVAs were computed for each of the 11 tests contained in CAT-ASVAB. The dependent measure in each analysis was the average of the two scores obtained from like-named tests of forms C1 and C2. The results and summary statistics for the nine adaptive power tests are provided in Table 12-2. As indicated, none of the power tests displayed significant mean differences.

Table 12-2. ANOVA Results and Summary Statistics (Power Tests)

			Means (*m*) and SD (*s*)								
Condition	**N**	**Statistic**	**GS**	**AR**	**WK**	**PC**	**AI**	**SI**	**MK**	**MC**	**EI**
1	210	*m*	.34	.27	.41	.02	–.71	–.62	.65	–.52	–.47
		s	.88	.96	.84	.91	.71	.77	.97	.93	.91
2	433	*m*	.28	.12	.28	–.02	–.75	–.77	.55	–.59	–.41
		s	.92	1.00	.90	.94	.74	.81	1.04	.93	.94
3	228	*m*	.27	.12	.27	–.03	–.80	–.72	.55	–.52	–.41
		s	.96	1.03	.92	.97	.71	.80	1.03	.89	.91
4	210	*m*	.32	.26	.33	.05	–.79	–.76	.71	–.52	–.44
		s	1.03	.99	1.01	.97	.74	.78	.92	.85	.95
5	228	*m*	.31	.22	.32	.05	–.69	–.68	.60	–.49	–.35
		s	.91	.97	.86	.89	.69	.76	.98	.89	.86
6	222	*m*	.33	.22	.33	.01	–.77	–.70	.65	–.57	–.39
		s	.85	.95	.91	.96	.74	.74	.92	.86	.89
7	218	*m*	.24	.18	.25	.00	–.72	–.73	.59	–.61	–.45
		s	.93	1.00	.87	.91	.71	.78	.94	.84	.88
8	224	*m*	.28	.29	.31	–.02	–.82	–.78	.60	–.57	–.48
		s	.87	1.00	.87	.97	.67	.74	1.00	.86	.92
9	217	*m*	.24	.08	.26	–.05	–.73	–.78	.56	–.66	–.43
		s	.88	.89	.91	.94	.69	.78	.96	.83	.90
10	218	*m*	.28	.23	.32	.04	–.81	–.75	.62	–.57	–.45
		s	.96	.92	.94	.94	.71	.77	.92	.79	.92
11	225	*m*	.32	.24	.34	.02	–.71	–.66	.56	–.52	–.35
		s	.95	.94	.91	1.02	.73	.78	1.00	.90	.94
12	217	*m*	.28	.24	.27	–.01	–.68	–.70	.63	–.46	–.35
		s	.94	.91	.88	.91	.76	.78	.98	.91	.92
13	213	*m*	.27	.23	.24	–.05	–.80	–.75	.64	–.57	–.52
		s	.94	.94	.90	.96	.65	.76	.98	.85	.92
ANOVA		*F* value	.27	1.12	.53	.31	1.01	.91	.55	.85	.75
		P value	.99	.34	.89	.99	.44	.54	.88	.60	.70

Table 12-3 displays results for the two speeded tests. For each test, three scores were examined:

Rate the proportion correct (corrected for chance guessing) divided by the mean response time,

RT the average response latency (seconds) computed from the answered (reached) items, and

P the proportion of correctly answered items calculated from reached items only.

The dependent measure was the average of these variables across the two CAT-ASVAB forms. As indicated in Table 12-3, significant mean differences for response time (RT), accuracy (P), and rate were found for NO. For CS, significant and marginally significant differences were found for response time (RT) and rate, respectively. Additional comparisons were made among speeded test rate-score means (Rate) to help relate the significant findings to specific hardware characteristics.

