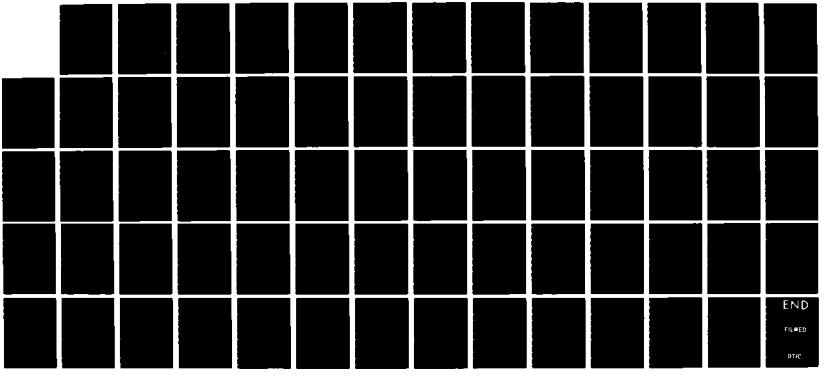
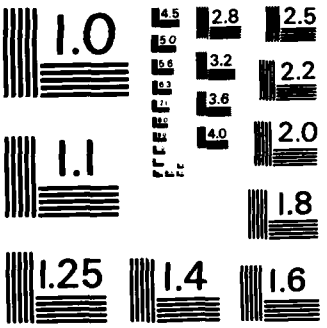
AD-A160 712 A PLAN FOR SCALING THE COMPUTERIZED ADAPTIVE ASVAB (VOCATIONAL APTITUDE B..(U) JOHNS HOPKINS UNIV BALTIMORE MD DEPT OF PHYSIOLOGY B F GREEN ET AL. 1/1

UNCLASSIFIED 30 NOV 83 ARI-RR-83-1 N00014-80-K-0304 F/G 5/10 NL





(2)

A Plan for Scaling the Computerized Adaptive ASVAB

Bert F. Green
R. Darrell Bock
Robert L. Linn
Frederic M. Lord
Mark D. Reckase

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 83-1	2. GOVT ACCESSION NO. AD-A160712	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Plan for Scaling the Computerized Adaptive ASVAB		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Bert F. Green R. Darrell Bock Robert L. Linn Frederic M. Lord Mark D. Reckase		8. CONTRACT OR GRANT NUMBER(s) N00014-80-K-304
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology The Johns Hopkins University Baltimore, MD 21218		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE: 63707N-NPRDC; 61153N(42) RR04204 TA: RR0420401 NR: 150-463-1; 150-463-2
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research (Code 442-PT) Arlington, VA 22217		12. REPORT DATE 30 November 1983 13. NUMBER OF PAGES 60
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This study was supported by funds from the Navy Personnel Research and Development Center, and the Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Latent trait theory, item response theory, tailored testing, adaptive testing, item characteristic curve theory, evaluation, ASVAB, computerized testing, computerized adaptive testing, ability testing, vocational aptitude battery, test scaling, test equating, test calibration.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A plan is offered for scaling the computerized adaptive test (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). The scaling will produce CAT scores that are comparable to the current paper & pencil ASVAB scores. The plan provides for scaling of a provisional CAT, Form 99, using data from field tests of prototype equipment, to be followed by scaling of final Form 100 on operational equipment. Each calibration includes recalibration of test items in the CAT environment, and scaling of revised CAT scores. Procedures are stated for composites, including AFQT, VE, and occupation speciality composites. IOT&E procedures are outlined.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A Plan for Scaling the Computerized Adaptive ASVAB

Bert F. Green
R. Darrell Bock
Robert L. Linn
Frederic M. Lord
Mark D. Reckase

EXECUTIVE SUMMARY

The U.S. Armed Services plan to use computerized adaptive testing (CAT) for the Armed Services Vocational Aptitude Battery (ASVAB), taken by all applicants for military service.

Computer test presentation has many advantages; test administration is more efficient and secure, because no printed tests are stored, and no answer sheets are needed. Adaptive testing achieves additional efficiencies by asking questions of a difficulty level appropriate for each candidate, as determined by his answers to earlier questions. In effect, a correct answer is followed by a more difficult question; an incorrect answer leads to an easier question. Each CAT consists of a large pool of possible questions (called items). Each item is characterized by three parameters - its precision, its difficulty, and its chance of being answered correctly by very low-ability candidates. A new test theory, more elaborate and precise than classical theory, manages the adaptive process.

The scores of the new CAT version of the ASVAB must be scaled so that they will be as nearly equivalent as possible with the scores on the paper and pencil (P&P) version of the ASVAB. Scaling, also called equating or norming, is regularly used when new P&P versions of the ASVAB are introduced. Converting from conventional paper-and-pencil (P&P) test administration to CAT administration poses two kinds of equating problems. First, the calibration of the CAT items must be verified, and recalibrated if necessary, because their parameters will have been obtained in the P&P environment. Second, the scores from the CAT tests, based on the recalibrated item parameters, must be scaled to be comparible to the scores from the corresponding P&P tests. This report recommends procedures for recalibrating item parameters in the CAT ASVAB item pools, and for equating scores on the CAT version of the tests with scores on the corresponding P&P ASVAB tests.

The plan for scaling the CAT version of the ASVAB is thorough and extensive, in an attempt to avoid an error like the misscaling of the ASVAB that happened in the late 1970's and was corrected in 1980. The report recommends a series of continued steps of data analysis and monitoring. In particular, various uncertainties indicate that any test scaling that is done before the operational equipment is in regular use will have to be readjusted after operational experience. The adjustments may be small but they may not be negligible, and cannot be ignored. Rather than making a series of adjustments to CAT during the early stages of implementation, a two-stage scaling process is proposed. First, a preliminary CAT battery would be calibrated and scaled on the prototype equipment, for early use in the operational phase. Second, a final CAT battery would be calibrated and scaled early in the operational phase, for use late in the first operational year.

The preliminary battery, here called Form 99, would use only a subset of items from the total operational item pool, and hence might not be completely equivalent to the eventual operational CAT, here called Form 100, with its full item pools. Also, the items in Form 99 would have been calibrated on prototype equipment rather than the operational equipment. Still, Form 99 would have been properly scaled to the P&P ASVAB, and would even serve as the link between the P&P ASVAB and Form 100.

The calibrations and scalings for the tests in Form 99 would be done using data collected during prototype field testing, so that the preliminary battery could be put in place as soon as CAT becomes operational. Form 99 would have some limitations, but it would be adequate for early use.

Calibration and scaling of Form 100, the eventual battery, would use data collected on the operational equipment during the initial months of actual CAT system implementation. This time period can also be considered a period of initial operational testing and evaluation (IOT&E) for Form 99, during which the equating of Form 99 can be checked. When Form 100 is in place, a final IOT&E period will be needed to check the equating of this full CAT battery.

When the CAT system is fully operational, it is anticipated that data for the calibration of new test items will be collected on-line by including trial items in the item bank with previously calibrated items. With these data, parameters of the new items can be estimated on the same scale as the operational items, so new items can be added without the need for further rescaling.

INTRODUCTION

The U.S. Armed Services are considering the introduction of computerized adaptive testing (CAT) into military accessions procedures. At present all applicants for military service take the Armed Services Vocational Aptitude Battery (ASVAB), a standard paper-and-pencil (P&P) test of cognitive abilities, skills and technical information. Testing efficiency can be greatly improved by using computer presentation of test questions (usually called items) and by choosing items for presentation to a candidate appropriate to his apparent skill level, as indicated by his response to previous items (Lord, 1980; Green, 1983; McBride, 1977).

Because this testing method is both highly efficient and relatively novel, careful evaluation is appropriate. A plan for evaluating the CAT version of the ASVAB has recently been prepared (Green et al, 1982.) That document also provides background discussion of the CAT technique, and the nature of the ASVAB. Other discussions of CAT can be found in Lord (1980), Urry (1981), and Weiss (1978, 1980, 1983.)

This report discusses the important question of how to scale the scores of the new CAT version of the ASVAB, so that as nearly as possible they will be comparable to the scores on the paper and pencil version of the ASVAB (PP-ASVAB). This process, called scaling, norming, or equating is regularly used when new P&P versions of the ASVAB are introduced. Special problems are encountered when changing to a very different method of item presentation.

The plan presented here includes a complex series of analyses that are designed to be cost-effective and yet to avoid the kind of serious scaling error that arose in scaling ASVAB Forms 5, 6, and 7. This error, which has been thoroughly documented elsewhere (Maier & Truss, 1983) resulted in accepting for service a large number of applicants who should properly not have been admitted, with the concomitant costs of extra training time, and increased attrition. This report describes a plan that tries to avoid the problems encountered at that time, as well as dealing with the additional problems inherent in the transition to a new form of test presentation.

Adaptive Testing

The principal idea of adaptive testing is simply that each test taker is asked questions that are appropriate for his or her level of skill or ability. It is inefficient to ask questions that are too easy or too difficult for the candidate, since those responses contribute very little information about that person's ability.

The method of adaptive testing has roots in early psychological measurement. Psychophysicists, beginning with Wundt, determined sensory thresholds by presenting stimuli at varying intensities according to the observer's ability to sense them. Binet, (1909), the originator of mental testing, asked each child questions appropriate to the child's age, and moved up or down the age scale depending on the child's answers. The process of choosing items appropriate to the child's mental ability can be viewed as adapting the test to the test-taker. Such a procedure is very difficult to manage if people are tested in groups rather than one at a time, so ordinary pencil-and-paper (P&P) tests present the same items to all test-takers. The items on group tests vary in difficulty over a range appropriate to the population being tested, so group tests are roughly matched to the population, but cannot be adapted to the individuals.

With a digital computer to present the test items, item-by-item adaptive testing becomes feasible. The computer can score each response immediately and can then select the next item that will be most appropriate for the candidate. Each candidate gets a set of items uniquely selected for him or her. More specifically, each person's first item has about medium difficulty for the total population. Those who answer correctly get a harder item; those who answer incorrectly get an easier item. After each response, the examinee's ability is estimated, along with an indication of the accuracy of the estimate. The next item to be posed is one that will be especially informative for a person of the estimated ability, which generally means an item of medium difficulty at that ability level. Normally, the process results in harder questions being posed after correct answers and easier questions after incorrect answers. The change in item difficulty from step to step is usually larger earlier in the sequence when less is known about candidate's ability; later in the sequence the difficulty changes less radically as the system tries to refine its estimate of the candidate's ability. The process continues, until there is enough information to place the person on the ability scale with a specified level of accuracy, or until some more pragmatic criterion is achieved. If desired, each candidate's score on a CAT can be estimated to the same level of accuracy. By contrast, high and low scores on a conventional P&P group test are typically less accurate than scores near the mean.

A CAT consists of a set of items, called an item pool or item bank, from which particular items are selected for presentation to the candidate. The precision of the CAT depends on the characteristics of the items in the pool. If the pool is not large enough, and is not well-matched to the ability distribution of the group being tested, the advantages of an adaptive test will not be fully realized. For example, if the adaptive procedure indicated that the next item for a particular person should be moderately easy, but there are no more moderately easy items, the system would have to settle for an item that is very easy, or for one that is moderately difficult, with the result that less information would be obtained than if an appropriate item had been available. Thus adaptive testing requires a sufficient supply of items at each ability level. If security considerations suggest that the items be varied, several alternative items are needed at each ability level; thus, large item pools are needed for adaptive tests.

Adaptive testing places new demands on psychometric test theory and method. Classical test theory is not adequate; methods appropriate for conventional P&P group tests will not work with adaptive tests. The most obvious problem is that the test score can no longer be the number of items answered correctly. In an ideal adaptive test, after the first few items, everyone will tend to answer about the same number of items correctly. The score must depend in some way on the characteristics of the items answered correctly.

Also the indices commonly used to judge the quality of the items are less appropriate. The ordinary index of item difficulty is the proportion of persons answering the item correctly, which is dependent on the population of test takers. Likewise, the ordinary indices of item discriminating power, such as the item-test correlation, are also dependent on the population.

Early work on adaptive testing is discussed in Harman, Helm & Loye (1968), Holtzman (1970), and Wood (1973). More recent accounts can be found in Weiss (1974, 1978, 1980), and Green (1983a,b). Applications have been discussed by Urry (1977), Lord (1977, 1980), and Kreitzberg & Jones (1980).

Item Response Theory

Classical test theory is not suited to adaptive tests. Classical theory supposes that all test-takers confront the same set of test items, as in the conventional P&P tests. Classical indices of reliability, validity, and item quality are relevant to a particular set of items and a particular population of test-takers. But an adaptive test presents different items to each taker, and is, in principal, independent of the particular population.

A theory that is appropriate for adaptive tests was developed by Rasch (1960), Lawley (1943), Tucker (1946), Lord (1952), Samejima (1969), Owen (1975), and others. This new theory, now called item response theory (IRT), was discussed by Birnbaum (1968) as latent trait theory[1] in Lord & Novick's (1968) major treatise on test theory. Hambleton & Cook (1977), and Warm (1978) give good introductions. More complete accounts of IRT have been given recently by Lord (1980), Urry & Dorans (1983), Urry (1981), and Hulen, Drasgow, and Parsons (1983).

The theory postulates that persons vary in the ability being assessed by the test, and that their abilities are distributed along a continuum labelled θ (theta) from low to high. The probability of answering an item correctly is assumed to vary with ability, starting at a low value for low-ability candidates and increasing as ability increases, up to certainty for persons of very high ability, as sketched in Fig. 1. In IRT the mathematical form of these curves is called the logistic curve, and the curves are called item characteristic curves (ICC), although some authors use the phrase item response function. Curves for different items vary in three respects: (a) the discriminating power, (b) the difficulty, and (c) the pseudo-chance level. These characteristics are represented mathematically by the parameters **a**, **b**, and **c** in the logistic equation.[2] In Fig. 1, Item 1 has a higher **a** than Item 2, because it is steeper and hence more discriminating. Item 1 has a lower **b** than Item 2, because Item 1's curve is to the left of Item 2 Item 1 is easier than Item 2 because its probability of being answered correctly is higher for many ability levels. Item 1 has a higher **c** than Item 2 because its probability of a correct answer is larger for very low ability levels.

---

[1] The term "latent trait theory" is used in the earlier literature, rather than "item response theory." "Latent" signifies that the ability or skill being assessed is inferred from the item responses, and is in this sense latent in the item responses; "trait" merely refers to a characteristic of the examinee that is sufficiently stable to be measured. However, some laypersons may interpret the terms "latent trait" in a non-technical sense as implying a fixed, inherited property of the individual not alterable by training. This interpretation is incorrect, and is in no way appropriate to tests of vocational skills and knowledge, so the neutral phrase "item response theory" is preferred.

