ASVAB Technical Bulletin No. 4

P&P-ASVAB Forms 23–27

Personnel Testing Division
Defense Manpower Data Center

March 2009
Revised May 2012

Table of Contents

Executive Summary vii

1. Introduction 1
1.1. Overview 1
1.2. History and Background of ASVAB 2
1.3. Purposes of ASVAB 2
1.4. Testing Programs 3
1.4.1. MEPCOM 3
1.4.2. ETP 3
1.4.3. STP/ASVAB CEP 4
1.5. Retesting 4
1.6. Advisory Groups 5
1.6.1. Defense Advisory Committee (DAC) on Military Personnel Testing 5
1.6.2. Manpower Accession Policy Working Group (MAPWG) 5
1.7. Program Oversight 5

2. Contents and Scoring of the ASVAB Battery 5
2.1. Subtests 5
2.2. Composites 6
2.2.1. AFQT Scores 6
2.2.2. Service Classification Composites 8
2.2.3. STP Composites 8
2.3. Major Changes to the Battery 8
2.3.1. IRT Scoring for P&P-ASVAB 8
2.3.2. Deletion of NO and CS from the Battery 9
2.3.3. Addition of AO to the Battery 10
2.3.4. Reordering the Subtests 11

3. Item Development and Form Assembly (Forms 23–26) 11
3.1. Timeline for Fielding Forms 23–26 12
3.2. Item Development 12
3.2.1. Sources of Items 12
3.2.2. Item Writing Procedures for New Items 13
3.2.3. Taxonomies 14
3.3. Item Review 15
3.3.1. Editorial Review 15
3.3.2. Sensitivity Review 15
3.3.3. Differential Item Functioning (DIF) Analyses 16
3.4. Item Tryouts 16
3.4.1. Development of Tryout Forms 16
3.4.2. Administration of Tryout Forms 16
3.4.3. Item Analysis 17
3.5. Final Form Assembly 18

4. Equating and Associated Studies (Forms 23–27) 19
4.1. ASVAB Equating/Scaling Overview 19
4.2. OPCAL 20
4.2.1. Item Overlap 21
4.2.2. Omega Issue 22
4.3. IOT&E 22
4.3.1. Scoring Methods/Score Scale Evaluation 23
4.3.2. Equating/Scaling Evaluation 25
4.4. Anchoring Study 27
4.5. Implementation of Forms 23–27 29

5. Norms for Forms 23–27 33
5.1. 1997 Norming Study 33
5.2. Test Form Equating 34

6. Statistical and Psychometric Properties of Forms 23–27 34
6.1. Subtest Moments 34
6.2. Subtest Intercorrelations 34
6.3. Item Parameters 37
6.4. Test Information Functions 40
6.5. Test-Retest Reliabilities 45

References 46

Appendices 49

Appendix A. History of the AFQT and the P&P-ASVAB 49

Appendix B. Timeline of Major Events in Recent ASVAB History 54

Appendix C. Service-Specific Composites 56

Appendix D. Military Installations used for Studies 58

Appendix E. AFQT Qualification Rates by Scoring Procedure 60

Appendix F. Subtest Theta Score (BME) Correlations 71

Appendix G. Subtest Number Right Score Correlations 76

List of Tables

2.1 P&P-ASVAB Content Summary 6
2.2 AFQT Categories 7

3.1 Timeline for Developing and Fielding Forms 23–26 12
3.2 Item Survival Rates for Forms 23–26 Tryout Pool 18

4.1 Forms Administered in the OPCAL 21
4.2 Forms Administered in the IOT&E 23
4.3 Comparison of Service Composite Qualification Rates for the Total Group 24
4.4 Comparison of Service Composite Qualification Rates by Subgroup 25
4.5 Form Qualification Rate Agreement for NR Equating and IRT Scaling 27
4.6 P&P-ASVAB Subtests in Old and New Orders 28
4.7 Forms Administered in Phase I of the Anchoring Study 28
4.8 Score Mean Differences and Effect Sizes Across Old and New Orders 30
4.9 Differences in CDFs for Standard Scores Across Old and New Orders 31
4.10 Differences in CDFs for Composite Scores Across Old and New Orders 32

