DOCUMENT RESUME

ED 222 532 TM 820 683

AUTHOR	Ree, Malcolm James; And Others
TITLE	Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10.
INSTITUTION	Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO	AFHRL-TR-81-49
PUB DATE	Feb 82
NOTE	19p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Aptitude Tests; *Armed Forces; Enlisted Personnel; *Equated Scores; Test Reliability
IDENTIFIERS	*Armed Forces Qualification Test; *Armed Services Vocational Aptitude Battery; Calibration

ABSTRACT

A calibration of the Armed Forces Qualification Test (AFQT) composite of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8a, 8b, 9a, 9b, 10a, and 10b to the metric of the AFQT Form 7a (AFQT-7a) and a comparison of these outcomes to the operational calibration tables implemented 1 October 1980 are presented. A sample of applicants for military enlistment was administered one form of ASVAB and the AFQT-7a in counterbalanced order. For analytic purposes, an edited sample (15,115 males) was separated into 6 samples based on the 6 forms of ASVAB administered. Data were collected at 20 geographically dispersed Armed Forces Examining and Entrance Stations on the 6 forms of ASVAB and the AFQT-7a. Each of the six samples was edited and scored, and descriptive statistics were computed. The root-mean-square and average absolute deviation measures, investigating the similarity of the equated scores across the forms, showed only small differences among the operational table and tables developed during this study. Forms 8, 9, and 10 of ASVAB were found to be parallel when equated to AFQT-7a, and a single conversion table was deemed appropriate for operational enlistment processing. (Author/PN)

AFHRL-TR-81-49

AIR FORCE HUMAN RESOURCES LABORATORY

CALIBRATION OF ARMED SERVICES VOCATIONAL APTITUDE BATTERY FORMS 8, 9, AND 10

By

Malcolm James Ree
John J. Mathews
Cecil J. Mullins
Randy H. Massey, Capt, USAF

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

February 1982

Interim Report for Period October 1980 — July 1981

Approved for public release: distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

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

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

This report has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-81-49
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): CALIBRATION OF ARMED SERVICES VOCATIONAL APTITUDE BATTERY FORMS 8, 9, AND 10
5. TYPE OF REPORT & PERIOD COVERED: Interim, October 1980 - July 1981
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Malcolm James Ree, John J. Mathews, Cecil J. Mullins, Randy H. Massey
8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Manpower and Personnel Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 77191804
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235
12. REPORT DATE: February 1982
13. NUMBER OF PAGES: 18
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.

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

18. SUPPLEMENTARY NOTES: SM Study Number 7728

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

ability testing
AFQT
aptitude tests
ASVAB
calibration
curve smoothing
equating
equipercentile
polynomial regression

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

The objective was to calibrate the Armed Forces Qualification Test (AFQT) composite of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8a, 8b, 9a, 9b, 10a, and 10b to the metric of the AFQT Form 7a (AFQT-7a) and to compare these outcomes to the operational calibration tables implemented 1 October 1980. A sample of applicants for military enlistment was administered one form of ASVAB and the AFQT-7a in counterbalanced order. From this target sample of 22,400, a "males only" sample of 15,115 was developed through data editing techniques designed to exclude females and cases with incomplete or unusable data. For analytic purposes, this edited sample was separated into six samples based on the six forms of ASVAB administered. Data were collected at 20 geographically dispersed Armed Forces Examining and Entrance Stations (AFEES) on the six forms of ASVAB and the AFQT-7a.

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 (Continued):

Each of the six males only samples was edited and scored, and descriptive statistics were computed. Percentiles for both the ASVAB and the AFQT-7a were equated and smoothed by a polynomial regression procedure. Each sample was split in half, and the equating and smoothing were repeated on each half sample. Since results were consistent among the large sample and the two half samples, they were accepted. In order to investigate the similarity of the equated scores across the forms, root-mean-square (RMS) and average absolute deviation (AAD) measures were computed between the various equating tables. A comparison of the forms found them to be equivalent when they were equated to AFQT-7a. The RMS and AAD measures showed only small differences among the operational table and tables developed during this study. Forms 8, 9, and 10 of ASVAB were found to be parallel when equated to AFQT-7a, and a single conversion table was deemed appropriate for operational enlistment processing.

PREFACE

This study was completed under the auspices of Personnel Qualification Systems which is part of a larger effort in Force Acquisition and Distribution. It was subsumed under project 77191804, "Maintenance and Improvement of Enlisted Selection and Classification Tests" and executed as part of the responsibility of the Air Force Human Resources Laboratory (AFHRL) as lead laboratory under the executive agent (Air Force) for Armed Services Vocational Aptitude Battery research and development.

An effort such as this, although under the direction of an individual, can be accomplished only through a team effort. The authors wish to express their appreciation to Roy Chollman, James Earles, A1C Jenny Hodge, and A1C Gerald Yates. A debt of gratitude is owed to Doris Black, who served to condense and translate the analytic requests into operational procedure for the Technical Services Division. Henry Clark wrought minor magic by convincing the computer to produce analyses prior to established due dates.