[2] In the three-parameter logistic model, the probability of Person i with ability θ_i responding correctly to item j, becomes

$$P_j(\theta_i) = c_j + (1-c_j)/(1+\exp(-1.7a_j(\theta_i - b_j))).$$

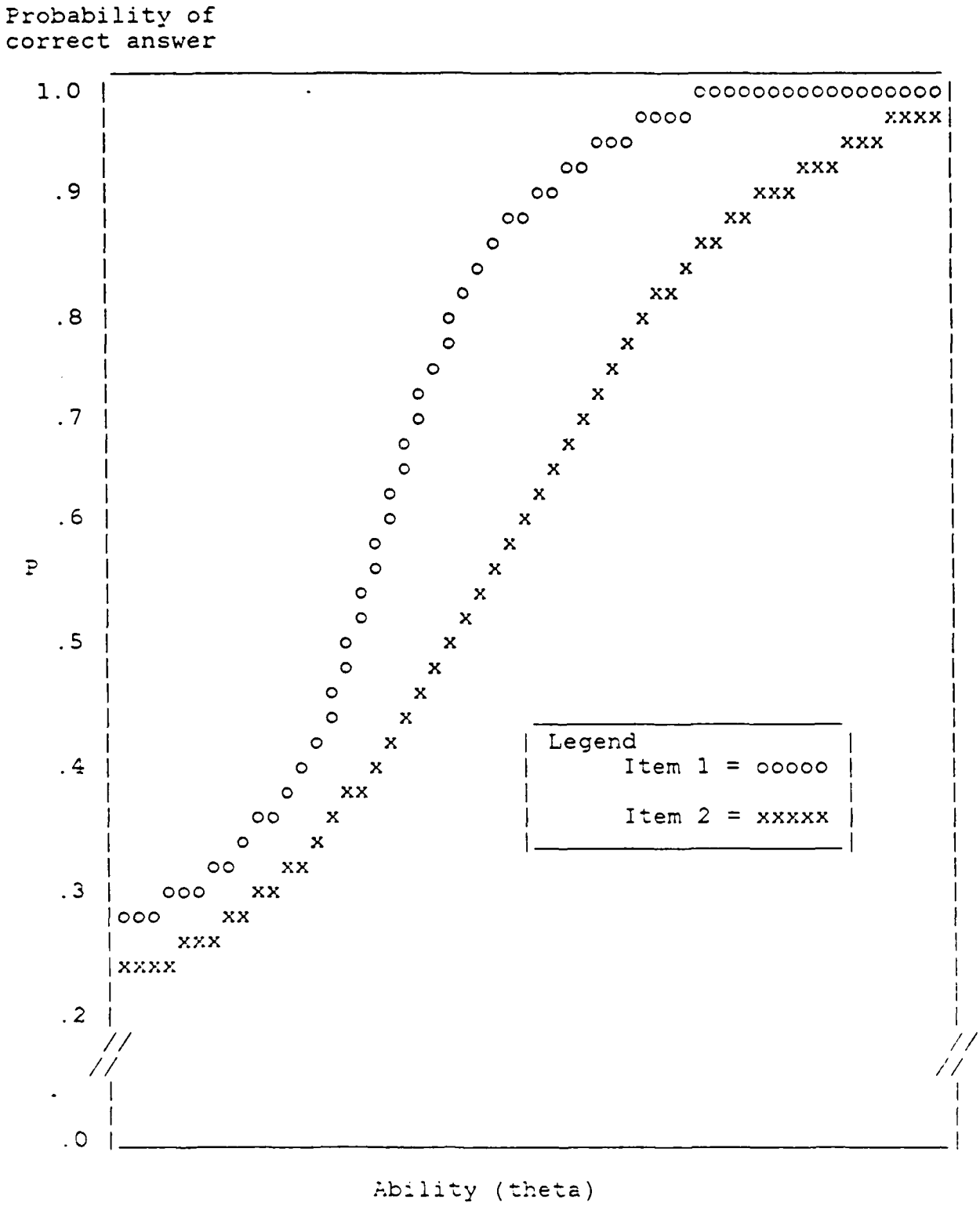


Figure 1. Sketch of item response curves for two items. The graph shows the probability of a correct answer as a function of ability level.

It might be supposed that for four-option items like those on many tests, c would be about .25. However, c is often found to be less than would logically be expected if wrong answers were random guesses. Not all examinees guess when they do not know the correct answer, and wrong answers may be due more to misinformation or incomplete information than to guessing. One study shows that on some four-alternative multiple choice tests, the c parameter varies from .10 to .35 or more, with a median of about .20 to .25. Another study finds that if all item response curves for a similar test are forced to have the same c value, a value of .10 is best (Bock & Mislevy, 1981). The third parameter, c, complicates item response theory enormously, and it would be an immense convenience to leave it out. Nevertheless, the three-parameter model is needed. The model does not fit multiple choice items well when $c = 0$.

The theory of statistical estimation provides a powerful way of describing the amount of information in an item, and in a test. The relative amount of information that an item provides about persons of various abilities is called the item information function. It can be shown that maximum information occurs in the vicinity of $\theta = b$, and that this information is proportional to a. This means, first, that a is indeed an index of discrimination, and second, that items with b-values near to a person's ability provide most information about that person. (The specific location of the maximum, and the specific information at that point depend in a complex way on c.)

An elegant feature of IRT is that the information in a test is the sum of the information functions of the individual items presented. Furthermore, test information is inversely related to the variance of measurement error, which permits an estimate of the error of measurement of each score, not just an average standard error of measurement for a population of scores, as in classical test theory. (Bayesian theory provides an equivalent result.)

A conventional non-adaptive test has a single fixed information function that is usually relatively high in the middle range of test scores, and relatively low at the extremes. Thus the accuracy of the test is considerably less for persons with high or low scores than for persons with scores in the middle range. In adaptive testing, items are chosen for each candidate so that the information function for that candidate will be maximum in the vicinity of his or her ability level. In an adaptive test, each item contributes substantially to the information function, so a given level of precision can be achieved with a much smaller number of items than would be possible with a standard fixed-item test. (Again, Bayesian theory provides a slightly different analysis but reaches the same conclusions.)

It is important to recall that at the start of the testing process we know little or nothing about the candidate's ability level. Consequently, in a tailored test, the first item presented is one that is appropriate for the average candidate. After each item response, an improved estimate can be made of the candidate's ability, and a more appropriate item can be selected for presentation. Each stage of the process yields a better estimate of the ability of the candidate, and also an estimate of the standard error of the estimate. The test can be stopped when this error becomes small enough, or the number of items to be presented can be fixed, and chosen so that, on the average, the level of precision is acceptable.

In adaptive testing, the estimate of ability and the choice of the next item require knowledge of the parameters of the item response curves - the a's, b's, and c's. Estimates of these values must have been determined before the testing process is begun. This is usually done by giving all of the items to comparable, large samples of candidates, in a conventional testing situation. If there are too many items for this to be practical, then overlapping subsets of items can be given to several different samples of candidates. Methods are then available for linking the estimates of item parameters. There is a large literature on parameter estimation; see for example Reckase (1978), Ree (1981), and Yen (1981).

CAT's emphasis on item parameters can be considered a refinement of common practice. Conventional test construction also uses knowledge of the characteristics of the available items. Indices of item difficulty and item discrimination are commonly obtained from pretest data. But these values are used in a somewhat informal way in constructing a conventional test, whereas the item parameters are a central part of the adaptive testing process, both for item selection and for test scoring.

The Current ASVAB

The current PP-ASVAB has six forms, which until recently were 8a, 8b, 9a, 9b, 10a, and 10b. Forms 8a and 8b have been retired, and have been replaced by Forms 9x and 9y, which are derived from 9a and 9b. Six new forms of the PP-ASVAB, 11a, 11b, 12a, 12b, 13a and 13b, have been prepared and scaled for introduction very soon.

The PP-ASVAB has 10 separately-timed sections as listed in Table 1. On each subtest, the observed score is the number of items answered correctly. All items offer four answer options, except CS which offers five answer options. Two of the tests - NO and CS - are highly speeded. Although the score is the number of items correct in the available time, examinees tend to make very few errors on the speeded tests. Four of the subtests, WK,

PC, AR, and NO are combined to form the Armed Forces Qualification Test (AFQT).

Table 1. Name, number of items, and time allowed for ASVAB subtests in Forms 8 through 14

Name	Number of items	Test minutes
GS - General Science	25	11
A F Q T		
AR - Arithmetic Reasoning	30	36
WK - Word Knowledge	35	11
PC - Paragraph Comprehension	15	13
NO - Numerical Operations	50	3
CS - Coding Speed	84	7
AS - Auto and Shop Information	25	11
MK - Mathematics Knowledge	25	24
MC - Mechanical Comprehension	25	19
EI - Electronics Information	20	9
Totals	334	144

The Armed Forces Qualification Test

The AFQT is based on the sum of raw scores on WK (word knowledge), AR (arithmetic reasoning), PC (paragraph comprehension), plus 0.5 * NO (numerical operations). The raw scores are summed, with NO scores being multiplied by 0.5. Any fractional score is rounded up to the next higher integer. Table 2 shows the means and standard deviations of the raw scores on the original 10 subtests for ASVAB 8, 9 and 10. The scores are highly intercorrelated, so the relative contributions of the subtests to the AFQT is moot. Means and standard deviations obtained in operational use of the test are shown in Table 2 (From Ree et al). (Note that the mean raw AFQT score is about .25 more than the weighted sum of the means. This is caused by the rounding. Roughly 1/2 of the candidates had their scores increased by 0.5 in the rounding process, - 1/2 of 0.5 is 0.25.)

The raw AFQT sum is transformed to a percentile scale, and is reported only as a percentile. This percentile scale has been scaled, as well as possible, to the original 1944 Armed Forces population, but the linkage is weak because the content of the test has changed somewhat. The AFQT raw scores tend to range from below chance to 105. The percentile equivalents have been smoothed, and an adjustment is made if necessary so that there is a raw score that translates to the 50th percentile.

The AFQT percentile distribution is divided into five categories:

AFQT Category	Percentile Range
I	93-99
II	65-92
III A	50-64
III B	31-49
IV	10-30
V	01-09

All services use the AFQT for initial screening. Individuals who score in AFQT Category V are not eligible to enlist. Otherwise, minimum AFQT scores are adjusted to maintain a fairly constant flow of applicants.

Table 2. Descriptive statistics for Raw Scores on ASVAB 8, 9, and 10 and AFQT-7a (m = mean, s = standard deviation)

Sub-test		ASVAB Form Administered 8a	8b	9a	9b	10a	10b
GS	m	15.29	15.10	14.61	14.59	14.66	14.74
	s	4.83	4.92	5.51	5.54	5.09	5.15
AR	m	16.47	17.13	16.92	17.28	17.93	17.09
	s	6.76	7.13	6.96	6.86	6.70	6.98
WK	m	24.64	23.44	23.53	23.72	22.99	23.43
	s	7.55	7.56	7.66	7.75	7.82	7.60
PC	m	10.08	9.84	9.27	10.02	9.59	10.02
	s	3.38	3.34	3.48	3.28	3.77	3.17
NO	m	34.52	34.75	34.29	33.93	35.03	34.58
	s	10.17	10.05	10.58	10.40	10.04	10.36
CS	m	41.29	41.27	41.42	41.70	42.34	42.08
	s	15.04	15.23	15.05	14.53	14.84	14.42
AS	m	15.25	15.24	15.77	15.74	15.77	15.83
	s	5.82	5.76	5.77	5.71	5.65	5.66
MK	m	11.32	11.14	11.24	11.20	12.33	12.35
	s	5.54	5.43	5.46	5.60	5.33	5.56
MC	m	14.44	14.14	14.28	14.32	14.45	14.27
	s	5.43	5.41	5.33	5.07	5.25	5.20
EI	m	11.50	11.46	11.94	12.05	12.06	11.75
	s	4.31	4.29	4.13	3.98	4.03	4.03
VE	m	34.72	33.28	32.80	33.73	32.58	33.46
	s	10.45	10.40	10.63	10.55	11.09	10.26
AFQT	m	68.69	68.02	67.10	68.22	68.27	68.29
	s	19.22	19.79	19.88	19.78	19.85	19.61
AFQT 7a	m	54.77	54.37	54.68	54.91	54.89	55.40
	s	20.80	20.94	21.02	21.05	20.77	20.82

* From Ree, M.J., Mathews, J.J., Mullin, C.J., & Massey, R.H., Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10. AFHRL-TR-81-49, February, 1982. Manpower and Personnel Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235.

The six forms of the ASVAB include six distinct parallel versions of each of the four subtests in the AFQT composite. The remaining subtests in the ASVAB have only three forms, each in two permuted versions. For example MC on 8a is the same as MC on 8b except that the items are rearranged, according to a simple algorithm. From the data in Table 2 it can be determined that some subtests of identical but permuted items are not equally difficult - the mean difference, although small, is sometimes statistically significant which may indicate that test timing, although liberal, is restrictive enough to cause different items to be reached. (An alternative explanation is that equivalent groups were not realized.) A difference of about .057 standard deviation is significant, since each form was taken by about 2500 cases. Of course, with 2500 cases, it takes very little perturbation to create statistical significance. For example, the means of MC on 8a and 8b differ by 0.4 raw score points, with a standard deviation of about 5.4. A difference of (5.4 x .057) = 0.3 would be statistically significant. The size of the difference is very small, but it is probably of the same order of magnitude as the size of equating errors, which tend to be .05 to .10 of the standard deviation with samples of this size (Lord, 1981a,b).

ASVAB Standard Scores

All ASVAB subtest scores are transformed to a scale with a nominal mean of 50, and a nominal standard deviation of 10. The scores on each of these subtests are scaled to equivalent tests on previous forms, the intent being to reference each subtest to the 1944 population. The scale is truncated at 20 and 80. Any score that would fall below 20 is changed to 20; any score that would fall above 80 becomes 80. (On Forms 8, 9, and 10, no subtest has any raw score that has a transformed score higher than 75 so truncation does not occur at the top.)

One additional standard score is created: VE (verbal) is the sum of the raw scores on WK (word knowledge) and PC (paragraph comprehension), scaled to a nominal mean of 50; and standard deviation of 10. VE is used extensively in composite scores (see below.) Indeed, WK and PC are never used separately, but only as VE.

Composites

All four services use composites of scores on several subtests to assess the suitability of the applicant for various military specialties - Air Force Specialty Codes (AFSC) in the Air Force; Military Occupational Specialties (MOS) in the Army; "ratings" in the Navy. All Services compute aptitude composite scores by summing the subtest standard scores, but from that point on,

procedures differ. The Navy uses these sums directly for classifying enlistees, while the other Services convert the sums to their traditional score scales. The Army and Marine Corps use a standard score scale with mean 100 and standard deviation 20. The Air Force uses a percentile score scale similar to the AFQT scale, except that the percentile scores are reported only in intervals of five units each. The score scales for aptitude composites, as for the AFQT, have been referenced to the 1944 World War II mobilization population.

Table 3 shows the composites currently used by each service. Because of the different scalings, the same composite in use by different services may yield different results. An effort is currently underway to examine the relative efficiency of all these composites, which are quite highly correlated. Table 4 shows the intercorrelations of ASVAB Standard Scores on the subtests; intercorrelations among the composites can readily be obtained from Tables 3 and 4.

Table 3. ASVAB Composites. Each composite is the sum of the indicated ASVAB standard scores. An entry of 2 indicates double weight of that test in the composite.