6.1 Forms 23–27 Subtest NR Score Means and Standard Deviations 35
6.2 Forms 23–27 Subtest IRT Score (BME) Means and Standard Deviations 36
6.3 IOT&E Phase 1 Subtest NR Mean Intercorrelations 36
6.4 IOT&E Phase 2 Subtest NR Mean Intercorrelations 37
6.5 Summary of Item Parameters for GS Forms 23–27 37
6.6 Summary of Item Parameters for AR Forms 23–27 38
6.7 Summary of Item Parameters for WK Forms 23–27 38
6.8 Summary of Item Parameters for PC Forms 23–27 38
6.9 Summary of Item Parameters for MK Forms 23–27 39
6.10 Summary of Item Parameters for AS Forms 23–27 39
6.11 Summary of Item Parameters for MC Forms 23–27 39
6.12 Summary of Item Parameters for EI Forms 23–27 40
6.13 Summary of Item Parameters for AO Forms 23–27 40
6.14 Test-Retest Reliability Estimates 45

A.1 AFQT History 1950–Present 50
A.2 P&P-ASVAB Forms History 1968–Present 52

C.1 Service-Specific Composites 57

D.1 Recruit Training Centers Used in Item Tryout and OPCAL Studies 59
D.2 MEPS Participating in Anchoring Study 59

E.1 Form 28C Qualification Rates by Scoring Procedure for Total Group (N=10735) 61
E.2 Form 28C Qualification Rates by Scoring Procedure for Females (N=2690) 61
E.3 Form 28C Qualification Rates by Scoring Procedure for Blacks (N=2877) 61
E.4 Form 28C Qualification Rates by Scoring Procedure for Hispanics (N=1293) 62

E.5 New Form Qualification Rates by Scoring Procedure for Total Group (AFQT = 31).....62
E.6 New Form Qualification Rates by Scoring Procedure for Total Group (AFQT = 50).....63
E.7 New Form Qualification Rates by Scoring Procedure for Total Group (AFQT = 65).....63
E.8 New Form Qualification Rates by Scoring Procedure for Total Group (AFQT = 93).....64
E.9 New Form Qualification Rates by Scoring Procedure for Females (AFQT = 31).....64
E.10 New Form Qualification Rates by Scoring Procedure for Females (AFQT = 50).....65
E.11 New Form Qualification Rates by Scoring Procedure for Females (AFQT = 65).....65
E.12 New Form Qualification Rates by Scoring Procedure for Females (AFQT = 93).....66
E.13 New Form Qualification Rates by Scoring Procedure for Blacks (AFQT = 31) 66
E.14 New Form Qualification Rates by Scoring Procedure for Blacks (AFQT = 50) 67
E.15 New Form Qualification Rates by Scoring Procedure for Blacks (AFQT = 65) 67
E.16 New Form Qualification Rates by Scoring Procedure for Blacks (AFQT = 93) 68
E.17 New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT = 31) 68
E.18 New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT = 50) 69
E.19 New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT = 65) 69
E.20 New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT = 93) 70

F.1 Correlations Among Subtest Theta Score Estimates: Forms 23A (Phase 1).....72
F.2 Correlations Among Subtest Theta Score Estimates: Forms 23B (Phase 1).....72
F.3 Correlations Among Subtest Theta Score Estimates: Forms 24A (Phase 2).....72
F.4 Correlations Among Subtest Theta Score Estimates: Forms 24B (Phase 2).....73
F.5 Correlations Among Subtest Theta Score Estimates: Forms 25A (Phase 2).....73
F.6 Correlations Among Subtest Theta Score Estimates: Forms 25B (Phase 1).....73
F.7 Correlations Among Subtest Theta Score Estimates: Forms 26A (Phase 1).....74
F.8 Correlations Among Subtest Theta Score Estimates: Forms 26B (Phase 1).....74
F.9 Correlations Among Subtest Theta Score Estimates: Forms 27A (Phase 2).....74
F.10 Correlations Among Subtest Theta Score Estimates: Forms 27B (Phase 2).....75