The authors also wish to express their appreciation to Jacobina Skinner and the other members of Publication Review Panel 2 for helpful comments on an earlier draft of this manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CALIBRATION OF ARMED SERVICES VOCATIONAL APTITUDE BATTERY FORMS 8, 9, AND 10

I. INTRODUCTION

The measurement of human characteristics has been a necessary part of selection and classification for military occupations for over 60 years. Like measurement of physical characteristics, such as length, weight, or density, no natural units of measure exist for psychological characteristics; rather, artificial units are established by consensus. One of the most frequently used units of measurement for human characteristics is the percentile equivalent. The percentile is reported in reference to some standard population or group. Ability tests used for military selection and classification are usually referenced to the 1944 mobilization base, and this is usually accomplished by equating new tests to old tests. Equating is the conversion of score units of one test to the score units of another test. The current study describes the referencing of Forms 8a, 8b, 9a, 9b, 10a, and 10b of the Armed Services Vocational Aptitude Battery (ASVAB) to the mobilization base metric, through the use of an anchor test.

There are two important reasons why current tests are equated to past tests. The first is to enable the testing agency to report on the relative distribution of scores on a year-to-year basis in a common metric. For example, the various military services like to be able to compare current accessions to past accessions on the same scale. The second reason is to provide a consistent meaning for cutting scores for selection and classification tests. In theory, a score for the new test at the 80th percentile can be said to be equivalent to a score at the 80th percentile on the past tests, and this equivalence becomes the definition of consistency.

When several forms of a test are to be operational simultaneously, it is an advantage if they are parallel, which allows the use of a single equating table. Gulliksen (1950) offers a definition of parallel tests which includes sameness of factor structure, equality of means, equality of variances, and equality of non-zero correlations with an external criterion. It also seems reasonable to include equivalence of skew and kurtosis (Ree, 1977), the third and fourth moments of the distribution, although little research exists in the area.

Parallel tests may be constructed by assigning items randomly to forms. This method is usually called "Randomly Parallel Forms." Or items may be matched on difficulty and/or discrimination, stratified, and then assigned randomly to one of a set of multiple forms. This procedure is called "Stratified Parallel." Analytic methods of constructing parallel forms also exist (Ree, 1976), but they tend to be intensive of computer time.

Using the Stratified Parallel method, Forms 8, 9, and 10 of the ASVAB were constructed to be parallel in terms of raw scores so that a single table might be used to convert raw scores on any of the six forms to percentile equivalents. The objective of this study was to determine if a single table were appropriate.

Calibration of Tests

Because two or more forms of a test can never be made precisely equivalent in range and level, it is necessary to render the forms interchangeable by equating. The equating procedure may be defined (Flanagan, 1951; Angoff, 1971) as converting the scoring units of one test to the scoring units of another.

In general, two procedures have been in common use: linear and equipercentile equating. Linear equating requires that equivalent Z-score transformations of the two tests represent the same cumulative proportion. Said differently, the shapes of score distributions should differ only trivially. Equipercentile equating, on the other hand, makes no such assumption of Z-score equivalence. The linear method offers the advantage of dealing with analytic statistics (means, standard deviations, etc.) which are verifiable. Equipercentile equating is preferable when the distributions differ and is often offered as the definition of equating (Jaeger, 1981). It should be noted that the linear and equipercentile approach coincide when both the distributions to which they are applied have the same shape.

Angoff (1971) uses the term "calibration" to describe the equating of tests of differing abilities. For example, the equating of a test of Word Knowledge to a test of Reading would be called "calibration." Therefore, it is appropriate to say that military selection and classification tests have been calibrated rather than equated. Angoff is somewhat critical of the calibration technique because a problem arises from the nature of calibration. It is repeatedly stated in the literature (Angoff, 1971; Flanagan, 1951; Jaeger, 1981) that calibrating does not lead to sample-unique solutions, as does equating, although empirical evidence is not offered. The non-uniqueness of the solution makes difficult the interpretation of several calibrations of the same test, or parallel forms of the test. Military selection and classification tests have frequently been calibrated, rather than equated. Form 8a of the ASVAB was linked via calibration to an anchor test using several differing subject groups ranging from high school students to new military recruits. The effects of calibrating, as opposed to equating, require further study in order to understand fully the consequences of the technique.

Three previous studies (Boldt, 1980; Maier & Grafton, 1981; Sims & Truss, 1980) were conducted which calibrated Form 8a to Armed Forces Qualification Test Form 7a (AFQT-7a). Because ASVAB Forms 8, 9, and 10 were constructed to be parallel by the method described previously as "Stratified Parallel Forms," it was reasoned that calibrating one form was tantamount to calibrating all forms. That is, because calibration sets raw scores of the calibrated test equivalent to raw scores on an anchor or target test, and because the raw scores of the six forms were constructed to be equivalent, then any one form may be calibrated, and the results should then be applicable to all the other forms. The crucial requirement is that the forms be parallel. If they are not, separate calibrations are required. The present study seeks to verify the results of the earlier calibration studies which produced the tables implemented 1 October 1980. These are referred to as the operational tables.