Composites	Specialty Symbol	ASVAB Tests VE	AR	MK	MC	GS	AS	EI	NO	CS
Air Force/ Army,Marines,Navy										
General/Gen.Tech.	G/GT	1	1							
Administ./Cler.	A/CL	1							1	1
Electronics	E/EL		1	1		1		1		
Air Force (only)										
Mechanical	M				1	1	2			
Navy (only)										
Submarine	SUB	1	1		1					
Crypto.Tech.Int.	CTI	1	1						1	1
Hospitalman	HM	1		1		1				
Mechanical	MECH	1			1		1			
Avia.Struct.Mech.	AM	1			1					
Basic Elec.	BE/E		1	2		1				
Machinery Rep'rman	MR		1		1		1			
BT,EN,GS				1			1			
Army (only)										
Surveill./Commun.	SC	1					1		1	1
Skilled Tech	ST	1		1	1	1				
Operators/Food	OF	1			1		1		1	
Field Artillery	FA		1	1	1					1
Combat	CO		1		1		1	1		
Maintenance Mech.	MM				1		1	1	1	
Army, Marines										
Gen'l Mech.	GM			1		1	1	1		
Marines (only)										
Field Artillery	FA	1	1				1			
Combat	CO	1					1		1	
Maintenance	MM	1		1		1			1	

Table 4. Intercorrelations(1) of ASVAB Subtests for Applicant Sample (N=2375)

	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI	VE
GS	100	69	82	73	47	46	70	63	71	75	82
AR	69	100	68	68	62	52	62	78	67	65	71
WK	82	68	100	80	50	49	68	61	68	75	98
PC	73	68	80	100	52	50	62	59	63	69	90
NO	47	62	50	52	100	62	40	57	42	43	53
CS	46	52	49	50	62	100	41	49	43	42	52
AS	70	62	68	62	40	41	100	50	74	73	69
MK	63	78	61	59	57	49	50	100	60	57	63
MC	71	67	68	63	42	43	74	60	100	72	69
EI	75	65	75	69	43	42	73	57	72	100	76
VE	82	71	98	90	53	52	69	63	69	76	100
Mean(2)	46.3	46.2	46.5	46.6	47.7	47.7	46.5	46.7	46.2	46.5	46.3
S.D.	9.6	9.3	10.2	10.3	10.3	10.0	9.8	9.0	9.5	9.7	10.1
Alpha(3)	.86	.91	.92	.81	(.72)	(.75)	.87	.87	.85	.82	.93

(1) Decimals omitted
(2) Means and standard deviations reported as standard scores with population mean of 50 and standard deviation of 10.
(3) Internal consistency reliability. (For NO and CS, parallel form reliabilities for a recruit sample are given.)

From Maier,M.H. & Grafton, F.C. Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8AX Alexandria, VA: Research Report 1301, U.S. Army Research Institute for the Behavioral and Social Sciences, January 1981. and Ree, M.J., Mullins, C.J., Mathews, J.J., & Massey, R.H. Armed Services Vocational Aptitude Battery: Item and Factor Analysis of Forms 8, 9, and 10. Brooks AFB, Texas: Air Force Human Resources Laboratory, AFHRL-TR-81-55, March 1982.

A given composite may be used for several purposes. For example, the composite of VE and AR is used by all services. In the Navy, this composite is used for at least six different ratings. In each case, the applicant must achieve a certain cut-off score to qualify for that rating, but there is some opportunity for leeway in the setting of the cut-off, depending on other considerations. Also, to qualify for some ratings, the candidate needs minimums on two or more composites. (For example, Advanced Technical Training (BT) requires MK + AS of at least 94 and VE + AR of at least 110.)

Each service allocates recruits to specialty schools by a complex process that attempts to match the recruit to the school in terms of the applicant's aptitudes and preferences as well as the need for recruits and for minorities in the various schools. In the Navy, allocation of recruits to specialty schools is carried out by the Navy's computerized job reservation system. The system offers the recruit five choices; if all are rejected, three more are offered. The net effect is that many specialty schools will have a severely restricted range of ASVAB scores - at least on their composites - since recruits who are not likely to complete training are rejected, and the matching process tends to reject recruits who are too highly qualified.

ASVAB Testing

Testing is done at several kinds of locations. There are 68 Military Entrance Processing Stations (MEPSs) located across the country, to which applicants must be transported. Then there are about 900 Mobile Examining Teams (METs) who are more easily accessible. Some are mobile teams of military personnel. Other teams are located at permanent sites, often in federal buildings, and are civilian employees of the Office of Personnel Management (OPM), which is under contract with the DoD. Finally, teams of examiners visit high schools to conduct the voluntary high school testing program. The voluntary nature of this program is variable - rates of eligible takers range from 0 in some schools, to 100% in schools where it is a part of the school's guidance program.

Equating the ASVAB

The AFQT for ASVAB 8, 9, and 10 was equated to AFQT 7a, a test used from 1960 through 1973, which in turn was calibrated with earlier tests, and so on, back to a population of military personnel obtained in December, 1944. The relation to the 1944 base is tenuous at best. Future forms of the test will be equated to a nationally representative sample of youth between the ages of 18 to 23. This representative sample was obtained in 1980, using ASVAB 8a, and provides a new base - a (weighted)

sample of military-age persons in 1980. (This change has important implications for score calibration because the national sample is 50% women, with a consequent lowering of scores on mechanical comprehension, auto and shop information and electronics information.)

When ASVAB 8, 9 and 10 were produced, a three-stage plan for data collection was implemented. First, items were pretested (in test booklet form) at Recruit Training Centers (RTCs.) This permitted item statistics to be obtained for the new items. Also the relationship of these items to those on the then-operational ASVAB were determined. From these statistics, six approximately equivalent forms were produced for each subtest. In January and February 1980, samples of applicants at the Military Entrance Processing Stations (MEPS, then AFEES) were tested on Form 8a as well as AFQT 7a, and the operational form of the ASVAB (6 or 7.) These data were used to calibrate ASVAB Form 8a with the old AFQT 7a. Most importantly, the AFQT portion of ASVAB Form 8a was calibrated with the old AFQT 7a. Two additional samples, one of recruits and one of high school students were also used. The calibration results for all three samples were nearly equivalent (Maier, 1981).

When Forms 8, 9 and 10 were made operational, in October 1980, operational data were collected during a period called Initial Operational Test and Evaluation (IOT&E). At that time all six forms were presented in an equivalent-groups design; AFQT 7a was also presented as an experimental test. These data were used to determine whether the corresponding subtests on the six new forms were practically parallel, and whether the calibrations of the subtests and the AFQT that had been obtained for Form 8a would hold for all forms. It was decided that the earlier calibrations were satisfactory, and that the six forms of the ASVAB were essentially raw-score-parallel. (Ree, Mathews, Mullins & Massey, 1981). In fact, as can be seen from Table 5 (Table 8 from Ree et al.) the six equated percentile scales for the 6 forms of the AFQT seldom differ by more than one percentile point from the average scale, although there are a few differences of 2 or 3 percentile points. These differences are well within the standard error of measurement on the AFQT, which is about 4 to 6 raw score points, or about 4 to 8 percentile points over the major part of the scale.

Procedures for equating new ASVAB Forms 11, 12 and 13 were more elaborate than the procedures used for Forms 8, 9 and 10. The items were written and pretested using Air Force recruits at Lackland AFB. The items were culled, and six 50-item versions were formed for each subtest. Response data were obtained from further samples of recruits from all services, 1000 per version, in an equivalent groups design. These data were used to determine the final items to be included in each subtest on each form, and to do preliminary equating.

The next step was to gather additional equating data both for applicants and recruits. Test booklets were prepared, containing different subsets of the ASVAB tests, as shown in Table 6. Booklets were designed to include one or more of the composites that must be calibrated, including the AFQT. One set of nine booklets was extracted from ASVAB Form 8a, which was not then in operational use. Another parallel set of nine booklets was extracted from ASVAB Form 11a. Each participating applicant at a MEPS was given one of the 18 possible booklets in a balanced equivalent groups design with 18 groups.

A balanced design was also used for recruits, but each recruit took a complete ASVAB as an experimental test. There were seven equivalent groups, who took either Form 8a or one of the six new forms, 11a, 11b, 12a, 12b, 13a, and 13b. The MEPS testing provided extensive data on one pair of forms only - 8a and 11a. whereas the data from the RTCs provided comparisons of all new forms and one earlier form.

Note that the data collection designs for both the MEPSs and the RTCs provide for equivalent group comparisons of tests identified as experimental and given under non-operational conditions. An informed-consent notice is regularly read before the administration of any non-operational tests, both at RTCs and at MEPSs. A few applicants at the MEPSs choose not to participate, but virtually no recruits opt out at the RTCs. Thus motivation conditions can be different for operational and nonoperational tests, thereby requiring a design in which all tests being compared are administered under similar conditions.

TABLE 5. AFQT Conversion Tables for Current Forms of the ASVAB (From Ree, et al, op. cit.)

AFQT Raw Score	AFQT Percentile for Form 8a	8b	9a	9b	10a	10b	Avg	AFQT Raw Score	AFQT Percentile for Form 8a	8b	9a	9b	10a	10b	Avg
0-17	1	1	1	1	1	1	1	62	28	28	29	28	29	29	29
18	1	1	2	1	2	1	1	63	29	29	30	30	30	30	30
19	1	2	2	2	2	1	2	64	30	30	32	31	31	31	31
20	1	2	2	2	2	2	2	65	31	31	33	32	32	32	32
21	2	2	3	2	3	2	3	66	32	32	34	33	33	33	33
22	2	3	3	3	3	2	3	67	33	33	36	34	34	34	34
23	2	3	3	3	3	3	3	68	34	34	38	36	36	36	36
24	3	3	4	4	3	3	3	69	36	36	40	38	38	40	38
25	3	4	4	4	4	4	4	70	38	40	42	40	40	42	40
26	4	4	5	5	4	4	4	71	40	42	44	42	42	44	42
27	4	5	5	5	5	5	5	72	42	44	46	44	44	46	44
28	5	5	6	6	5	5	5	73	44	46	48	46	46	48	46
29	5	6	7	6	6	6	6	74	48	48	49	48	48	49	48
30	6	6	7	7	6	6	6	75	49	49	50	49	49	50	49
31	6	7	8	7	7	7	7	76	50	51	51	50	50	51	51
32	7	7	8	8	8	7	8	77	51	51	52	51	51	52	51
33	7	8	9	8	8	8	8	78	52	52	54	52	52	54	53
34	8	8	9	9	9	9	9	79	54	54	56	54	54	56	55
35	8	9	10	9	9	9	9	80	56	56	58	58	56	58	57
36	9	10	10	11	10	10	10	81	58	58	60	60	58	61	59
37	9	10	11	11	10	10	10	82	60	60	62	61	60	61	61
38	10	11	11	11	11	11	11	83	61	61	63	62	61	62	62
39	11	11	12	12	11	11	11	84	62	62	65	63	62	63	63
40	11	12	12	12	12	12	12	85	63	63	67	65	63	65	64
41	12	12	13	13	12	12	12	86	65	65	70	67	65	67	67
42	12	13	13	13	13	13	13	87	70	70	72	70	70	70	70
43	13	13	14	14	13	14	14	88	72	72	74	72	72	72	72
44	13	14	14	14	14	14	14	89	74	74	76	74	74	74	74
45	14	14	15	15	14	15	15	90	76	76	78	76	76	76	76
46	14	15	15	15	15	15	15	91	78	78	80	78	78	78	78
47	15	15	16	16	16	16	16	92	80	80	81	80	80	80	80
48	15	16	17	16	16	16	16	93	81	81	82	81	81	81	81
49	16	16	17	17	17	17	17	94	82	82	83	82	82	82	82
50	17	17	18	17	17	17	17	95	83	83	85	83	83	83	83
51	17	17	18	18	18	18	18	96	85	85	87	87	85	85	86
52	18	18	19	18	18	18	18	97	87	88	88	88	87	87	87
53	18	19	20	19	19	19	19	98	88	88	89	89	88	88	88
54	19	20	21	20	20	20	20	99	90	89	90	90	89	89	89
55	20	21	22	21	21	22	21	100	91	90	92	91	90	90	91
56	21	22	23	22	22	23	22	101	92	91	93	92	91	91	92
57	22	23	24	23	23	24	23	102	93	92	94	93	92	92	93
58	23	24	25	24	24	25	24	103	94	93	95	94	93	93	94
59	24	25	26	25	25	26	25	104	95	94	96	95	94	94	95
60	25	26	27	26	26	27	26	105	96	96	97	96	95	95	96
61	26	27	28	27	27	28	27	(105 is highest possible score)							

Table 6 - MEPS Test Booklet Composition

The AFQT is in Booklets 5 and 6.
Every subtest appears at least four times.
The numbers are minutes of actual testing time,
not including the time for instruction, practice, etc.

Book	(11) GS	(36) AR	(11) WK	(13) PC	(3) NO	(7) CS	(11) AS	(24) MK	(19) MC	(9) EI	Total Time
1	X	X						X		X	80
2	X		X	X	X		X		X		68
3	X		X	X				X	X		78
4	X						X	X	X	X	74
5		X	X	X	X	X	X				81
6		X	X	X	X	X					70
7		X			X		X		X	X	78
8		X				X	X		X	X	82
9		X				X		X	X		86

The final step in the equating process is to check the adequacy of the derived scales in the operational setting. For ASVAB Forms 11, 12 and 13, this will occur in October 1984, during their IOT&E period. It should be noted that the concept of an IOT&E period was first implemented in 1980, when Forms 8, 9 and 10 were introduced. At that time there was no need for any readjustment.

Future plans call for all ASVAB scores to be referenced to the 1980 national probability sample, using Form 8a. ASVAB Forms 11, 12, and 13, which are scheduled for introduction in October, 1984, have been equated to ASVAB Form 8a. The new CAT version of the ASVAB must either be made comparable to ASVAB 8a directly, or through comparability with Forms 11, 12, and 13.

CAT ASVAB Systems

An experimental CAT system, based on the Apple III computer, was developed at the Navy Personnel Research and Development Center (NPRDC) in 1982. Earlier experimental CAT systems had been in operation, but for purposes of this discussion, we shall refer to this complete system as the Experimental CAT System. The system now includes four units, each with eight testing stations; each station can present a complete ASVAB.

The experimental system uses Owens-Bayesian estimation of ability (theta) and administers 15 items in each power test except PC, which has only 10 items. NO and CS items are presented for a fixed time; initially 3 and 7 minutes respectively; these time limits have since been changed, since many recruits finished the test within the limit. Except for NO and CS, discussed below, the items appear one at a time on a standard video display screen. Four answer options are offered - the respondent presses one of a small set of keys on a keyboard [a standard terminal keyboard with a metal panel over most of the keys]. The choice appears on the screen. The respondent presses a "verify" button to go on, or he may change his response. In the experimental system, three NO items appear at one time on the screen. They must be answered in order. As the subject responds, his choice appears in the answer box for that item on the screen; when all three items have been answered, and when the subject has verified his responses, the screen presents three new items. The same general process is used for coding speed (CS), except that seven items appear per screen.

In the experimental CAT system, item selection is based on a maximum likelihood information analysis, using a table (the "info table") that lists the most informative items for each theta value from -2.25 to 2.125 in steps of 0.125 theta units. The Owens-Bayesian procedure is used for test scoring (theta estimation). The stopping rule is a fixed number of items. At

present, 15 items are administered for every power test, except PC, for which 10 items are used. Initially, item selection involved a random choice from the ten best[3] items, at the current theta level, that had not previously been administered to this candidate. Later, this rule was changed so that the first item is selected from the five best items at theta = 0; the second item is chosen from the best four unused items at the theta level resulting from the first item response; the third item is chosen from the best three unused items at the theta level implied by the first two responses; and the fourth from the best two. The fifth and subsequent items are always the best unused item at the appropriate theta level.