G.1 Correlations Among Subtest Number Right Scores: Forms 23A (Phase 1) 77
G.2 Correlations Among Subtest Number Right Scores: Forms 23B (Phase 1) 77
G.3 Correlations Among Subtest Number Right Scores: Forms 24A (Phase 2) 77
G.4 Correlations Among Subtest Number Right Scores: Forms 24B (Phase 2) 78
G.5 Correlations Among Subtest Number Right Scores: Forms 25A (Phase 2) 78
G.6 Correlations Among Subtest Number Right Scores: Forms 25B (Phase 1) 78
G.7 Correlations Among Subtest Number Right Scores: Forms 26A (Phase 1) 79
G.8 Correlations Among Subtest Number Right Scores: Forms 26B (Phase 1) 79
G.9 Correlations Among Subtest Number Right Scores: Forms 27A (Phase 2) 79
G.10 Correlations Among Subtest Number Right Scores: Forms 27B (Phase 2) 80

List of Figures

6.1 Test Information Functions for GS Forms 23–27 41
6.2 Test Information Functions for AR Forms 23–27 41
6.3 Test Information Functions for WK Forms 23–27 42
6.4 Test Information Functions for PC Forms 23–27 42
6.5 Test Information Functions for MK Forms 23–27 43
6.6 Test Information Functions for AS Forms 23–27 43
6.7 Test Information Functions for MC Forms 23–27 44
6.8 Test Information Functions for EI Forms 23–27 44
6.9 Test Information Functions for AO Forms 23–27 45

B.1 Timeline of Major Events in Recent ASVAB History 55

Executive Summary

The Armed Services Vocational Aptitude Battery (ASVAB) is a cognitive aptitude battery measuring verbal ability, mathematical ability, science and technical knowledge and skills, and spatial ability. The ASVAB is administered to all applicants for enlistment in the military and in the nation's high schools as part of a career exploration program sponsored by the Department of Defense (DoD). The ASVAB is administered in two modes, paper-and-pencil (P&P) and a computerized-adaptive version, commonly called the CAT-ASVAB. By law, DoD is not allowed to accept into active duty an applicant who scores below the 10th percentile on the Armed Forces Qualification Test, a composite of four subtests. Each Service has its own classification composites that are used to qualify applicants for specific military occupations. Norms for the ASVAB have been developed from data gathered on three occasions: 1944, 1980, and 1997.

This technical bulletin describes how P&P-ASVAB Forms 23–26 were developed and outlines the supporting research studies. The technical bulletin also describes characteristics of Form 27, a retired operational form that was reordered, renamed, and re-equated alongside Forms 23–26, for use in the event of test compromise.

Thousands of new items were written in 1991 and 1992 to develop P&P-ASVAB Forms 23–26. The item writers were guided by content area taxonomies and target distributions of difficulty for each subtest. After editorial review for taxonomic coverage, estimated difficulty level, format, style, and sensitivity, the items were assembled into overlength forms and administered to recruits in tryout studies in 1992 and 1993. Final forms were assembled with the best items, as determined using classical and item response theory (IRT) statistics. The forms were matched to the 1980 norming form and to each other with respect to test information functions; the matching step involved item-swapping.

Forms 23–27 were administered to recruits in 1997 and 1998 to obtain interim equating transformations. The interim transformations were then used operationally in 2000 and 2001 to qualify a limited number of applicants for enlistment. Data from the applicant administration were used to develop the final equating transformations. Forms 25–26 were implemented in January 2002 in the Enlistment Testing Program and Forms 23–24 were implemented in July 2002 in the high school Student Testing Program. Form 27 was made available for operational use in the Enlistment Testing Program in January 2002.

A number of major changes to the battery coincided with the implementation of the new forms, including (a) changing the scoring method from number right to IRT scoring; (b) changing the order of subtest administration; (c) removing two speeded subtests, Numerical Operations (NO) and Coding Speed (CS), from the battery; and (d) adding a spatial ability test, Assembling Objects, to the battery. These changes are all discussed.