In order to determine if the assumptions underlying the procedures for calibrating ASVAB-8a and thereby Forms 8b, 9a, 9b, 10a, and 10b were acceptable, an Initial Operational Test and Evaluation (IOT&E) was undertaken. The IOT&E was begun shortly after the test was put into operation for selection and classification of candidates for military enlistment.

II. METHOD

The Tests

Forms 8, 9, and 10 of the ASVAB are multiple aptitude batteries comprised of 10 subtests. Eight of the subtests are power subtests, while two are speeded subtests. Table 1 shows the name, the number of items, and whether the subtest is power or speeded. These forms differ from the previous ASVAB forms by the inclusion of Paragraph Comprehension (PC) and Coding Speed (CS) subtests, by the combination of Automotive Information and Shop Information into a single subtest (AS), and by the deletion of subtests measuring Space Perception, Attention to Detail, and General Information. The overall administration time for any of the forms is about 180 minutes, and in operation, the test is answered on a machine scannable answer sheet.

Table 1. **Name and Number of Items for Power and Speeded ASVAB Subtests in Forms 8, 9, and 10**

Name	Number of Items	Power/Speed
General Science (GS)	25	Power
Word Knowledge (WK)	35	Power
Arithmetic Reasoning (AR)	30	Power
Paragraph Comprehension (PC)	15	Power
Numerical Operations (NO)	50	Speed
Coding Speed (CS)	84	Speed
Auto-Shop Information (AS)	25	Power
Mathematics Knowledge (MK)	25	Power
Mechanical Comprehension (MC)	25	Power
Electronics Information (EI)	20	Power

The Armed Forces Qualification Test (AFQT) composite is used for military enlistment qualification and is comprised of PC, Word Knowledge (WK), Arithmetic Reasoning (AR), and Numerical Operations (NO) subtests. All subtests are unit weighted except for NO, which is weighted by one-half.

The AFQT-7a served as the anchor test. This test was previously used for enlistment qualification but has been inactive for several years. It was chosen as the anchor test because its content is close to that of the test used in the 1944 mobilization base development testing. It is not believed to be compromised, and an earlier form (Form 3) of the ASVAB was calibrated against it.

The AFQT-7a has 100 items evenly distributed in the ability areas of WK, AR, Boxes (B), and Tool Knowledge (TK). The first two, WK and AR, are similar to the like-named subtests in the current AFQT portion of the ASVAB. The latter two, B and TK, are not found in the current AFQT portions of the ASVAB. It is the disparity in the ability areas measured which leads to labeling the equating effort a "calibration" and which leads to the problem of non-unique solutions.

Administration of Tests to Subjects

A sample of subjects was drawn to provide for equal geographical representation. Data collection took place in 20 Armed Forces Examining and Entrance Stations (AFEESs). Table 2 shows the locations of the AFEESs and the number of subjects at each. Each subject took the AFQT-7a and one form of the ASVAB, which was used for qualification for military enlistment. The AFQT-7a was administered on a separate answer sheet. The ASVAB and AFQT-7a tests were administered in counterbalanced order by reversing order of their administration each day from that employed the previous day. Tests were also administered at locations affiliated with the AFEES, called Mobile Examining Team (MET) sites and Office of Personnel Management (OPM) sites.

Table 2. **AFEES Sites and Sample at Sites[a]**

AFEES	Subjects
Chicago	1,500
Cleveland	1,300
Atlanta	800
Baltimore	1,600
Boston	1,300
Jacksonville	1,400
Los Angeles	2,600
Montgomery	900
Newark	1,400
Philadelphia	1,400
Richmond	1,200
St. Louis	1,400
Spokane	500
Denver	600
Houston	600
Phoenix	500
Portland	400
San Diego	600
Minneapolis	1,200
Omaha	1,200
Total	22,400

[a]Sites included AFEES, MET, and OPM locations for test administration.

Data Editing

All answer sheets were visually inspected for completeness of information and stray marks. The ASVAB uses a three-part answer sheet which is optically scannable and has precoded numbers on each sheet to keep the triplet set intact during operational scanning. There is also an optically scannable social security account number (SSAN) grid. These operational ASVAB answer sheets which had been scanned at AFEES were then rescanned and the required triplets of answer sheets were merged. The AFQT-7a answer sheets were also scanned and merged with the records of the ASVAB for each subject. Because only males were represented in the World War II (1944) mobilization base, female subjects were deleted from the original sample to leave a "males only" sample of applicants.