Specific procedures for the experimental CAT systems were chosen by NPRDC staff for experimental purposes. These procedures are not necessarily those that will be followed in the operational CAT, nor are they necessarily those recommended by Green, Bock, Humphreys, Linn, Lord, & Reckase (1982), or by the present committee. Nevertheless the results from the experimental system have provided extremely valuable data for evaluating the actual use of CAT in realistic settings.

The development of CAT into an operational system is being coordinated by the Computerized Adaptive Testing Interservice Coordinating Committee (CATICC). Three contractors, who won an initial design competition, are currently designing systems. Two or three prototype systems will then be built and field-tested. After careful evaluation, one will be chosen. That contractor will then build and install the operational system, which may be slightly different from the prototype.

All systems are designed to display a multiple-choice item on a screen, and to provide a way for the examinees to indicate their responses. Two systems use keys, a third uses a light pen for responses. Differences in display legibility can be appreciable. For example, the experimental system's display has two sizes of characters. The small size, which is used for paragraph comprehension items, is considerably less legible than the larger size. Also, the experimental system currently displays black characters on a white ground, which reduces legibility considerably because the ground is not uniform. The experimental system's display does not have adequate resolution for some of the diagrams accompanying items on the tests of mechanical knowledge and technical information. The display characteristics of the prototype CAT systems are not yet known. The systems may have other differences that make them not interchangeable, and, of course, an item looks much different on a display screen than on a printed page.

---

[3] In this context, the best items are those that, at the current theta level, produce the largest increase in test information.

The Experimental Items (X Pools)

Two sets of items have been developed for use with the CAT systems. They have been called the prototype item pools and the operational item pools, but here they will be called the X pools and the O pools, respectively (to avoid the complication that the prototype system will use the operational items whereas the experimental system uses the prototype items.)

The X item pools were developed for experimental work, and are used on the experimental system. Data used in estimating item parameters were obtained by administering booklets of test items at the MEPS sites, in a non-operational format and with informed consent. Item level data were also obtained for each person from the operational ASVAB. Appendix B shows the design of the study for five of the subtests. For each of these subtests shown (AR, WK, PC, GS, and MK), the data were analyzed in one huge run of LOGIST, in which theta scores were determined for all test-takers and parameters were estimated simultaneously for all relevant items in the X pools and all relevant items in the six current versions of the ASVAB (8a, 8b, 9a, 9b, 10a, 10b). The item parameters were obtained using a modification of LOGIST 2b (1976); the modifications were mainly in the treatment of the **c** parameter. Similar modifications have since been made for **c** in LOGIST 5 (Wingersky, 1983), which was not then available.

The X item pools for the speeded tests, NO and CS, are reproduced from ASVAB Form 8a. The items developed for MC, mechanical comprehension, were deemed to be not parallel in content to the ASVAB, and were not used. Instead, the items in the X MC pool are all of the MC items on the current ASVAB forms that could satisfactorily be put on the computer. The X pool contains about 70 MC items.

The CAT version of ASVAB treats the auto-shop test (AS) as two separate tests, auto information (AI) and shop information (SI); the two scores will be combined before equating. For AI, SI, and EI,[4] the item pools have well-estimated item parameters, but the item-level response data for the ASVAB was not available when the calibrations were made, so the parameters are not linked to item parameters for the present ASVAB.

The number of paragraph comprehension items in Pool X is much smaller than intended because many of the PC items would not fit on the computer display screen. The item parameters for MC, AI, SI, and EI were obtained with Logist, but are not linked with the P&P scale.

---

[4] NPRDC personnel are not completely satisfied with the content of the EI items, though here the problem is not as severe as with the mechanical comprehension items.

To summarize, the data for five subtests are about as planned - GS, AR, WK, MK and PC. All parameters were obtained using data from paper and pencil administration, and parameters are on the same scale as parameters for the corresponding P&P tests. The X item pools for AI and SI are large but the X pools for EI and MC are small. The parameter estimates for AI, SI, EI, and MC are not linked to current ASVAB items through the same LOCIST run. The speeded tests, NO and CS, are copies. Items from P&P ASVAB tests are used in MC and PC.

CAT Validity Study

The experimental system described above is being used by a contractor, Rehab Group, Inc., to study the validity of CAT scores for training school performance. CAT and P&P ASVAB data are being collected on recruits destined for certain Navy specialty schools. So far, the CAT has been administered to about 1400 Navy recruits, 200 to 250 destined for each of six advanced training schools for certain ratings. For each rating, the recruits have also been retested on those subtests of the ASVAB that contribute to the composite that is used to establish qualification for that rating. ASVAB Form 9a is used if the recruit had not initially taken Form 9, otherwise Form 8a is used. Item-level data are thus available for the P&P ASVAB as well as the CAT. Original ASVAB scores are also available, but not item-level data. Because the full CAT ASVAB was not available at the start of testing, only 600 recruits have been tested on the full battery of 11 subtests. The remaining 800 were tested on seven to ten subtests. In every case, the recruit was tested on the CAT form of the subtests included in the composite used to select his specialty.

The six Navy ratings, and the associated composites are shown here:

Mess Management Specialist	VE + AR
Hospital Corpsman	VE + MK + GS.
Radioman	VE + NO + CS
Hull Maintenance technician	VE + MC + AS
Electronics Technician	MK + EI + GS + AR
Sonar Technician - Surface	MK + EI + GS + AR

In February, 1983, the experimental CAT equipment was moved to a Marine Corps Recruit Depot for data collection on a similar schedule, with different occupational specialties. Later plans include the testing of Air Force recruits, and Army recruits.

In the first phase of data collection, NO and CS were presented with the same time limits as in the paper and pencil versions. However, the test proceeded much more swiftly on the computer; many examinees answered all items in the designated time. Reduced time limits have been used for subsequent groups of recruits.

The data from the first group of recruits are being analyzed, and will be used to obtain a provisional equating of the experimental CAT ASVAB to the P&P ASVAB, following the methods proposed below. These data, although fine for validity studies, are not ideal for equating, because they are limited to recruits. Nevertheless, they will permit examining the implications of some of the proposals.

CAT Item Pool O - The Operational Items.

Plans for the O item pools are well along. The item pools for the various subtests have been developed[5] and formed into test booklets. These booklets have been administered as experimental tests at various MEPS sites; item-level data for the operational ASVAB (Forms 9a, 9b, 10a, 10b, 10x, 10y) taken by each respondent have also been collected. Item parameters for all items are being obtained. We note that our proposal depends upon item parameters being available simultaneously for the operational ASVAB P&P tests and the O pool for CAT.

It should be noted that item content on the ASVAB is specified by reference to current ASVAB forms. The statement of work for item development contracts has specified only that items should be similar in content to the current ASVAB. Content is checked by having the contractor supply items of each content area for screening. This practice does not provide adequate control over item content. The difficulty with the content of some of the X item pools is one example of the need for stricter control. One welcome aspect of the contract for the O item pools requires the contractor to prepare detailed specifications for the content of each test area. When the contractor's specifications have been accepted by the ASVAB technical committee, they can guide future item development.

One issue that needs attention is the specification that the items for CAT be suitable for computer presentation. Diagrams should be simple and paragraph comprehension items should be short.

---

[5] Data have been collected at the MEPSs by the contractor, Dr. David Vale (Project Director, Stephen Prestwood) of Assessment Systems Corporation, 2233 University Ave., St. Paul, MN 55114. Dr. Vale will furnish item parameters, possibly using his own estimation program, about which very little is known. Some simulations are planned to compare his parameter estimates with Ancilles-X and with LOGIST.

CAT Prototype Field Trials

The prototype CAT systems should be ready for field tryouts late in 1984 or early in 1985. Three different vendors are designing system configurations. Two or three[6] of these designs will be chosen for field trial, for which prototype systems will be made. These systems will have different display and response devices, including (a) a graphics terminal with a separate key pad; (b) a special CRT monitor and a light pen; and (c) a CRT monitor with seven response buttons in the bezel. All of the designs have better graphics resolution than the Apple III experimental system.

CAT System Implementation

After the field trials are completed, there will be several months for evaluation and decision. Assuming a positive decision, installation of the actual operational units would begin soon after contracts are let, and would extend over a period of from 15 to 24 months. The operational equipment would not necessarily be identical with prototype versions; the field tests may indicate the need for changes.

---

[6] This report assumes that two vendors will be asked to build prototype systems for field test. If three prototypes are procured, appropriate changes would be needed.

SCALING PROPOSALS

Converting from conventional paper-and-pencil (P&P) test administration to CAT administration poses two kinds of scaling problems. First, the calibration of the CAT items must be checked, because their parameters will have been obtained in the P&P environment. If discrepancies are found to be larger that would be expected from normal statistical variation, then the items will have to be recalibrated, and the tests will have to be rescored using the new parameters. Second, the scores from the CAT tests must be scaled for comparability with the scores from the corresponding P&P tests.

This report recommends procedures for recalibrating item parameters in the CAT ASVAB item pools, and for scaling the scores on each CAT ASVAB test for comparability with scores on the corresponding P&P ASVAB test. A two-stage process is proposed. First, a preliminary CAT battery would be calibrated and scaled on the prototype equipment, for early use in the operational phase. Second, a final CAT battery would be calibrated and scaled early in the operational phase, using the operational equipment. The final CAT battery should be ready for use late in the first operational year.

Scaling, Equating, and Equity

The process of scaling the scores from two tests for comparability is often called "equating", but that term will be used sparingly here. Equating, in the broad sense used here, means that a candidate's expected score is the same on both tests. A candidate should not care which form of the test is used, because his score will be then same on both tests, except for random measurement error.

Some psychometric experts (see Holland & Rubin, 1982) use equating in a much stricter sense. Two forms are equated in this strict sense if for each candidate the expected scores, and the accuracy of those scores is the same on both forms. In an adaptive test, low and high scores will be nearly as accurate as average scores, whereas on a conventional test, low and high scores are less accurate (subject to more variation) than average scores. Thus an adaptive test and a conventional test cannot be equated in the this strict sense. In fact, few if any tests of any sort can be equated by this strict definition.

The purpose of equating is to insure that it is a matter of indifference to the candidate which form of the test is taken. In principal, a difference in accuracy could affect a test taker's preference for one form or the other. Most candidates want an accurate score. However, a candidate who expects to score too low to be accepted might prefer to take a chance with an inaccurate test since his score might be sufficiently wrong on the high side to get him accepted. On the other hand, a candidate who expects to be just barely acceptable wants to avoid the chance occurrence of an error that would lower his score below the cut-off. Note that this strategic consideration depends on the cutoff being fixed and the candidate having an accurate estimate of both the cut-off and his own score. If the cut-off should be lowered just a bit, the candidate who had preferred an inaccurate score will

suddenly find himself on the other side of the cut-off, wishing for a more accurate test. Finally, note that this analysis of the strategy of betting on measurement errors supposes that a candidate wants to get accepted, whether or not he is qualified. From the services' viewpoint, more accuracy is always better.

A more elaborate Bayesian analysis would consider the prior distribution of potential cut-off scores and the prior distribution of the candidate's potential test score. In general, if the mean of the latter exceeds the mean of the former, the candidate should prefer the more accurate test, otherwise the less accurate test. However, the relative gains from this choice are very small, even if the candidate's priors are objectively correct. Even this small potential gain is lost if the priors are wrong. In any case, the various gains and losses all stem from trying to take advantage of measurement errors. On the average, a candidate's scores on the two tests will be the same.

The same issue arises for assignments to various specialty schools. Again, at each decision point, the person expecting to be below the cut-off should prefer more measurement error, whereas the person expecting to be above cut-off should prefer less measurement error. Since these cut-offs differ, and are all higher than the AFQT entrance criterion, many candidates will be on different sides of at least two cutoffs, and more if there are several choices. Trying to assess all these options, still in terms of lucky errors, would tax a large computer, let alone a recruiter or applicant.

Comments in Maier and Truss (1983) and hearsay from many unnamed sources indicate that recruiters may try to take advantage of any edge that they perceive. If the option of P&P or CAT versions of the ASVAB exists, as it will during the time that the system is being implemented, recruiters may steer applicants toward CAT or toward P&P sites. They could learn that a low scorer's chances may be slightly better on the P&P version because of more room for chance success at the low end of the scale. On the other hand, they might hear candidates saying that the CAT ASVAB is easier, not noticing that the scores are no higher. But they are much more likely to favor the P&P version because of some aspect of test compromise. They may be able to coach the applicant on some questions. Consequently the niceties of relative measurement error are not likely to have much influence on a recruiter's decision. Certain knowledge of even one item on a test is worth more than the potential gain from trying to play the measurement error odds.

Since the only possible inequity is an inequity in knowing how to "play the odds" with possible random errors, and since that knowledge is only useful to some hypothetical person who knows both his own expected score and the exact cut-off value in current use, the broad meaning of equating is sufficient protection of equity. No practical inequity results because the CAT ASVAB, by measuring more accurately, provides scores that are not, strictly speaking, completely exchangeable with the P&P ASVAB. The very small gains that might in principal occur are difficult to realize in practice. Further, anyone trying to act on these marginal considerations is about as likely to be harmed as helped. Equity is better served by making all scores nearly equal in accuracy and by maintaining test security, both of which are better done with CAT.

The effect of item calibration errors

The item parameters must be checked to insure that the CAT scores correlate as highly as possible with the P&P scores and so that CAT can be as efficient as possible. To understand the effect of item calibration errors, suppose that all items are easier in CAT mode than in paper and pencil mode, but are just as discriminating. Suppose that the a values are unchanged, but that the b value for each item should be reduced by a constant d. Consider first the case where c=0 for all items. Then, after a number of item responses have been made, the estimated theta score will be too high, by the constant d. Since both the item locations and the thetas are too high by the same constant amount, appropriate item selection will occur. However, the first item selected will be for someone with a theta value lower than intended, by the constant d. For example, if the system intended to start with an item appropriate for someone with theta = 0, an item appropriate for someone at -d would be selected. A more difficult first item would be more efficient. Although the effect of the inappropriate first item is quickly overcome, the net result is a slightly larger score uncertainty after a fixed number of items. Also, if Bayesian scoring is used, the scores will be regressed to the wrong values, so there will be some bias in the score estimates, as well as a loss in efficiency. However, the effect of this bias is mainly to distort the theta scale, which can be corrected at a later step, when the final CAT test scores are equated.

Now consider the effect introduced because c is not zero. In this case, to a small (and unknown) extent, the item selection will be wrong if the b's are wrong. The error probably amounts to a non-linear scale distortion, which would combine with the distortion caused by using the wrong mean in a Bayesian method of scoring, and which could be largely corrected by score equating.

Inefficiencies of unknown, but probably small, size will thus occur if the item parameters are wrong in a consistent way, but there will be very little distortion or change in the equated score scale. The same remarks can be made about an over-all change in a values. If all a's are 10 percent too small, the scale of theta will be correspondingly expanded.