Table 12-3. ANOVA Results and Summary Statistics (Speeded Tests)

			Means (*m*) and SD (*s*)					
			Numerical Operations			Coding Speed		
Cond	***N***	**Statistic**	**Rate**	**RT**	**P**	**Rate**	**RT**	**P**
1	210	*m*	21.64	2.83	.93	10.33	5.28	.89
		s	5.43	.76	.07	3.20	1.46	.15
2	433	*m*	21.83	2.89	.94	10.27	5.38	.89
		s	5.91	.82	.06	3.18	1.39	.16
3	228	*m*	21.09	2.97	.94	9.81	5.54	.88
		s	5.38	.82	.06	3.49	1.39	.17
4	210	*m*	19.50	3.10	.91	9.88	5.40	.89
		s	5.33	.79	.09	3.05	1.41	.17
5	228	*m*	21.63	2.85	.94	10.37	5.49	.92
		s	4.90	.66	.06	3.01	1.33	.13
6	222	*m*	23.78	2.66	.94	10.91	5.14	.90
		s	6.29	.76	.05	3.16	1.31	.13
7	218	*m*	22.74	2.79	.94	10.72	5.19	.90
		s	6.23	.83	.06	3.12	1.43	.14
8	224	*m*	21.66	2.87	.94	9.73	5.43	.87
		s	5.62	.75	.07	3.59	1.38	.18
9	217	*m*	21.39	2.90	.93	10.36	5.40	.90
		s	5.47	.84	.07	3.09	1.45	.14
10	218	*m*	21.47	2.91	.93	10.21	5.35	.89
		s	5.30	.87	.07	2.97	1.36	.16
11	225	*m*	22.07	2.84	.93	10.47	5.30	.90
		s	6.40	.75	.08	3.30	1.41	.15
12	217	*m*	21.39	2.94	.94	10.08	5.52	.90
		s	5.65	.78	.05	3.00	1.45	.14
13	213	*m*	21.52	2.86	.93	10.27	5.29	.89
		s	5.38	.68	.07	3.31	1.33	.16
ANOVA		*F* value	6.22	3.67	2.94	2.50	1.68	1.28
		P value	.00	.00	.00	.00	.06	.22

Table 12-4 displays ANOVA results for the six research issues. The second column displays those conditions included in each ANOVA. The results for NO (columns 3 and 4) indicate significant effects for “input device,” portability,” and “input-device invariance.” Note, however, most significant effects can be attributed to the Dell subnotebook used in Conditions 3 and 4 (full keyboard and keypad conditions, respectively). An inspection of the means for Condition 3 and 4 (Table 12-1) indicates that this computer provides the lowest rate scores among all 13 conditions. This may have been due to the monitor which consisted of a liquid-quartz display. As indicated in the bottom row of Table 12-4, by excluding the Dell subnotebook Condition, non-significant mean differences were found when the same input-device (key pad) was used across remaining notebook and desktops computers (Conditions 1, 5, 9, 10, and 13).

The results for CS also display significant effects for "portability." However unlike NO, no effect of input device is observed, and the portability effect does not appear to be directly related to the Dell subnotebook computer. Some characteristic difference between desktop and notebook computers (other than input device) appears to affect mean rate scores on CS. Because of the inconsistency of these results, it is difficult to attribute the exact cause of the difference to a specific hardware characteristic.

Table 12-4. ANOVA for Selected Comparisons (Speeded Tests)

		Numerical Operations		**Coding Speed**	
Factor	**Conditions**	**F value**	**P value**	**F value**	**P value**
A. Input Device	5,6,7	7.71	.001**	1.76	.173
B. Color Scheme	11,12	1.38	.240	1.75	.187
C. Monitor	10,13	.01	.928	.04	.841
D. CPU	9,10	.02	.881	.27	.602
E. Portability	1,4,9	9.91	.001**	1.59	.205
	2,3,7	4.41	.012*	4.40	.013*
	8,11	.51	.477	5.30	.022*
F. Input Device Invariance	1,4,5,9,10,13	5.25	.001**	.76	.577
	1,5,9,10,13	.08	.987	.10	.981

** p < .05; ** p < .001*

Discussion

Among the five hardware dimensions examined, none were found to affect the psychometric properties of the adaptive power tests contained in the CAT-ASVAB. This result is noteworthy, since it suggests that some future changes in input device, color scheme, monitor, CPU, and portability may not necessarily lead to changes in reliability, construct validity, or the score scale of the adaptive power tests. Thus some variation in hardware may be permissible without the need for separate power test equating transformations.