Three other editing procedures were employed. First, to determine if the correct form of the test (8a to 10b) was specified on the answer sheet, a check was performed by scoring the first four items in the NO, CS, and WK subtests. Twelve items in all were scored. The NO and CS are speeded subtests, and the WK subtest has the easiest items first. It was reasoned that any examinee's score of 6 or less was suspect and should be examined further. This was accomplished by applying each of the six form-specific scoring keys for these 12 items to the answer sheet and comparing the magnitude of the scores from the various key sets. For example, if the subject coded "Form 8a" on the answer sheet and obtained a score of 2 from the Form 8a key, but when scored on the Form 10b key obtained a score of 11, then the entire test was scored using the 10b scoring key. If, on the other hand, low scores were found for all forms, then the key for the form indicated by the examinee was retained.

The second data editing procedure was designed to see if differences existed among types of testing sites: AFEES, MET, and OPM. This was accomplished by inspecting the mean and standard deviation of the absolute differences, by type of test site, between the scores on the AFQT-7a and the AFQT portion of the six forms of the ASVAB. Systematic deviance in a type of testing site would indicate that data from that kind of site should be discarded.

The third and final check was to investigate the bivariate scatter plots and standardized residuals devolved from regressing scores for each ASVAB-AFQT on scores on AFQT-7a, scores on each AR on Math Knowledge (MK), and scores on each NO on CS. These three sets of variables allow investigation of consistency of responding between the first and second halves of the ASVAB for both power and speeded tests as well as between a test actually used for military enlistment qualification (ASVAB) and a test (AFQT) given for equating purposes only. Each pair of variables is highly correlated. Examinees with standardized residuals outside of the range of $\pm$ 2.50 were identified for further scrutiny. They were located on the appropriate scatter plot and were deleted if it was reasonably clear from visual inspection that they represented true outliers by being substantially away from the bulk of the scatter.

Sample

From the original sample collected at the AFEES, MET, and OPM sites, females and those who failed the data editing were removed. Six male-only samples were created based on the form of ASVAB administered. Random half-samples were selected within each of the six male-only samples created for Forms 8a through 10b. These half-samples were established in order to cross-validate results and to investigate consistency of various estimates made in the equating process.

Equipercentile Equating and Calibrating

It is appropriate to specify that Forms 8, 9, and 10 of ASVAB were calibrated using AFQT-7a as a standard. The plan identified as "Design II" by Angoff (1971) was used for each pair of composites to be calibrated.

Test calibration was accomplished using raw scores on the ASVAB-AFQT and on the AFQT-7a as a starting point. For each raw score distribution of ASVAB-AFQT and AFQT-7a, sample dependent percentiles from 1 to 99 were computed in unit intervals. This is essentially a raw score to raw score procedure. Previous equatings using ASVAB-AFQT raw score to AFQT-7a percentile equivalents only, rather than ASVAB-AFQT raw score to AFQT-7a raw score were deemed insufficient, as information was lost when raw score point intervals were collapsed. The raw score to raw score procedure was used because it is more widely accepted and more efficient. After the raw score equivalents were established, it was necessary to smooth the resulting line. This smoothing was accomplished by using the analytic procedure of polynomial regressions up to the third order. The fit of the regression was used to determine the best curve.

The creation of half-samples was especially useful in determining the relative stability of the quadratic and cubic regression weights. Each smoothing was accomplished three times, and the weights were retained only if they remained relatively constant. The cases in which higher order weights did not remain constant were smoothed by the first order polynomial, as it always remained constant. Table 3 provides an instructive example using invented data. The samples 1 through 3 on the left show instances where the weights (W_i) are stable and thus are acceptable to smooth the equating line. The fourth, fifth, and sixth samples show an instability of weights due to capitalization on chance fluctuation, which causes the high order polynomials to be rejected. Note how the values in the columns marked "W2" and "W3" fluctuate in these later samples but not in samples 1 through 3. This kind of instability of weights should be the basis for rejection of the polynomial. Note also how the standard error of estimate (SEE) decreases substantially as the higher order terms are entered in samples 1 to 3, but not in samples 4 through 6. This consistency and reduction of SEE is indicative of a better fit. Three additional points are worthy of note. First, the R^2 is observed to change only in the trivial third decimal place, and little emphasis should be placed on it. Secondly, the standard error of estimate is appropriate for determining fit. Finally, care must be exercised not to interpret the R and R^2 as correlations between raw scores for subjects. These indexes reflect the covariation of the equated percentile points in a distribution and must be expected to be quite high. One advantage of this method of smoothing is that it is analytic and reproducible, thereby avoiding the myriad pitfalls of hand smoothing.