In short, a consistent mis-scaling of item parameters has the same effect as a population of unexpectedly high ability, or an unexpectedly homogeneous population. The effect on an adaptive test is small, and in principle is corrected by equating.[1]

---

[1] A full technical analysis would note that any effect that changes the theta scale by a linear transform can be made transparent to the system by a corresponding transform of the item parameters. If

$$\theta^* = m\theta + d$$

then

$$a^* = a/m$$
$$b^* = mb + d$$

On the other hand, if the items are differentially affected by presentation mode, with some items easier in CAT mode, and others not changed in difficulty, then the items selected by the CAT procedure may be inappropriate. Scores would then be less informative and less precise. Under these circumstances, there is no alternative but to reestablish the parameters of each item in the CAT context.

There are several reasons why CAT item parameters might be inaccurate. First, the data used in the initial estimation of the parameters will be obtained under special testing conditions, with possibly less motivated test-takers. Second, the data will be obtained in paper and pencil format, not in computer format. Third, the paper and pencil test cannot be administered adaptively. These three effects should be general, affecting all items equally. As noted above, general effects can be easily corrected.

On the other hand, the computer display format may favor some items over others. Items with especially long stems may be affected differently from short easy-to-read items. Items with diagrams may be affected more than items without diagrams. Differential effects are troublesome.

Very little evidence exists concerning the effect of computer terminal presentation upon the difficulty of test items. The typical short item can probably be read about as well on a CRT screen as printed on paper. The buttons for responding may be slightly unfamiliar, but the chance of putting down the answer on the wrong space on the answer sheet is greatly reduced in the computer format. One early experiment at NPRDC using timed tests found no difference in a vocabulary test but a difference favoring the P&P format for a reasoning test (Sacher and Fletcher, 1978). A recent experiment at NPRDC using an arithmetic reasoning test found a small but clear difference in favor of the printed form. Preliminary evidence from the CAT version of the ASVAB suggests no difference at all, for the typical item (Allred & Green, 1984).

If it could be established with certainty that the CAT item parameters are correct in the sense of being on the same scale as the P&P version, then the item calibration step is unnecessary. The only problem would be to equate the CAT theta scores to the ASVAB standard scores now in use. This can be done from any large heterogeneous sample, using ordinary equating procedures, as discussed below. However if there is doubt that the item parameters apply to the CAT format, then the items must be recalibrated, and the CAT test responses rescored before proceeding to establish scale equivalence.

Because the relation of CAT testing to conventional P&P testing is not completely clear, it is recommended that both item calibration and score scaling be done, in sequence. First the item parameters must be checked. If any parameters are changed, then the tests must be rescored using the new parameters, obtaining new theta scores. Second, the new theta scores must then be scaled to the P&P ASVAB, using standard equating methods.

Item recalibration

The details of item recalibration are discussed in Appendix C. Here the general outline of the procedure will be indicated. The basic idea is to use the parameters of the items on the operational ASVAB, which are considered known and fixed, together with each applicant's responses to the CAT items, to fit an ICC for each CAT item, on the theta scale of the operational P&P ASVAB. The fitted curve for each CAT item can then be compared with the ICC obtained for the item in its P&P form in the initial calibration. When this is done for a set of items from a given test the relation of the new and old ICC's can be determined.

At present, item parameter estimates are available, on a common scale, for the CAT operational items in their P&P form and P&P ASVAB Forms 9a, 9b, 10a, 10b, 10x, and 10y. The parameter estimates for the P&P ASVAB forms are each based on a small number of cases. However ample data are, or can be, available to get good parameter estimates of the items parameters for each form on its own separate scale. These are the estimates that should be used in further analyses, but they should be put on the common scale by the Lord & Stocking technique (1983). The new estimates of the CAT item parameters in the computer environment will also be on the common scale. Appendix C suggests ways of achieving this. The differences between the item parameters in P&P mode and CAT mode provides an indication of the CAT-P&P mode effect.

The data normally obtained from administering the CAT tests will not be sufficient for item recalibration. First, CAT item selection avoids items too easy for the candidate, so there will be very little data for recalibrating the c parameter. The c parameter reflects the lower portion of the item response curve, where the probability of a response is low. Only when the current theta value of a candidate is seriously overestimated will the item be presented to persons whose final theta value is actually in the lower portion of the curve. The data will be ideal for estimating the a and b parameters Second, some items are not used very often, so a great many tests would have to be given in order for enough responses to accumulate in the ordinary course of CAT testing. Preliminary data from the CAT validity study show that for a recruit sample, only about 60-70 items are presented to at least 1% of the test-takers, only about 45 items are presented to at least 10% of the recruits, and only about 25 items to as many as 25% of the recruits.

Three methods are suggested for coping with the data sparseness. First, only a subset of items can be recalibrated. Below we propose defining an initial CAT form called Form 99 with only 50 items per test pool. This smaller number of items will be adequate, and will simplify item recalibration.

A second method of coping with sparse data is to add some item presentations to each test. After the CAT items for a test have all been presented, a few more items can be presented for item calibration purposes only. These items would not be used to get the candidate's theta score that will form the basis of the test scaling process. Just how many items to add depends on other decisions, and is discussed below.

A third method of coping with sparse data relates to the c parameter. Almost all of the data available for each item from its natural use in CAT testing will be data in the informative range of an item's ICC. The item will seldom be presented to someone for whom the item is too easy (a probability of correct response above .90) or too difficult (a probability of correct response of about c). If estimates of the c parameter are required, it will be necessary to present each item to at least 100 test takers with theta scores in the low range of the item's ICC. A schedule of item presentations could be devised; note that the extra items are needed only for the low-scoring applicants, presenting a dilemma for equity. For purposes of rough calculation, we could suppose that 1000 cases out of a target sample of 2000 cases would yield useful data. For an extra 100 responses per item, we would then need one extra item per test for each 10 items being reestimated.

The alternative method of coping with the c parameter is to fix it at some reasonable value, such as its value in the P&P mode. The c parameter has little if any direct influence on either item selection or item scoring. Its effect is indirect, through its effect on the a and b parameters. Any particular ICC can be fit in its influential region, by a curve with any moderately low c value, by suitable adjustments in the a and b parameters. It is the ICCs, not the parameters themselves that govern the adaptive process and the test scores. The authors of this report are divided on this point, but the concensus is that it will be at least adequate, (and some would say completely satisfactory) to use the c values obtained in the original P&P item calibration. We thus propose that in item recalibration, where data are sparse, that the original c values be treated as fixed for each item, and only the a and b parameters refit.

This procedure does not imply that we expect no changes in examinees' chance performance on items in the CAT environment. In fact, some changes are likely. Because CAT does not permit skipping a presented item, more guessing is to be expected. The increment would be small, because there is not a great deal of skipping on the current P&P ASVAB. It is not yet clear whether computer presentation in itself either encourages or discourages guessing. Further, it is not yet clear whether any added guessing will be random, or whether the more popular distractors will attract more responses. The former would push c toward .25, whereas the latter would tend to reduce c below its P&P value. Finally, whatever the effect of the computer on guessing, the amount of guessing will be less in the adaptive test because items are selected so that respondents are seldom faced with the need to guess. Thus, the correct value for c is of very minor consequence either in item selection or item scoring, so pragmatically, the P&P values may as well be used.

Various plans have been put forward in Appendix C for establishing the new parameters. Some adcitional research is needed to determine the best procedure for estimating parameters from the CAT data. Nevertheless it must be emphasized that the methods proposed are adequate. Also, any remaining difficulties with the resulting scale will be corrected by the next step of equipercentile equating, discussed below. The main differences among the various psychometric procedures proposed would be in the eventual efficiency of the CAT system.

Equipercentile equating

When the item parameters have been readjusted, the tests must be rescored. The rescored thetas from the CAT must then be equated with the scores from the P&P version. This requires a sample of examinees who have taken both the CAT version and a conventional P&P form of the ASVAB. Each P&P test provides two scores, a raw score and an ASVAB standard score derived from the raw score. (WK and PC raw scores are combined to yield a single VE standard score.) Only standard scores are used operationally, so the CAT tests must provide equivalent standard scores. The standard scores are used in the occupation specialty composites. However, raw scores are used in obtaining the AFQT and VE composites, and will be needed for AS, so an equivalent of raw scores must be provided. However, the equivalent raw scores need not be reported.

Occasionally, references to AFQT raw scores or other raw scores are encountered. In retrospect, with the advantage of hindsight, it would have been better never to have used raw score composites and never to report them except for research purposes. The use of raw scores should be discouraged, and attempts to provide published raw score equivalents from the CAT should be discouraged as well.

To obtain a CAT equivalent for the ASVAB standard scores, the regular procedure would be to use equipercentile methods to equate the rescored CAT thetas with the P&P derived ASVAB standard scores. Thus, for all tests except WK, PC, AI, and SI, an equivalent standard score should be provided by directly scaling CAT theta to P&P-derived standard scores using equipercentile methods. There is no need for an intermediate step of obtaining raw scores. An intermediate step will be useful in checking the equating by means of IRT methods, but this will be a check of the primary method.

CAT scores can be scaled to P&P derived standard scores from any test form, on the assumption that the test forms already yield comparable standard scores. However, it would be better to have a single P&P form, to reduce the variability of the equating, and it would be best if the P&P form were 13c (a.k.a. 8a), because then CAT could be scaled directly to a known norm, without intervening calibrations. If different P&P forms are involved, there will not be enough data to permit repeating the procedure for each different P&P form, so the form differences will have to be ignored, at least when scaling CAT Form 99. The details of equipercentile equating are discussed in Appendix A.

In addition, it will be necessary operationally to obtain equivalent raw scores for the tests that are involved in the AFQT, VE and AS composites. For the equating sample only, it will be useful to obtain equivalent raw scores on all the adaptive tests in the battery, with equipercentile methods. Equating of AFQT, VE, and AS will be discussed here. The use of equated raw scores to check the recalibrated item parameters, and incidentally to check the use of IRT theory in the adaptive process will be discussed in connection with the evaluation of the equating.

Special scaling procedures

AFQT: The AFQT is currently derived from a combination of raw scores on four tests: WK, PC, AR, and NO. The safest scaling procedure will be to obtain equivalent raw scores on each test from the CAT scores on these tests and to combine these equivalent scores by the usual formlua - WK + PC + AR + 0.5NO. Then this combined score must be scaled with the corresponding AFQT percentile score from the P&P test via equipercentile methods.

Two methods are suggested for use in obtaining equivalent raw scores from the CAT thetas. Method A is to obtain an expected true score on some form of the ASVAB for which item parameters are available on the common CAT scale. An expected true score can be found for each possible theta by simply summing the probabilities of correct response to the P&P items, over the items in the test. This is a theoretical curve computed from the item parameters of the P&P test. Each person's theta can be transformed to an expected true score by the theoretical curve. The variance of the expected true scores should be adjusted by a multiplying constant so that it is the same as the known raw score variance of the P&P test scores. These scores are not equivalent to raw scores in detail, but are a good basis for forming composites.

Method B would be to scale the CAT thetas to the P&P raw scores by equipercentile methods, using the equating sample. Here, at least for the field test data, there will not be enough data to scale each P&P form separately, so the combined forms would have to be used, unless all cases in the sample had scores on the same P&P test. A single form is much preferable in this method, but a mix of forms could be used if necessary.

The result of either method is a set of raw score surrogates that must then be combined following the usual AFQT formula. The combined score will then have to be transformed to an equivalent AFQT percentile score.

VE.: The VE score from the P&P test is based on the sum of raw scores on WK and PC on the P&P test. For CAT, equivalent raw scores on WK and PC should be obtained, using either method proposed above for AFQT. The equivalent scores for an individual should be added, and the result should then be scaled with the VE standard score from the P&P test, via equipercentile methods, to obtain an equivalent VE standard score from the CAT.

The AI and SI tests.: The new AI and SI do not have an exact counterpart in the P&P ASVAB. The single P&P test is represented in CAT by the two tests AI and SI because of concern that AS might not be a unidimensional test. Nevertheless the present ASVAB battery provides one score, so the two CAT scores will have to be combined somehow. For purposes of item recalibration, each should use the AS test on the P&P ASVAB as its counterpart.

Several procedures are possible for combining the AI and SI scores. The simplest would be to scale the AI and SI CAT theta scores to raw scores on the P&P AS test by either Method A or B. At this point the scores should

be combined. We recommend weighting the equated raw scores equally. An alternative would be to use weights from the multiple regression of the two equated raw scores on the P&P ASVAB standard scores.

A neater but more elaborate alternative would be to derive two raw scores on the P&P AS test by dividing the items into AI and SI items. Each item would have to be classified as either auto- or shop-related. Then the corresponding CAT thetas could be scaled to the P&P raw scores by Method A or B, and then the two scores combined. Again, we recommend equal weighting. The combined score from the CAT tests can then be scaled to the P&P ASVAB standard score by equipercentile equating.

Occupation specialty composites.: The scores for occupational composites are normally obtained by adding ASVAB standard scores using weights shown in Table 3. For the equating samples, each composite should be obtained both from the operational P&P standard scores and from the scaled CAT standard scores. Equipercentile scaling should be applied to the composites. If the interrelations (covariance structure) of the scaled standard scores from the CAT tests is about the same as those from the P&P tests, there should be very little change in the composite scales, but they must be checked. Equivalence of the two sets of scores cannot be assumed, but must always be checked.

Summary of procedure.

The complete set of steps recommended for the equating of each CAT power test to its corresponding P&P ASVAB test is listed below. Here the term "CAT score" is taken to mean CAT theta for the adaptive tests and computer produced scores for the speeded tests. This analysis is to be done on a sample of cases who have taken both CAT and P&P versions of the battery.

1. For each adaptive test, item parameters are reestimated. Adjustments are made if necessary.

2. Any test for which item parameters have been adjusted must be rescored.

3. For each test except WK, PC, SI, and AI, obtain equivalent standard scores from CAT by equipercentile scaling of corresponding CAT scores and P&P standard scores. Prepare tables for operational transformation of CAT scores to equivalent ASVAB standard scores.

4. Equate raw score composites.

    a. For each test in the ASVAB except CS, obtain equivalent raw scores by either Method A or B above. (Equivalent raw scores will be used for checking the scaling of all adaptive tests.) For WK, PC, AR, NO, AI, and SI, prepare tables for operational transformation of CAT scores to equivalent raw scores.

b. AFQT: Combine equivalent raw scores by the standard formula (WK+PC+AR+0.5NO). Scale this combined score from the CAT with the AFQT percentile scores from the P&P test, using equipercentile methods. For operational use, prepare a table to transform the combined equated raw scores to the equivalent AFQT percentile scores.

c. VE: Add equivalent raw scores on WK and PC. Equate this combined score from the CAT with the VE standard scores from the P&P test. For operational use, prepare a table to transform the combined equated raw scores to the equivalent VE standard scores.

d. AS: Add equivalent raw scores AI and SI from the CAT test. Equate this combined score with the AS standard score from the P&P test. For operational use, prepare a table to transform the combined equated raw scores to the equivalent AS standard scores.

5. Check each test except AI, SI, NO and CS by obtaining the monotonic transformation from CAT theta scores to expected number right scores on the corresponding P&P ASVAB tests. This should be linearly related to the equated raw scores.

6. Form occupational composites of equated standard scores from the CAT. Equate each composite to the corresponding composite from the P&P scores using equipercentile methods.