1. Introduction

The Armed Services Vocational Aptitude Battery (ASVAB) is a cognitive aptitude battery measuring verbal ability, mathematical ability, science and technical knowledge and skills, and spatial ability. The ASVAB is administered to all applicants for enlistment in the military, and results are used for determining enlistment eligibility, job placement, and the awarding of enlistment bonuses. It has been used as the single selection and classification battery for enlistment testing since 1976. The ASVAB also has been administered in the nation's high schools since 1968 as part of a comprehensive career exploration program sponsored by the Department of Defense (DoD).

Although the subject matter (content), administration conditions, and normative score scale of the ASVAB, as well as the methods used in its development, have changed over the years, the fundamental purpose of the battery has remained constant since 1976: to select applicants into the U.S. military and classify them into jobs.

1.1. Overview

This technical bulletin describes how paper-and-pencil (P&P) ASVAB Forms 23–26 were developed; how the forms were assembled and scored; how the battery was scaled, equated, and normed; and provides evidence of reliability, validity, and test fairness. An equating of Form 27 alongside Forms 23–26 is also described in the bulletin.[1]

The ASVAB is also administered as a computerized adaptive test (CAT-ASVAB); however, the focus of this bulletin is on P&P-ASVAB Forms 23–27. Thus, there is only incidental discussion of the CAT-ASVAB. Development, implementation, and evaluation of CAT-ASVAB item pools are documented in other bulletins in this series:

- *ASVAB Technical Bulletin No.* 1 (DMDC, 2006) documents item pools 1–2.
- *ASVAB Technical Bulletin No. 2* (DMDC, 2009) documents item pools 3–4.
- *ASVAB Technical Bulletin No. 3* (DMDC, 2008) documents item pools 5–9.

Also see Sands, Waters, and McBride (1997) for an in-depth recounting of the development of CAT-ASVAB from the very earliest stages through the partial implementation at five testing sites in 1992 and the nationwide implementation at all of DoD's main enlistment testing facilities in 1996–97.

Section 1 of this bulletin provides a brief history of ASVAB, describes the purposes of ASVAB, discusses the DoD testing programs, and explains provisions for program advisory groups and program oversight. Section 2 describes the subtests, composite scores, and changes to the battery accompanying implementation of the new forms. Section 3 focuses on procedures for the development of Forms 23–26. Section 4 summarizes the equating and linking studies and other

[1] Form 27 was created by renaming and re-equating a previously retired operational form (Form 15), and thus did not undergo development simultaneously with Forms 23–26. Form 27 is reserved for use in the event of test compromise.

associated studies that were carried out in support of implementation of Forms 23–27. Section 5 summarizes the 1997 norming effort. Section 6 provides psychometric and statistical specifics for the Form 23–27 subtests.

Forms 25–26 were implemented in January 2002 in the Enlistment Testing Program and Forms 23–24 were implemented in July 2002 in the high school Student Testing Program. Form 27 was made available for operational use (in the case of test compromise) in the Enlistment Testing Program in January 2002, along with Forms 25–26.

1.2. History and Background of ASVAB

During World War I, the Army evaluated the mental aptitude of potential recruits with a test of general mental ability, called the Army Alpha. During World War II (WWII), the Army General Classification Test (AGCT) and the Navy General Classification Test (NGCT) were the test batteries used for screening potential military recruits and enlistees. Additional classification tests, including an Air Force battery, were developed early in WWII to serve as measures of specialized aptitudes related to technical fields (Maier, 1993). The earliest administration of the ASVAB was in 1968 in DoD's high school Career Exploration Program.