Table 3. Example of Smoothing by Polynomial

Sample	Type	R^2	W1	W2	W3	SEE
			Composite 1			
1	Full	.9987	1.045			2.618
1		.9999	.981	.056		1.072
1		.9999	.970	.049	.051	.674
2	Half	.9985	1.050			3.012
2		.9999	.980	.060		1.401
2		.9999	.970	.051	.050	.801
3	Half	.9989	1.055			3.000
3		.9999	.980	.058		1.300
3		.9999	.971	.049	.052	.790
			Composite 2			
4	Full	.9999	1.061			2.710
4		.9999	.982	.311		2.600
4		.9999	.961	.032	.202	1.930
5	Half	.9981	1.059			2.950
5		.9999	.931	.103		2.710
5		.9999	.929	.009	.001	2.070
6	Half	.9992	1.072			2.870
6		.9999	.901	.081		2.650
6		.9999	.918	.050	.400	1.800

Table Generation

The ultimate goal of this effort is to produce tables for each ASVAB AFQT composite from Forms 8a through 10b and to determine if a single table for each composite is applicable across the set of six forms. The tables were generated by picking the appropriate smooth curve form and evaluating it at each raw score point for the range of the AFQT composite. This yielded six equating tables, one for each ASVAB form. An average table was created from these six. Several deviation indexes were computed to make comparisons among these tables and the operational table. These indexes were the root-mean-square (RMS) deviation and average absolute deviation (AAD). Additionally, the similarity between classification into mental categories (see Grunzke, Guinn, & Stauffer, 1970) by the operational table and the six form-specific tables was investigated by computing a two-way frequency table of classification.

III. RESULTS AND DISCUSSION

Data Editing

The check to determine if the correct form (8a through 10b) was coded produced 427 subjects requiring scrutiny. Table 4 shows the number of cases, by form, which were identified for verification. For all the forms, 32 cases were deleted, 51 had form changes, and 344 were left unchanged.

Table 4. **Number of Subjects Flagged by Key Verification by Test Form**

		Subjects	
Form	Total	Not Key Flagged	Key Flagged
8a	2650	2561	89
8b	2529	2477	52
9a	2625	2549	76
9b	2527	2467	60
10a	2510	2429	81
10b	2438	2369	69

By way of example, four cases displayed in Table 5 are instructive. Case 1 was changed to 8b because of the low score on 8a compared to the high score on 8b. Case 2 was deleted because having a one or zero on all scoring keys indicated the examinee was unlikely to have been trying very hard. Case 3 was deleted because it was impossible to determine which test the examinee was administered, as the form coded on the answer sheet had the lowest score of the six. Case 4 was kept, despite the low scores, since the score for the form coded was the highest.

Table 5. **Example Cases from Key Verification Procedures**

		Scores for Forms					
Case	Form Coded	8a	8b	9a	9b	10a	10b
1	8a	4	10	2	1	3	5
2	8b	1	0	0	1	1	0
3	9a	3	4	2	3	4	3
4	9b	1	4	3	6	3	1

The second data editing procedure of investigating differences among types of testing sites by comparison of absolute differences on AFQT-7a and ASVAB-AFQT revealed no systematic differences. Consequently, all site types were deemed appropriate for inclusion in the study.

The third and final check was to investigate the bivariate scatter plots and standardized residuals devolved from regressing scores for each ASVAB-AFQT on scores on AFQT-7a, scores on each AR on MK, and scores on each

NO on CS. Examinees with standardized residuals outside of the range of $\pm$ 2.50 were identified for further scrutiny. Each was located on an appropriate scatter plot, and the score was deleted if it was clear that the examinee represented a true outlier by being substantially away from the bulk of the scatter. It was observed, for example, that some examinees displayed high scores on AR but very low scores on MK. This is an illogical situation that might be accounted for by having obtained some answers for the AR subtest, which is in the qualification portion of the ASVAB, but not for MK. It might also be an indication of faltering motivation on the later MK test. In either case, the examinee should not be in the sample. Figure 1 shows this condition. The observations within the dotted boundaries were subject to scrutiny and potential deletion. Only 132 subjects were removed during this procedure. The final sample was comprised of 15,115 male subjects.