In operational computer testing, the tables generated in the equating steps will be used as follows.

1. An ASVAB standard score is obtained from each CAT test score except WK, PC, AI and SI, directly from the tables obtained in Step 3 above.

2. For AFQT, VE and AS:

   a. An equivalent raw score is obtained from each person's CAT scores on AI, SI, AR, WK, PC, and NO, using the tables generated in Step 4 above for use in AFQT, VE and AS composites.

   b. For AS, VE and AFQT, equivalent raw scores are combined: AS = AI + SI; AFQT = WK + AR + PC + 0.5 NO; VE = WK + PC.

3. For occupational composites, ASVAB standard scores are combined as in Table 3, and the combined values are transformed to equated occupational composites using the tables from Step 5 above.

Evaluating the equating

Equivalent raw scores and IRT expected true scores.: The first major evaluation of the equating is the relation of the equivalent raw scores to the expected true scores on the P&P test. As described above as Method A, IRT theory provides a way by which it is possible to compute the expected true score of each test-taker on the P&P tests from a knowledge of the item parameters of the P&P test items and the CAT theta of the test-taker. From the two sets of item parameters, a monotonic function can be generated, showing the expected true score for any possible theta. The expected true scores for a sample can then be compared with the actual raw scores on the P&P test. For this analysis, each examinee must have taken the CAT and the P&P ASVAB form whose item parameters were used to generate the function. The relationship should be linear and the means should be equal. The relation of the standard deviations should be predictable from the precision (reliability) of the test scores, since true scores are regressed to the extent of the average standard error of measurement. The extent of the similarity provides a check on the theory and a check on the revised item parameters. If more than one P&P form was used in the equating sample, this comparison must be done separately for each different P&P test form, and item parameters must be available for the items on each form, so it would be best to use only one form, if possible. If any significant departures from the theory are found, explanations should be sought.

Other evaluations of the equating: The equating can be evaluated in three additional ways. First, a scatter plot of equated CAT scores vs. P&P scores can be produced. These scatter plots are expected to be linear and clustered around a 45 degree line. Second, the correlation coefficient should be nearly as high as the reliability of the ASVAB tests. These two analyses can be done for each test in the ASVAB, both for the raw scores and the ASVAB standard scores. It can also be done for each of the composites, including AFQT, VE, as well as the various specialty composites (Table 3).

A third analysis is appropriate only for the AFQT and the specialty composites. For each of these scales, a fourfold table can be produced for each known decision point and the proportion of disagreement of classification can be determined. These proportions should be very small.

THE PROPOSAL FOR TWO CAT FORMS

The various uncertainties mentioned above suggest that any test equating that is done in advance of having operational equipment in regular use will have to be readjusted after operational experience. The adjustments may be small but they may not be negligible, and cannot be ignored in the equating plan.

Rather than making a series of adjustments to CAT during the early stages of implementation, it seems more reasonable to designate an intermediate test, here called Form 99. This test would use a subset of items from Pool O, and hence would not be equivalent to the eventual operational CAT, here called Form 100, with the full item pools. The items in Form 99

would have been checked and possibly recalibrated using data obtained on prototype equipment rather than the operational equipment, but the tests in Form 99 would have been properly equated to the corresponding tests in the P&P ASVAB; indeed Form 99 would serve as the link between the P&P ASVAB and Form 100.

The calibrations and equatings for the tests in Form 99 would be done using data collected during prototype field testing. This preliminary battery would be ready to be put in place when CAT becomes operational, but it would have limited item pools.

Calibration and equating of Form 100, the eventual battery, would use data collected on the operational equipment during the initial months of operational implementation. This time period can also be considered an IOT&E period for Form 99, during which the equating of Form 99 can be checked. Form 100 will replace Form 99 as soon as possible, probably by the end of the first year of installation. At that time, a final IOT&E period will be needed to check the equating of this full CAT battery.

When the CAT system is fully operational, it is anticipated that data for the calibration of new test items will be collected on-line by including trial items in the item bank with previously calibrated items. With these data, parameters of the new items can be estimated on the same scale as the operational items, so new items can be added without the need for further equating.

Form 99

A complete CAT battery can be assembled using only about 50 of the 200 items in each test's O pool. This preliminary CAT battery would be administered on an experimental basis during field-testing.

The items in Form 99 can be designated naturally from the item parameters that will have been determined for the O pools. For each test in the battery (except CS and NO), the operational system will include an item selection table (the "info table" in the experimental system) that indicates which items are available for presentation, as a function of the candidate's current theta estimate. Only about 25% to 50% of Pool O will actually be in the table, and these items, or many of them, should constitute Form 99. With little loss, the table can be limited to 50 items per test.

For item recalibration, at least 1000 responses per item would be desirable. If the tests were known to be equivalent, so that the CAT environment had little if any effect, then 500 cases would be sufficient to establish that fact, because the ICC's could be determined by regression on the theta values, which could be treated as known. However, when that issue is in question, then 1000 cases would be desirable, because iterations (of the LOGIST sort) will be needed. Data from the experimental CAT system indicates that the tests with short stems and no diagrams - GS, AR, WK, and MK will not be much affected by CAT mode,. However there is some chance that the other tests - MC, AI, SI, EI, and PC - will show a larger effect. It

may be that the graphics on the prototypes are so good that these items will not differ much from their P&P counterparts, but the conservative approach is to expect some differential effect of mode.

For test scaling, 2500 cases are desirable per test. Taking 2500 as a base, an analysis described in Appendix D suggests that adding two (2) items to each test where minor effects are expected will yield about 500 cases per items, while adding six (6) items will yield about 1000 cases per item. The analysis is quite rough, and better estimates can be obtained, but the order of magnitude won't change much. The analysis aiso shows, as might be expected that the problems will be with the extreme items. It will surely be necessary to halve the test-taking group, and adding additional easy items to the less able group, while adding difficult items to the tests of the more able group.

How will the added items affect the battery? Assuming that four of the tests (GS, AR, WK, and MK) can be recalibrated with 500 responses per item, whereas the others (PC, MC, AI, SI, and EI), need 1000 responses each, 38 additional items would be needed, extending the test by about 25 minutes.

Partial CAT batteries.: If time prohibits giving each examinee a complete CAT battery in any phase of the plan, equivalent data can be obtained for scaling by obtaining four times as many cases, each receiving only a partial battery. Each examinee would be administered a subset of experimental CAT tests that could be completed by almost all examinees in 70 minutes. Different combinations of subtests would be administered to different samples with the requirement that the combination of subtests for all composites be administered to a minimum of 2000 examinees. The following four experimental CAT batteries of various combinations of the subtests should be sufficient to cover all composites:

CAT Subtest/ (Estimated time)

Battery	WK (6)	PC (10)	AR (18)	NO (3)	CS (7)	MC (13)	AS (10)	MK (16)	EI (7)	GS (8)	Total Time
1	X	X	X	X	X	X	X				67
2			X	X	X	X		X	X	X	72
3	X	X				X	X	X	X	X	70
4	X	X	X	X		X	X		X		67

Every composite in Table 3 above is represented in at least one of the four batteries. Every pair of tests is represented in at least one group. (The presence of NO in Battery 2 is only for the purpose of pairing it with GS and MK.) Times shown are for the prescribed test; added items take added time.

Partial tests would not only be troublesome operationally, but would be less desirable for scaling. It would be better if all candidates had the same context for test responding. However the main constraint is the need to rescale each separate composite. That is a real need, and must not be lost sight of. Important career decisions rest on the composites.

rescale each separate composite. That is a real need, and must not be lost sight of. Important career decisions rest on the composites.

In order to recalibrate the occupational specialty composites, samples of 2500 per battery would be desirable,. However, since changes are expected to be minimal, 2000 would be an acceptable compromise. This would give much more than the needed minimum for individual tests.

Selection of field test sites.: Because the data obtained during the CAT system field tests will be used for equating Form 99 as well as for item recalibration, the selection of field test sites is relevant to equating. Many factors must be considered in making the site selection. For purposes of equating, it is important that applicants have a wide range of ability with a sufficient number of low as well as high scoring examinees. The applicants at these sites should also be reasonably representative of the national applicant pool in terms of race, gender, and educational background. It is important to note that these applicants will have to take Form 9 or 10 of the ASVAB, rather than one of the then current forms (11, 12 or 13). Above it was assumed that item parameters for Form 99 and the P&P test are available on the same scale, and we understand that this is only true for Forms 9 and 10.

If feasible, each vendor should install equipment in two sites, to protect against some unforeseen idiosyncrasy of a particular site. This could be accomplished by rotating equipment between sites halfway through the field tests. Assuming that two vendors participate in the field tests, the location of equipment might follow the schedule outlined below:

	Field Test Sites	
Vendor	First Half	Second Half
A	1	2
B	2	1

The experimental CAT and operational ASVAB should be administered in counterbalanced order with half the applicants taking the CAT first and the other half taking the ASVAB first. Item response data from both an operational P&P version and CAT version are needed for 2500 applicants per vendor.

The requirement of 2500 CAT examinees per prototype can be translated into equipment needs, and planned usage, in many ways. Various strategic considerations are important and are better known to operational personnel. In some plans, CAT testing will occur for only some of the applicants being processed at the field test sites. In these plans, CAT test takers should be chosen on a random schedule, to insure a proper sample of applicants.

The data analysis will be done in the period between field testing and operational implementation, and may as well be done for both prototypes, although only one scaling will be used.

As this report is being prepared, the extent of field testing is currently unclear. Field testing may extend over a period of many months. From the point of view of accurate test scaling, the more observations the better. Also, an extended field testing may mean that testing in the MET sites is more likely, which would tend to boost the number of low-scoring applicants. For purposes of test scaling, the more low scorers the better.

A serious problem with field tests in general, and extended field tests in particular, is the inevitable desire of the manufacturers to make hardware and software adjustments. Some of these are unavoidable. Most will not affect test scores. Slight improvements in display legibility, slight changes in system latency in providing the next item, and the like, probably have little effect. But a substantial change in legibility, or a change in the graphics display system could have a pronounced effect. And even a minor change in the response system will have a clear effect on the speeded tests.

On-site engineers and programmers will have a strong tendency to do things to the system. On-site data collectors and test supervisors should strongly discourage this tendency. To be sure, we want the best system that can be devised, but we also want a test ready to go when the operational units appear, and this will not be possible if the prototype system keeps changing.

IOT&E for Form 99

During the initial stage of operational implementation, the equating should be checked by the method of equivalent groups. In this method, two equivalent groups of test takers take two different forms of the test. The two score distributions of the two groups should be the same if the two forms are properly equated.

Within-MEPS equivalent groups.: Normally, equivalent groups are formed by assigning successive applicants, or successive groups of applicants to the alternate forms. This might be difficult to accomplish during system implementation. Perhaps the MEPSs could use CAT and P&P on alternate days. We recommend a sample of 2500 cases.

Between-MEPS equivalent groups: An adequate alternative would be to match MEPS, and to consider as equivalent, samples from two matching MEPS. This requires some additional assumptions, and the equivalence is less controlled, but the procedure would be much easier operationally. For each of the first four MEPS in which CAT is installed, two other MEPS could be identified that now yield very nearly the same P&P test score distributions as those from their matched target MEPS. Before CAT installation, detailed P&P data could be collected from all 12 MEPSs that will be target or matching MEPS in the design. For each set of three MEPSs, one target and two matching MEPSs, score distributions would be determined for each test in the battery. Then, when CAT has been installed in the target MEPS, data will again be collected. Score distributions should bear the same relationships to each other after CAT as before CAT. Any alteration in the CAT score distributions would be presumed to be due to inadequate CAT scaling. If the pattern occurs over two or three target MEPSs, then adjustment is required.

This procedure assumes no population drift across MEPSs during the interval of CAT installation. It also assumes that the target MEPSs are neither more nor less popular as a result of having the new CAT test equipment.

We recommend IOT&E samples of 1000 from each participating site. Note that this will not involve any experimental testing; all tests will be administered in operational conditions.

As the plans for implementation of CAT unfold, additional opportunities for comparisons within MEPS may occur. In order to take advantage of any such possibilities, data from these MEPS that are taken before installation should be as exhaustively complete as possible, to permit any kind of matching that might turn out to be possible and desirable.

Choosing the design.: We defer the selection of the actual design for IOT&E to persons with more experience in the testing program. We believe that either procedure would work. A disadvantage of the within-MEPS equivalent groups design is that data for item recalibration of Form 100 would be delayed, since the CAT tests would not be given as often as possible.

The main reason for avoiding a within MEPS design at this stage in system implementation is to avoid the inevitable manipulations of the recruiting personnel. We wonder whether random assignment of persons to groups can be achieved. Personnel are sure to have a variety of reasons for preferring one or the other test for the marginal candidates. A between-MEPS design avoids most of these problems. If everyone at a MEPS and associated MET are using the computerized test, then no undesirable selection can take place.

Form 100

The items in Form 100 are a superset of the items in Form 99, namely all the items that are in the final item selection ("info") tables.

During the first six months of CAT operation, some or all of the sites should be designated to provide data for Form 100 by adding items to each test until data are collected to recalibrate the items in the enlarged item pools. Since the operational equipment may be slightly different from the prototype equipment, the procedure for Form 99 should in general be repeated for Form 100 using the new equipment. However, at this point it would be very desirable to recalibrate each item. Because of large test volumes, adding two items per test should produce enough data.[2] For Form

---

[2] The suggestion of adding two items to each test is based on the present procedure of stopping the test after a fixed number of items. If a variable stopping rule is employed, then different numbers of items might be added to tests of different lengths. Suppose, for example, that the algorithm used from 12 to 20 items, depending on the consistency of the item responses - technically depending on the test information. Then 4 items might be added to the tests for those who would otherwise stop after 12 items, whereas progressively fewer items would be added to the longer

100, there will be no corresponding P&P scores. ICC regressions must necessarily be determined against CAT thetas. The problem of using CAT thetas for CAT ICC's will by then have been solved either theoretically or empirically. When the item parameters have been satisfactorily verified or adjusted, then a sample of 2500 cases can be rescored, and the thetas from Form 99 and Form 100 can be equated by the equipercentile method. The tables relating CAT theta to ASVAB standard scores can then be adjusted for Form 100 thetas.

IOT&E for Form 100 is absolutely critical. It checks the entire process. For this reason, we recommend that as a final step, IOT&E for Form 100 should consist of obtaining equivalent samples of 2500 applicants who will take the CAT and the current P&P test. In principal, the same procedures could be used here as for Form 99. However, we strongly recommend a within-MEPS design at this step, as the best way to get equivalent samples. Applicants at chosen MEPS could be assigned on the basis of odd or even social security numbers to either the CAT or P&P version of the ASVAB. The first 2500 applicants completing the CAT version and the first 2500 applicants completing the P&P version at those sites would be used for the IOT&E scaling checks.

At the same time, the equating of Form 100 to Form 99 should be checked using data obtained from other MEPSs. Here equivalent groups are the responsibility of the software alone. Since Form 100 includes all of the items in Form 99 but uses a different item selection table and possibly different scale conversion tables, the computer can administer either form as needed.

On-line CAT maintainence

Once Form 100 is installed, it will be possible to introduce an on-line item-bank maintainence and updating system. The accuracy and stability of the item calibrations can be checked as data accumulate, and new items can be calibrated relative to the dimensions defined by the existing item pool. Eventually, the original items will be replaced, and all items in the bank will have been calibrated on line.