However, some effects on rate scores were observed for the two speeded tests. For NO, these significant effects appeared to be caused by differential effects of hardware on both response latency and accuracy. Furthermore, scale location of the rate score was influenced by the type of input device. Some input devices appeared to allow for faster responding, which resulted in higher rate scores. When the same input device was on desktop and notebook computers, no differences in psychometric score properties were identified. For CS, “portability” effects were identified—causing differences in scale location between desktop and notebook computers. Although the difference appears to be related to response speed rather than to response accuracy, it is difficult to attribute the exact cause of the difference to a specific hardware characteristic.

Although the results suggest that computer-administered power tests are insensitive to hardware changes, prudence should be exercised when altering any characteristic of an existing test with an established score scale, or when considering the exchangeability of scores obtained from different hardware configurations. This caution grows out of experiences with paper-and-pencil tests, where seemingly trivial differences, such as differences in line length or spacing can have a related effect on observed score distributions. When considering variation in hardware among computer administered tests, it may be useful to consider the following two factors.

1. ***To what extent is the test speeded?*** To the extent that speed influences test scores, hardware is likely to have an increasing effect on the score scale. Among the 11 tests studied here, there was a clear demarcation between power and speed. Although each of the nine power tests had an associated time limit, these time limits typically allow (in a military applicant population) over 98 percent of all examinees to complete all questions. Thus, any small differences in response times caused by different hardware are unlikely to result in an increase in the frequency of unanswered items. Conversely, for the two speeded tests, scores are determined by dividing the percent correct by the item latencies. For these tests, it is very obvious how different hardware may cause different response times. However, the issue becomes more complicated when changes are being considered for power tests that have completion rates somewhere between the two extremes, say 90 percent. If the power test is sufficiently speeded, it is conceivable that latency-related hardware changes may increase the numbers of incomplete tests by a large enough amount to significantly alter the score scale.

2. ***To what extent is the item appearance dependent on the hardware?*** In the current study, the item appearance on different computers was almost identical. The same software was used to administer the adaptive tests on different computers. In each condition, VGA monitors were used. Although both text and graphics were presented, the position and relative dimensions of all text and graphics remained relatively constant across conditions. The software presented text using a standard DOS fixed-width font, which resulted in identical line breaks and spacing across conditions. Variations involving more extensive alterations in appearance (i.e., changes in font and line breaks) may have larger effects than the ones identified in this study.

Although the adaptive power test results are encouraging, caution should be exercised when generalizing these results to other tests and other hardware configurations. Some meaningful (but small) effects may have been present but were not detected because of insufficient power. In some instances, small changes in the score scale can have important consequences for selection decisions. The samples used in this study may not have been large enough to detect small but important effects caused by different hardware. A useful and important follow-on study would (a) consist of a small number of conditions (say, one desktop and one notebook condition), and (b) employ large samples (say, 2,500 subjects per condition). If present, such a study could detect these small but important effects of hardware on the score scale. If this future, large sample study replicates the current findings, then added confidence can be given to the hardware-invariance property attributed to adaptive power tests.

References

Bloxom, B. M., McCully, R., Branch, R., Waters, B. K., Barnes, J. D., & Gribben, M. R. (1993). *Operational calibration of the circular-response optical-mark-reader answer sheets for the ASVAB* (Report 93-009). Monterey, CA: Defense Manpower Data Center.

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves* (Report 86-24). Alexandria, VA: Center for Naval Analyses. (NTIS No. AD-B1-3 889)

Ree, M. J ., & Wegner, T. G. (1990). Correcting differences in answer sheets for the 1980 Armed Services Vocational Aptitude Battery reference population. *Military Psychology, 2,* 157-169.

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.

Spray, J. A., Ackerman, T. A., Reckase, M. D. & Carlson, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261-271.