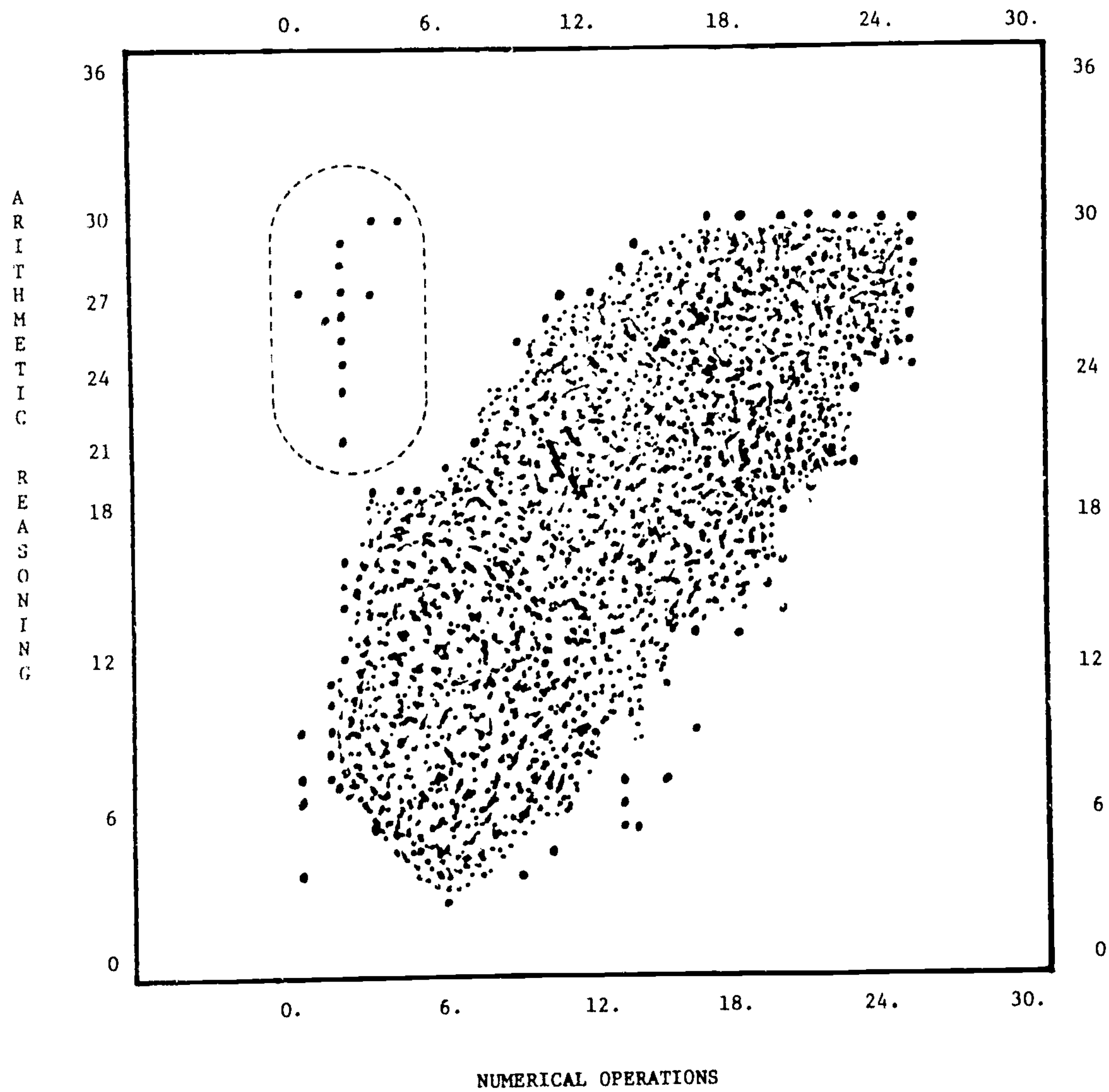

Figure 1. **Scatter Plot of Arithmetic Reasoning and Numerical Operations Test Scores.**

Analysis Samples

Table 6 displays the sample sizes for each of the six male-only samples.

Table 6. **Number of Subjects by ASVAB Form**

Form	Number of Subjects
8a	2,621
8b	2,506
9a	2,587
9b	2,500
10a	2,484
10b	2,417

Descriptive Statistics

Table 7 shows the descriptive statistics for each of the ASVAB subtests, the ASVAB-AFQT, and AFQT-7a. As can be seen, the means ($\bar{X}$) differ relatively little, as do the standard deviations (σ). Cumulative frequency distributions of the scores are of the same general shape with few differences among them.

Table 7. **Descriptive Statistics for ASVAB 8, 9, and 10 and AFQT-7a**

	ASVAB Form Administered											
	8a		8b		9a		9b		10a		10b	
Sub-Test	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
GS	15.29	4.83	15.10	4.92	14.61	5.51	14.59	5.54	14.66	5.09	14.74	5.15
AR	16.47	6.76	17.13	7.13	16.92	6.96	17.28	6.86	17.93	6.70	17.09	6.98
WK	24.64	7.55	23.44	7.56	23.53	7.66	23.72	7.75	22.99	7.82	23.43	7.60
PC	10.08	3.38	9.84	3.34	9.27	3.48	10.02	3.28	9.59	3.77	10.02	3.17
NO	34.52	10.17	34.75	10.05	34.29	10.58	33.93	10.40	35.03	10.04	34.58	10.36
CS	41.29	15.04	41.27	15.23	41.42	15.05	41.79	14.53	42.34	14.84	42.08	14.42
AS	15.25	5.82	15.24	5.76	15.77	5.77	15.74	5.71	15.77	5.65	15.83	5.66
MK	11.32	5.54	11.14	5.43	11.24	5.46	11.20	5.60	12.33	5.33	12.35	5.56
MC	14.44	5.43	14.14	5.41	14.28	5.33	14.32	5.07	14.45	5.25	14.27	5.20
EI	11.50	4.31	11.46	4.29	11.94	4.13	12.05	3.98	12.06	4.03	11.75	4.03
VE	34.72	10.45	33.28	10.40	32.80	10.63	33.73	10.55	32.58	11.09	33.46	10.26
AFQT	68.69	19.22	68.02	19.79	67.10	19.88	68.22	19.78	68.27	19.85	68.29	19.61
QT-7a	54.77	20.80	54.37	20.94	54.68	21.02	54.91	21.05	54.89	20.77	55.40	20.82

Note. AFQT-7a is denoted by QT-7a.