As noted above, since the new items are placed on the equated scale, no further equating of Form 100 is needed. However, some method should be devised for regularly checking scale comparability as the items change.

The speeded tests

Early results from the experimental system indicate that the speeded tests, NO and CS, can be answered considerably faster on the computer. When P&P time limits are used, too many applicants get the maximum score (called a "ceiling effect".) In fact there is some ceiling effect on the P&P versions of the CS test.

---

tests.

The "quick fix" is to reduce the time limit for the computer tests, so that the score distributions will be more nearly like the P&P score distributions. Sympson (private communication) has recommended choosing the time limit for which the resulting scores correlate highest with the corresponding P&P ASVAB speed test scores. This is practical because the computer records response times amd elapsed times periodically, so a shorter time limit can be imposed upon the scores artificially.

The response times recorded by the computer permit an even better measure of performance speed than is possible in the P&P mode. The P&P test is limited to scoring the number of correct responses within some fixed time interval. The time limit is critical, but is necessarily timed manually in P&P administration, and there is no good way to keep a clever examinee from adding one or two answers after the time limit. Thus there is necessarily some error variation due to manual administration.

By contrast, the computer can accurately time the response to each item, as well as the responses to a group of items presented on the screen together. In the experimental system, three NO items appear at one time on the screen, to be answered in order. As the subject responds to each item, his choice appears in the answer box for that item on the screen. When all three items have been answered, and when the subject has verified his responses, the screen presents three new items. The same general procedure is used for the coding speed (CS) test, except that seven items appear per screen.

The computer permits several measures of speed to be obtained. Number of correct responses per unit time and its inverse, average time per correct response, can be obtained, on an item by item or screen by screen basis. A more sophisticated procedure could record time per screen or time per item, and stop when a stable value was reached. In the short term, a variety of computer measures can be evaluated, to find the measure that relates best to the corresponding P&P measure. In the long run, the measures should be examined for their predictive utility. It should be noted that response time is often the most revealing performance measure in cognitive science research. Response times have often been very diagnostic in this research.

Evidence from the first available data obtained in the CAT validity study mentioned above (Greaud and Green, 1984), indicates that the rate of correct responding - number of correct responses per minute - has excellent score distributions. It has the additional major advantage that total testing time affects its accuracy but not its magnitude. Further, it constitutes a refinement of the present measure, which is also a rate of responding, in the sense of the number of items correct in the allotted total time. We recommend that for Forms 99 and 100, as well as for the CAT validity study, correct responses per minute be used as the CAT measure.

Evidence suggests that good reliability can be obtained from 35 NO items and 56 CS items. In the computer form of the test, everyone should get the same number of items, but their times to completion will vary, and not everyone will answer all items correctly.

In scoring the computer speed tests, we recommend disregarding the response to the first item, whcih tends to be too long. Then, all items with times less than some set minimum (0.5 sec. is recommended) should be disregarded. Then the mean and standard deviation of response times can be calculated for this test-taker, and all items disregarded with a time greater than 3 standard deviations above this person's mean response time. The total testing time is then calculated as the elapsed time minus the disregarded times; the remaining time is the divisor for the number of correct responses (eliminating those disregarded) to get correct responses per minute.

This measure will have to be equated to the P&P raw score scale, and thence to the ASVAB standard score scale, by equipercentile score equating. This to some extent destroys the excellent psychometric properties of the measure, but that cannot be helped. At some future time, when CAT content is reconsidered, and a complete rescaling is done, then the new scale for the speeded tests can be used in its CAT form.

The times on the speeded tests are extremely sensitive to the response mechanism, and to minor aspects of the display. There is every reason to expect the calibration to be equipment-specific. Any change in equipment will require a recalibration of the speeded tests.

The Joint Services Selection and Classification Working Group (JSSCWG), formerly the ASVAB Working Group, is studying the speeded tests carefully. There have been numerous problems with the tests, partly stemming from their sensitivity to time limits, response form, and other administrative details. Computer presentation and recording promises much more accurate measures, although as noted elsewhere, the speeded tests will still be sensitive to response mode.

IMPLEMENTATION PROBLEMS

The most serious problem that we see in instituting Form 100 is in spreading item use across the pool of available items. The present algorithm concentrates item use on the very best items. In fact, unless some changes are made, Form 100 will actually be Form 99, in that the only items being used are the Form 99 subset. Some mechanism such as additional random item selection, or selection dependent only on b value, or temporary and frequent retirement of over-used items, or some more sophisticated technique is needed to provide the claimed improvement in test security. Otherwise the best items of medium difficulty are likely to be presented to nearly all examinees.

A related problem is how to deal with repeat test takers. Using a very restrictive rule for item selection will result in repeating items on repeat test taker tests, unless special steps are taken. The simplest solution is to require all retakes to use the P&P versions, at least while Form 99 is in place. The same problem occurs for Form 100 unless item selection rules are relaxed. Should the system keep track of the particular items given to each applicant? Should the system be capable of omitting a list of items from the item selection table?

One way to meet the repeater problem is to keep the item selection algoritm sufficiently loose that the expected overlap of items for any given repeater is small, and can be ignored. Another solution would be to maintain a series of non-overlapping or nearly non-overlapping item selection tables and to select among these at random for each first-time test-taker, but to record the table identification. This amounts to creating different CAT forms. The equivalence of forms seems assured, although it would have to be checked.

Final decisions about other CAT procedures have not yet been made. A final decision is needed about the stopping rule for tests. The final decision on item selection procedures and the stopping rule for the adaptive test must be made before prototype testing. Or, to be more precise, whatever rules are picked for the prototype tests must be kept in force until Form 100 is established. A change between Form 99 and Form 100 would complicate the equating, and a change during prototype testing would seriously compromise the equating process. It would be best to wait until Form 100 was established before making any structural changes. Any such changes would require a rechecking of the equating.

The above considerations have strong implications for the software. Of course, computer records must be kept for all prototype test-takers, for all CAT test takers during the first year of CAT implementation, and for designated applicants in the operational CAT. Data collection for equating purposes, as well as other analyses, will require that two sets of records be kept for designated applicants - their operational records, and their experimental records.

The software must be able to present additional items in a test, according to some prearranged schedule; possibly dependent upon the applicant's score. By implication, the test scores must be based only on the items in the regular

sequence, not on the extra items. The software must be able to present one of two (or more) forms according to some prearranged schedule. Each form may have its own score calibration tables. For example, during IOT&E for Form 100, both Form 99 and Form 100 will be administered as operational forms.

The software and hardware must be able to record response times, at least to 1/60 second accuracy. (Our impression is that the main cost of a timer is in the registers for its read-out, not in the clock itself, so millisecond processing may represent very little additional outlay.) Future applications are certain to call for timing; there is some probability that 1/60 second may not suffice; one millisecond accuracy would surely suffice.

We have not investigated CATICC plans for eventual system operation. It is clear that a software group will be necessary not only to solve operational problems as they arise, but also to modify the system software for experimental data collection.

There may be some interest in modifying test procedures as CAT progresses. Perhaps a different stopping rule would be better; time limits on the speeded tests may seem wrong; the "verify" button may seem unnecessary; The light pen might work better if its area of sensitivity were changed; the test instructions might be clarified. Computer specialists are likely to want to make such adjustments and may not realize the possible effects on the equated scores. It is thus important to emphasize that any modifications to test procedures may affect the equating, and must be checked before being altered.

CHRONOLOGICAL SUMMARY

1. Before Field Test Phase.

   a. Use item parameters to establish Form 99, 50 items per test.

   b. Set up a schedule for adding items to prototype CAT tests to provide data for item calibration.

2. During Field Testing. For each vendor, test 2500 examinees on an extended CAT, monitoring data collection and making necessary revisions in schedules regularly. These candidates should take Form 9A of the operational ASVAB.

3. After Field Testing. Recalibrate and equate Form 99 for each vendor.

   a. Recalibrate items for Form 99.

   b. Rescore all CAT tests using the new item parameters.

   c. Scale CAT thetas to ASVAB standard scores for all tests via equipercentile equating.

   d. Scale CAT thetas to equivalent P&P raw scores on all tests execpt CS.

   e. For AFQT, VE, and AS, combine equivalent raw scores from the CAT and scale the results to ASVAB AFQT percentile, VE and AS standard scores, respectively, from the P&P form.

   f. Check the scaing.

   g. Scale the occupational composites obtained from the equated standard scores obtained from CAT with the standard scores obtained from the P&P form.

   h. Now Form 99 is ready for operational use.

4. Before Operational Implementation (assuming a between-MEPS design.)

   a. As soon as MEPS have been selected for initial installation of CAT, designate the first four as target MEPS. For each, select two matching MEPS with ASVAB score distributions that closely match the score distributions for the target MEPS.

   b. Obtain complete test data on 2500 cases from each target and matching MEPS.

   c. Plan schedule of data collection for Form 100 item recalibration. That is, determine how many additional items are needed, and schedule the CAT system so that appropriate items are added to each test.

5. Operational implementation - the first six months. Perform IOT&E on Form 99, and collect data for calibrating and equating Form 100.

   a. IOT&E on Form 99. Obtain data on 2500 CAT cases form each target MEPS and 2500 P&P cases from each of the matching MEPS. Use these data and those obtained earlier to make any necessary changes in Form 99 equating tables.

   b. Collect additional data for item recalibration on Form 100 by adding items. Monitor data collection and make necessary adjustments to the data collection schedule regularly.

6. Operational implementation - next five months. Analyze the data for Form 100.

   a. Recalibrate items for Form 100 in the same way as for Form 99, but using ICC curves based on CAT thetas using the best currently available method.

   b. Equate Form 100 thetas to Form 99 thetas via equipercentile equating and make implied changes in Form 99 equating tables, so they will refer to Form 100 thetas.

7. Operational implementation - twelfth month. IOT&E on Form 100. Obtain equivalent samples of at least 2500 cases each for CAT Form 100 and P&P ASVAB to check on final equating of Form 100. Here we recommend using both CAT and P&P within the same MEPS on a pseudo-random schedule.

LIST OF ABBREVIATIONS

AFQT	Armed Forces Qualification Test (a part of the ASVAB)
AFSC	Air Force Specialty Code
ASVAB	Armed Services Vocational Aptitude Battery

The tests in the ASVAB are:

AI	Auto Information
AR	Arithmetic Reasoning
AS	Auto-shop
CS	Coding Speed
EI	Electronic Information
GS	General Science
MC	Mechanical Comprehension
MK	Mathematical Knowledge
NO	Numerical Operations
PC	Paragraph Comprehension
SI	Shop Information
VE	Verbal (a combination of WK and PC)
WK	Word Knowledge

CAT	Computerized adaptive test
IRT	Item response theory
ICC	Item characteristic curve (Same as IRC)
IRC	Item response curve (Same as ICC)

In IRT an ICC is defined with the following symbols:

a	item discriminability
b	item difficulty
c	item pseudo-chance level
θ	theta - the ability of a test-taker
theta	θ

P&P	Paper and pencil (conventional group test)

NPRDC	Navy Personnel Research and Development Center

APPENDIX A

Details of equipercentile equating.

Equipercentile equating assumes the existence of a frequency distribution f(x) that is considered the target and a frequency distribution g(y) that is to be transformed to match f(x) as closely as possible. Note however, that all that can be done is to determine what x-value corresponds to each y-value. The more different values of y and x, the better; a plentitude of y values is especially welcome.

When using a matched group design, the observed P&P distributions should be smoothed, even if they include thousands of cases. A weighting often used was originated by Cureton & Tukey (see Angoff, 1961). Smoothing may best be done on the plot of corresponding percentiles. Many smoothing algorithms have been studied, but none is good enough to displace manual smoothing by human judgment, as described by Angoff(1982). Equating is then just a matter of determining which scores on y yield the same percentile as the scores on x. When equating composites, the CAT composite distributions can be smoothed in the same manner.

The scaling or equating that will be done for CAT uses a one-sample set of data. The same persons took both the P&P test and the CAT test. In this case, a two-way plot of CAT theta vs ASVAB raw or standard scores is possible. When developing a scaling, the theta scale is being adjusted to the P&P derived score, so that after transformation, the plot is as nearly as possible a straight line with a slope of 45 degrees. For natural intervals (or convenient intervals) on the P&P scale, the same number of persons can be marked off on the effectively continuous theta scale. After the function is determined, it should be smoothed, as described above. The ends will be difficult to smooth analytically. Some experts recommend graphical smoothing, othres recommend no smoothing at all. Note that it is not reasonable to smooth each distribution separately, because of the paired nature of the data.

The main advantage of one-sample calibration is that the resulting scatterplot of P&P test scores vs. sclaed CAT scores will show the amount of devaiation from a perfect relationship. The plot will, of course, be linear in the sample, haing been sclaed to be linear, so the relationship must be checked in an independent sample.

APPENDIX B. ITEM CALIBRATION DESIGN FOR CAT X ITEM POOL.

The following table is typical of the design layout. Other tests used fewer booklets, and therefore had more persons per line, but the general plan is similar for other tests.

Data Layout for Arithmetic Reasoning
(Revised 10/19/81)

	(13)	(14)	(15)	(16)	(17)	(18)							
	CAT Booklets						ASVAB Forms						
Group	B1	B2	B3	B4	B5	B6	8A	8B	9A	9B	10A	10B	N
1	X						X						333
2	X							X					333
3	X								X				333
4	X									X			319
5	X										X		333
6	X											X	305
7		X					X						333
8		X						X					333
9		X							X				333
10		X								X			333
11		X									X		333
12		X										X	333
13			X				X						333
14			X					X					333
15			X						X				333
16			X							X			325
17			X								X		319
18			X									X	333
19				X			X						333
20				X				X					333
21				X					X				333
22				X						X			333
23				X							X		333
24				X								X	322
25					X		X						299
26					X			X					333
27					X				X				333
28					X					X			333
29					X						X		333
30					X							X	332
31						X	X						333
32						X		X					333
33						X			X				333
34						X				X			295
35						X					X		294
36						X						X	333
Cases 11=	1956	1998	1976	1987	1963	1921	1964	1998	1998	1938	1945	1958	(11801)
Items 11=	35	35	35	35	35	35	30	30	30	30	30	30	(390)

APPENDIX C. ITEM RECALIBRATION DETAILS.

(Note: Dr. Brad Sympson offered many creative and useful suggestions; his contributions to this section are gratefully acknowledged. Many, but not all, of his suggestions could be included, so the committee bears the responsibility for the section.)

Item parameters will be available from initial item calibration of CAT items (in P&P format) together with items from current P&P Forms (9A, 9B, 10A, 10B, 10X, 10Y). CAT items will have been administered on a voluntary basis, whereas the P&P form will have been administered operationally. Nevertheless it is this calibration that will establish the common scale. The parameter estimates for the P&P ASVAB forms are each based on a small number of cases. However ample data are, or can be, available to get good parameter estimates of the item parameters for each P&P form on its own separate scale. These are the estimates that should be used in further analyses, but they should be put on the common scale by the Lord & Stocking technique (1983). The Lord-Stocking procedure starts with two sets of paramter estimates of the same items, and finds a scale transformation of one set of parameters so that both sets yield as nearly as possible equivalent thetas.

During field tests, data will be collected for CAT items administered on prototype equipment, and for a P&P form administered operationally. We recommend using a single form, such as Form 9A, as the operational form. (At the timeof field testing, P&P Forms 11, 12, & 13 will be in operational use, so special arrangement will be needed for using Form 9A.) Some recalibration methods would not need the old form, whereas others would. We believe that using a form that had been used when the items were first calibrated will provide most flexibility. However any single form could be used. Enough data exist, or can be obtained to get excellent parameter estimates for its items on its own separate scale. But the forms are scaled for equivalence so we can put the items on the common scale by the Lord-Stocking technique. The major problem is to estimate the item parameters for each CAT item under consideration, on this common theta scale.

One approach starts by obtaining the regression of item responses on the CAT theta scale. Here the CAT estimated thetas can be considered fixed, and the item regressions can be obtained either by ordinary logistic regression, by maximum likelihood (ML) estimation, or by some Bayesian method. The thetas estimated by the CAT system are Bayesian estimates; if a different method for estimating parameters is used it would be wise to reestimate the CAT thetas with a comparable method before proceeding. There are only two parameters (a and b) to be estimated, unless the data have been obtained for estimating c. The ML regression is easily obtained by Newton-Raphson iteration.

This procedure has the difficulty that the item response is part of the estimate of theta, so a degree of self-correlation exists. This leads to a steeper curve, and higher a value than would be found if true theta were available. Of course that is always true when thetas and item parameter are estimated jointly, but the self-correlation is small when the number of items is

large. In CAT the theta estimate is based on only a small number of items, (15, at present) the effect is quite large.

One solution to this problem is to estimate theta, for the purposes of this analysis only, from the n-1 (14 in this case) items excluding the item under consideration. The scoring method should be whatever method was used in the original parameter estimation. If, for example, Logist was used, then maximum likelihood scoring should be used.

Instead of basing the candidate's theta estimate solely on the CAT responses, theta could be estimated from responses to both CAT and the P&P test, since the item parameters for these tests are on the same common scale. This would provide a better basis for a score that is being treated as known in the regression analysis.

Another problem arises because the procedure uses estimated thetas. The a values, i. e., the curve steepness, will be reduced by the errors in thetas because these errors will regress the p-values toward the p-values for the average theta scores (assuming a unimodel symmetric theta distribution). In the original parameter estimates, approximately 50 items were used in estimating theta, resulting in a somewhat more accurate estimate. The regression might be raw, or it might be smoothed by taking account of the differential accuracy of the thetas, using procedures of Bock & Aitken, Samejima, or Levine.

The resulting parameters from this regression analysis will be on the common scale. They can be compared with the original parameters, and a decision made about adjustmets to the parameters of some, if any items. If the results indicate a systematic effect of mode of presentation, a systematic adjustment should be made. If a few items show significant departures from the expected ICC, reasons should be sought. It is hoped that few if any items will require change.

For tests where an item by mode interaction is a realistic prospect, enough data will have been collected for a more complete reestimation. This procedure, recommended by Sympson (private communication) would be to use both CAT responses and P&P responses together to estimate both thetas and new CAT item parameters, holding the c parameters constant.) LOGIST-V can do this presumably other estimation programs could do it as well. This procedure has the advantage of using all the data to estimate thetas. t The resulting parameters will be on an arbitrary scale, and wil have to put on the common scale by a Lord-Stocking adjustment, based only on the P&P item parameters. An alternative would be to hold the P&P parameters constant in the analysis, which can be done in LOGIST, but Lord (private communication) reports poorer results with this method.

After the Lord-Stocking transformation, the old and new versions of the CAT parameters should be compared to determine which item parameters need to be changed. It may seem best to use the new parameters entirely, or it may seem better to change only some item parameters. Every effort should be made to understand the reason for the difference. The actual items should be examined, both in the original booklets and on the screen, to look for any anomalies.

The above procedures are designed for separate analyses of data from each prototype field test, or fro the single prototype, if only one is tested. Brad Sympson has pointed out that the item parameter estimates can be markedly improved by pooling data across prototypes whenever that is justified by separate analyses.

The plan assumes that many items will seem to be equivalent on the two systems, in which case there is merit in combining the data from the two field testings. The plan recognizes the possibility that for other items, system differences are possible. Two sets of parameters are obtained for the items that are system-specific, but only one set of parameters is obtained for each item that is not system-specific. Two successive analyses of the field-test data are contemplated - one to identify the system-specific items, and one to estimate parameters. Tables C1 and C2 show the proposed plans.

This plan takes advantage of the fact that two sets of comparable data will be collected, and provides a way whenever possible, of combining this data. We note that this analysis can be done as well as the separate analyses proposed in the main report. It would be valuable to do both analyses for whatever light each can shed about the other. .

The plan shown uses two P&P forms, one of which was used in the original CAT O-pool calibration, and one in current use. An alternative preferred by the committee is to use only one P&P form, in order to have better data for later steps in the scaling process. However, operational requirements may force the use of all current forms of the P&P test. In any case the P&P test would be the same for both systems, and thus would form the link to get all parameters on a common scale. The Lord-Stocking procedure can be used to make this the common scale.

The great advantage of this analysis is that we have every reason to expect most items to be the same on the two systems, which means that twice as much data are avaiable for recalibrating the items.

Table from Sympsons letter.

Form 100 recalibration.: Recalibrating the Form 100 items can be done online either using new on-line alibration procedures under development, or using the item ICC regression methods, discussed above, that assume theta fixed. There is no need to recalibrate the Form 99 items, unless of course, some problem is detected in Form 99 IOT&E. The additional items will then be on the common scale. The operational Form 99 scale calibrations can be used for Form 100 unless Form 100 IOT&E detects a problem. Form 100 will have less error and be more secure than Form 99, because of the larger item pool, but it will be scaled to be equivalent to Form 99 automatically.

APPENDIX D. HOW MANY ADDED ITEMS?

Current data can provide a rough idea of the number of item responses that will be needed in addition to those that happen normally in CAT. Here only the data from the AR test in the experimental system are considered, for a sample of 1382 Navy recruits. The item response frequencies are shown in Table D-1 as a function of b-value for all items in the experimental AR pool. Column 1 of Table 2 shows Total responses, grouped by b interval and rounded. Column 2 simulates the addition of an equal number of tests from lee qualified candidates by reversing the distribution of responses. Relatively few item responses were obtained from recruits for items with b values less than 0. For the poorer candidates, we might expect a reverse skew. The two frequencies have been added, scaled back proportionally to yield 2500 total candidates, and rounded, yielding Column 4. An assumed distribution of b values is given in Column 5. Additional responses needed to obtain 500 per item and 1000 per item are shown in Columns 6 and 7.

To fill out to 500 per item, 3400 item responses are needed; 2 per test yields 5000, which provides some leeway for the various errors in this rough analysis. For 1000 repsponses per item we need 14600 additional responses; 15000 can be obtained by adding 6 items per test.

These estimates are very rough. Better estimates can be obtained with simulation, and should be pursued, but the general pattern is not likely to change.

Table D-1 Distribution of item responses by b-value.

b interval	frequency 0	1-99	100-199	200-299	300-399	400-499	500-599	600-699	700-899
below -1.5	12	4	1						
-1.5	3	2							
-1.3	8	1	1						
-1.1	4	1	2						
-0.9	4	1	1						
-0.7	6	1	1	4		1			
-0.5	7	1	1	1	3	1			
-0.3	10	1	1			2		1	
-0.1	9				1		1	1	
0.1	13	2	1	1		1	2	2	
0.3	16	1		1		2		1	1
0.5	14					1		1	3
0.7	6				1	1		1	2
0.9	1	3	1				1	1	2
1.1			2	1	1	1		2	
1.3			2	2					
1.5			1						
above		1		1					

REFERENCES

Allred, L.A. & Green, B.F. Analysis of Experimental CAT ASVAB test data. Department of Psychology, the Johns Hopkins University, Baltimore, MD, 21218, January, 1984.

Angoff, W. H. Scales, norms, and equivalent scores. In Thorndike, R.L. (Ed.) Educational Measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Angoff, W.H. Summary and derivation of equating methods used at ETS. In Holland, P.W. & Rubin, D.B. (Eds.) Test equating. New York: Academic Press, 1982.

ASVAB Working Group. History of the Armed Services Vocational Aptitude Battery (ASVAB). Washington, DC: Office of the Assistant Secretary of Defense, (Manpower Reserve Affairs and Logistics.) March, 1980.

Binet, A. Les idees modernes sur les infants. Paris: Ernest Flamorion, 1909.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958.

Birnbaum, A. Some latent trait models and their uses in inferrring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), Statistial Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

Bock, R.D., & Aiken, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Bock, R.D., & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Bock, R.D., & Mislevy, R.J. Data quality analysis of the Armed Services Vocational Aptitude Battery. Chicago: National Opinion Research Center, August, 1981.

Department of Defense. Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics.) Washington, DC, March 1982.

Greaud, V.A. & Green, B.F. Analysis of speeded test data from experimental CAT system. Dept. of Psychology, Johns Hopkins University, Baltimore, MD 21218, Jan., 1984.

Green, B.F. Adaptive Testing by Computer. In Ekstrom, R.B. (Ed)., Measurement, Technology, and Individuality in Education. New Directions for Testing and Measurement No. 17. San Francisco: Jossey-Bass, 1983a.

Green, B.F. The promise of tailored tests. In Wainer, H. & Messick, S.A. (Eds.), Principles of Modern Psychological Measurement. A Festschrift in Honor of Frederic Lord. Hillsdale, NJ: Erlbaum, 1983b.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.B., & Reckase, M.D. Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery. Department of Psychology, The Johns Hopkins University, Baltimore MD, 21218, May 15, 1982.

Hambleton, R.K., & Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 38, 75-96.

Harman, H., Helm, C.E., & Loye, D.E. (Eds.), Computer-assisted testing. Princeton, NJ: Educational Testing Service, 1968.

Hendrix, W.H., Ward, J.H., Jr., Pince, M., Jr., & Harvey, D.L. Pre-enlistment person-job match system. AFHRL-TR-79-29, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, September, 1979.

Holland, P.W. & Rubin, D.B. (Eds.) Test Equating. N.Y.: Academic Press, 1982.

Holtzman, W.H. (Ed.), Computer-assisted Instruction, Testing, and Guidance. New York: Harper & Row, 1970.

Hulin, C.L., Drasgow, F., & Parsons, C.K. Item Response Theory. Application to Psychological Measurement. Homewood, Ill.:Dow-Jones Irwin, 1983.

Jaeger, R.M., Linn, R.L., & Novick, M.R. A review and analysis of score calibration for the Armed Services Vocational Aptitude Battery. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics), June, 1980.

Kreitzberg, C.B., & Jones, D.H. An empirical study of the Broad-Range Tailored Test of Verbal Ability. RR-80-5, Educational Testing Service, Princeton, NJ, May, 1980.

Lawley, D.N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh. Series A, 1943, 61, 273-287.

Lord, F.M. A theory of test scores. Psychometrika Monograph #7, 1952.

Lord, F.M. A broad-range tailored test of verbal ability. Applied *Psychological Measurement*, 1977, 1, 95-100.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F.M. The standard error of equipercentile equating. RR-81-48, Educational Testing Service, Princeton, NJ, November, 1981a.

Lord, F.M. Standard error of an equating by item response theory. RR-81-49, Educational Testing Service, Princeton, NJ, November, 1981b.

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

Maier, M.H., & Grafton, F.C. Aptitude composites for ASVAB 8, 9, and 10. U.S. Army Research Institute, Alexandria, VA, May, 1981b.

Maier, M.H., & Grafton, F.C. Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8AX. Research Report 1301, U.S. Army Research Institute, Alexandria, VA, January, 1981.

Maier, M.H., & Truss, A.R. Original Scaling of ASVAB Forms 5/6/7: What went wrong. Research Report CRC-457, Center for Naval Analyses, Alexandria VA 22311, March 1983.

McBride, J.R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J.R. Adaptive verbal ability testing in a military setting. In Weiss, D.J. (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Department of Psychology, University of Minnesota, Minneapolis, MN, September, 1980.

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

Rasch, G. Probabilistic models for some intelligence and attainment tests. Dansmarks Paedogogiske Institute, Copenhagen, Denmark, 1960.

Reckase, M.D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study. Catalog of Selected Documents in Psychology, 1978, 8, 71. Ms. 1737.

Ree, M.J. The effects of item calibration sample size and item pool size on adaptive testing. Applied Psychological Measurement, 1981, 5, 11-19.

Ree, M.J., Mathews, J.J., Mullin, C.J., & Massey, R.H., Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10. AFHRL-TR-81-49, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, February, 1982.

Ree, M.J., Mullins, C.J., Mathews, J.J., & Massey, R.H. Armed Services Vocational Aptitude Battery: Item and Factor Analysis of Forms 8, 9, and 10. AFHRL-TR-81-55, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, March 1982.

Rubin, D.B. Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 1980, 75, 801-816.

Sacher, J. & Fletcher, J.D. Administering paper-and-pencil tests by computer, or the medium is not always the message. In Weiss, D.J. Proceedings of the 1977 Computerized Adaptive Testing Conference. Department of Psychology, University of Minnesota, Minneapolis, MN 55455, July, 1978.

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

Stocking, M.L. & Lord, F.M. Developing a common metric in item response theory.Applied Psychological Measurement, 1983, 7, 201-210.

Sympson, J.B., Weiss, D.J., & Ree, M.J. Predictive validity of conventional and adaptive tests in an Air Force training environment. AF HRL-TR-81-40. Air Forces Human Resources Laboratory, Brooks Air Force Base, TX, March, 1982.

Tucker, L.R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Urry, V.W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Urry, V.W. Tailored testing, its theory and practice. Part II: Ability and item parameter estimation, multiple ability application, and allied procedures. NPRDC TR 81 Navy Personnel Research and Development Center, San Diego, CA, November, 1981.

Urry, V.W., & Dorans, N.J. Tailored testing, its theory and practice. Part I: The basic model the normal ogive submodels, and the tailored testing algorithms. NPRDC TR 83-00, Navy Personnel Research and Development Center, San Diego, CA, April, 1983.

Warm, T.A. A primer of item response theory. Technical Report 940279, U.S. Coast Guard Institute, Oklahoma City, OK, December, 1978.

Weiss, D.J. Strategies of adaptive ability measurement. Res. Rep. 74-5. Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1974.

Weiss, D.J. (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Department of Psychology, University of Minnesota, Minneapolis, MN, 1978

Weiss, D.J. Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement. Research Report 79-6, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, MN, 1979.

Weiss, D.J. (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Department of Psychology, University of Minnesota, Minneapolis, MN, 1980

Weiss, D.J. (Ed.) New Horizons in Testing. Latent trait test theory and computerized adaptive testing. N.Y.: Academic Press, 1983.

Wingersky, M.S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In Hambleton, R.K. (Ed.) ERIBC Monograph on applications of item response theory. Vancouver, B.C.: Educational Reseach Institute of British Columbia, 1983.

Wood, R. Response-contigent testing. Review of Educational Research, 1973, 43, 529-544.

Yen, W.M. Using simulation results to choose a latent trait model. Applied Psychological Measurement, 1981, 5, 345-362.

END

FILMED

12-85

DTIC