Equating

All of the AFQT composites were calibrated using the AFQT-7a as the standard and were smoothed using polynomial regression with the constraint that the curve exhibit positive monotonicity. This meant that the curve was not permitted to turn downward, which would have provided two percentile points for a single raw score.

Each composite was calibrated in a full sample and two randomly selected half samples. The smoothing was applied to each subsample independently, and all three were used to decide on the appropriate smoothing on the basis of consistency among the samples and reduced standard error of estimate.

It is worth noting that the analytic procedure automatically provides a measure of fit, the standard error of estimate. Hand smoothing, as used in previous equating studies of ASVAB-8a, does not provide such an index without laborious computation. A goodness-of-fit of the equating curve for the previous studies was not assessed. This is one of the drawbacks to the nonanalytic method used previously.

Tables for the AFQT Forms

The tests were quite similar in frequency distribution and relationship to the calibration standard of AFQT-7a. This led to generally equivalent conversion tables for all six forms. Table 8 shows the conversions of each of the forms and the average correspondence of the six forms to the percentile standard or metric of AFQT-7a.

Table 8. Conversion Tables for Each Form

Raw Score	Percentile for Form 8a	8b	9a	9b	10a	10b	Overall[a] Avg
0-17	1	1	1	1	1	1	1
18	1	1	2	1	2	1	1
19	1	2	2	2	2	1	2
20	1	2	2	2	2	2	2
21	2	2	3	2	3	2	3
22	2	3	3	3	3	2	3
23	2	3	3	3	3	3	3
24	3	3	4	4	3	3	3
25	3	4	4	4	4	4	4
26	4	4	5	5	4	4	4
27	4	5	5	5	5	5	5
28	5	5	6	6	5	5	5
29	5	6	7	6	6	6	6
30	6	6	7	7	6	6	6
31	6	7	8	7	7	7	7
32	7	7	8	8	8	7	8
33	7	8	9	8	8	8	8
34	8	8	9	9	9	9	9
35	8	9	10	9	9	9	9
36	9	10	10	11	10	10	10
37	9	10	11	11	10	10	10
38	10	11	11	11	11	11	11
39	11	11	12	12	11	11	11
40	11	12	12	12	12	12	12
41	12	12	13	13	12	12	12
42	12	13	13	13	13	13	13
43	13	13	14	14	13	14	14
44	13	14	14	14	14	14	14
45	14	14	15	15	14	15	15
46	14	15	15	15	15	15	15
47	15	15	16	16	16	16	16
48	15	16	17	16	16	16	16
49	16	16	17	17	17	17	17
50	17	17	18	17	17	17	17
51	17	17	18	18	18	18	18
52	18	18	19	18	18	18	18

Raw Score	Percentile for Form 8a	8b	9a	9b	10a	10b	Overall[a] Avg
53	18	19	20	19	19	19	19
54	19	20	21	20	20	20	20
55	20	21	22	21	21	22	21
56	21	22	23	22	22	23	22
57	22	23	24	23	23	24	23
58	23	24	25	24	24	25	24
59	24	25	26	25	25	26	25
60	25	26	27	26	26	27	26
61	26	27	28	27	27	28	27
62	28	28	29	28	29	29	29
63	29	29	30	30	30	30	30
64	30	30	32	31	31	31	31
65	31	31	33	32	32	32	32
66	32	32	34	33	33	33	33
67	33	33	36	34	34	34	34
68	34	34	38	36	36	36	36
69	36	36	40	38	38	40	38
70	38	40	42	40	40	42	40
71	40	42	44	42	42	44	42
72	42	44	46	44	44	46	44
73	44	46	48	46	46	48	46
74	48	48	49	48	48	49	48
75	49	49	50	49	49	50	49
76	50	51	51	50	50	51	51
77	51	51	52	51	51	52	51
78	52	52	54	52	52	54	53
79	54	54	56	54	54	56	55
80	56	56	58	58	56	58	57
81	58	58	60	60	58	61	59
82	60	60	62	61	60	61	61
83	61	61	63	62	61	62	62
84	62	62	65	63	62	63	63
85	63	63	67	65	63	65	64
86	65	65	70	67	65	67	67
87	70	70	72	70	70	70	70
88	72	72	74	72	72	72	72

Table 8. (*Continued*)

Raw Score	Percentile for Form 8a	8b	9a	9b	10a	10b	Overall[a] Avg	Raw Score	Percentile for Form 8a	8b	9a	9b	10a	10b	Overall[a] Avg
89	74	74	76	74	74	74	74	98	88	88	89	89	88	88	88
90	76	76	78	76	76	76	76	99	90	89	90	90	89	89	89
91	78	78	80	78	78	78	78	100	91	90	92	91	90	90	91
92	80	80	81	80	80	80	80	101	92	91	93	92	91	91	92
93	81	81	82	81	81	81	81	102	93	92	94	93	92	92	93
94	82	82	83	82	82	82	82	103	94	93	95	94	93	93	94
95	83	83	85	83	83	83	83	104	95	94	96	95	94	94	95
96	85	85	87	87	85	85	86	105	96	96	97	96	95	95	96
97	87	88	88	88	87	87	87								

[a]Overall average based on conversion values prior to rounding to integers.

In order to determine if the ASVAB conversion tables truly differ, measures of deviation of subject percentile scores were computed using the operational, average, and form-specific table. These measures were RMS and AAD between pairs of interest. Table 9 shows the RMS and AAD for the AFQT. Although there are some differences among forms, the magnitudes of the differences are quite small. This is quite consistent with the two previous analyses and reinforces a picture of relatively small differences.

Table 9. Deviation Measures Comparing Use of One Versus Six Conversion Tables

	ASVAB AFQT Composites 8a thru 10b						
	Test Form						
Comparison	Pooled	8a	8b	9a	9b	10a	10b
			AAD				
O vs. P	.92	.79	.83	1.31	.68	.67	.98
O vs. A	.56	.88	.47	.65	.16	.25	.53
A vs. P	.65	.62	.64	.65	.67	.65	.65
			RMS				
O vs. P	1.25	1.25	1.24	1.48	.95	1.27	1.39
O vs. A	.87	1.04	.75	1.32	.40	.53	.84
A vs. P	.91	1.88	.90	.91	.92	.92	.92

Note. O = Optimum or 6 tables
P = Present operational table
A = Average of 6 tables from present study

It should be noted that the values for RMS exceed those for AAD, indicating that a few relatively large errors (four percentile points for one raw score in AFQT) exist. Inspection of the tables indicates that these deviations are generally limited to very low score ranges. This is probably attributable to guessing answers to the test items.

Table 10 shows the deviations across the five mental category boundary lines for the 15,115 subjects in the study. The comparison in Table 10 is between the conversion table put into effect 1 October 1980 and the form-specific tables developed in the present study (six tables in all). Off-diagonal entries are deviations.

Table 10. **Classification by Mental Category Based on One Versus Six Tables**

	Category by Operational Table				
Category by Six Tables	V	IV	III	II	I
V	934				
IV	177	5015			
III		121	5199	224	
II				3045	156
I					244

The proportion of deviations crossing boundaries can be computed by dividing the sum of the off diagonals by the sum of all the entries; it is 4.5%. In order to evaluate this percentage, a similar computation was done on the 8a form alone (not shown). The comparison was between the operational table outcomes and those from the specific table for 8a from the current study. The number of deviations across category lines was 2.4%. This value is useful as it presents an estimate of the expected deviations. Clearly the 4.5% representing the comparison of the present table versus the six tables is relatively small.

It was also deemed appropriate to investigate the number of deviations which were 1, 2, 3, or more percentiles in magnitude. Table 11 shows the deviations crossing categories. As may be observed, most of the deviations are not greater than one percentile point. Relatively few ever assume the magnitude of three percentile points and none are greater. It should be noted that for 14,437 subjects no deviations were observed.

Table 11. **Deviation of Percentile Scores across Category Lines**

		Size of Deviations		
Category	N	1 point	2 point	3 point
IV-V	177	71%	29%	
III-IV	121	69%	21%	
II-III	224	75%	25%	
I-II	156	35%	43%	22%

IV. CONCLUSIONS

Forms 8, 9, and 10 of ASVAB were found to be parallel when equated to AFQT-7a, and a single conversion table was deemed appropriate for operational enlistment processing.

REFERENCES

Angoff, W. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education. 1971, 508-600.

Boldt, R. F. *Scaling of the AFQT composite of the Armed Services Vocational Aptitude Battery Form 8 in a high school population.* Technical Memorandum 80-3. Washington, D.C.: Directorate for Accession Policy, Office of the Secretary of Defense, September 1980.

Flanagan, J.C. Units, scores, and norms. In E.F. Lindquist (Ed.), *Educational measurement.* Washington, D.C.: American Council on Education, 1951, 695-763.

Grunzke, M.E., Guinn, N., & Stauffer, G.F. *Comparative performance of low- ability airmen.* AFHRL-TR-70-4, AD-705 575. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, January 1970.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Jaeger, R.M. Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement,* 1981, 18, 23-38.

Maier, M. H., & Grafton, F. C. *Scaling Armed Services Vocational Aptitude Battery (ASVAB) Form 8a.* Working Paper MRPL-81-1. Alexandria, VA: Army Research Institute for the Behavioral Sciences, 1981.

Ree, M.J. *Development of statistically parallel tests by analysis of unique item variance.* AFHRL-TR-76-41, AD-A025 848. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1976.

Ree, M.J. *The effects of item-option weighting on the reliability of a perceptual ability test.* Paper presented at the 85th Annual Convention of the American Psychological Association, San Francisco, 1977.

Sims, W. H., & Truss, A. R. *Normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 using a sample of service recruits.* CRC 438. Alexandria, VA: Center for Naval Analyses, December 1980.