From 1950–1973, enlistment qualification decisions for all Services were based on the AFQT, which was modeled after the AGCT. From 1973–1976, each Service obtained examinees' AFQT scores from the Service's own classification battery. The administration of three separate batteries (Army, Navy, and Air Force/Marine Corps) became burdensome, and in 1976, joint-Service testing began with the introduction of the ASVAB into enlistment qualification procedures. The history and subtests of the AFQT and the P&P-ASVAB are outlined in Tables A.1 and A.2 in Appendix A. Figure B.1 in Appendix B graphically displays major events in the recent history of P&P-ASVAB and CAT-ASVAB. A detailed history of AFQT and P&P-ASVAB is found in Maier (1993).

Norms for the ASVAB have been developed on three occasions: 1944, 1980, and 1997. The AGCT and the NGCT were combined in the 1940s and administered to a sample of recruits and commissioned officers from all Services; this group became known as the WWII mobilization population and served as the WWII reference population (Maier, 1993). In 1980, DoD teamed with the Department of Labor (DoL) to administer the ASVAB to a nationally representative sample of American youth of military-eligible age (U. S. Department of Defense, 1982). The work was called the Profile of American Youth 1980 (PAY80). The 1997 norms, called PAY97 (Segall, 2004), also were developed in a joint project with the DoL. The PAY97 study is summarized in Section 5.

1.3. Purposes of ASVAB

ASVAB subtest scores are combined in various ways into composite scores. The AFQT, developed from four ASVAB subtests, remains the measure used for determining enlistment eligibility. The AFQT also is used to determine eligibility for enlistment bonuses, to facilitate manpower management, and to report on the quality of accessions (enlistees) to the Congress. In

addition, each Service has its own composites that are used to qualify applicants for positions in training schools and to make job assignments (Maier, 1993).

1.4. Testing Programs

DoD has two major testing programs[2]: The Enlistment Testing Program (ETP) and the Student Testing Program (STP). The STP is known more formally as the ASVAB CEP. The administration of the ASVAB in both programs is the responsibility of the U. S. Military Entrance Processing Command (MEPCOM).

1.4.1. MEPCOM

MEPCOM has primary responsibility for administering and scoring the ASVAB subtests, and for handling other processing activities necessary for bringing an individual into the military. Additional ASVAB-related MEPCOM responsibilities include training test administrators (TAs), maintaining and ensuring the accuracy of the optical scanners used to process the answer sheets, accurately reporting subtest and composite information to the appropriate Services, and guaranteeing the overall security of the tests.

1.4.2. ETP

During fiscal year 2011, MEPCOM administered 460,000 enlistment ASVAB tests (USMEPCOM, n.d., para. 4). Enlistment testing takes place at the 65 Military Entrance Processing Stations (MEPS) and the 400+ Mobile Examining Team (MET) sites, both of which are under the administration of MEPCOM. The MEPS are DoD's joint-Service processing facilities, staffed by military and civilian personnel. Their job is to determine whether applicants meet the high mental, moral, and medical standards established by law and policy. For candidates who are deemed qualified for military service, the MEPS are also where they will meet with Service counselors, negotiate and sign enlistment contracts, and "swear in" by taking an entrance oath. Finally, enlistees "ship" to basic training from MEPS.

MET sites, each one associated with a specific MEPS, are satellite units used for ASVAB testing only. Such sites are housed in a variety of settings, such as National Guard facilities. An examinee that takes the ASVAB at a MET site and elects to enlist must go to a MEPS to complete in-processing.

ASVAB is administered in two modes, P&P-ASVAB and CAT-ASVAB. CAT-ASVAB employs the item response theory (IRT) three parameter logistic (3-PL) model and computes scores based on the posterior Bayesian modal estimate (BME) of examinee aptitude. It has been operational at all MEPS since 1997. Close to two-thirds of applicants for enlistment take the CAT-ASVAB, and the remaining examinees take the P&P-ASVAB. The MEPS administer only CAT-ASVAB; several MET sites also administer CAT-ASVAB, while the remaining MET sites

[2] In addition, active duty members of the military take special forms of the ASVAB as part of the process of changing jobs, or specialties, within the military, in what is called the In-Service Testing Program.

administer P&P-ASVAB only. Efforts are continuing in support of expanding CAT-ASVAB administration to additional MET sites.[3]

The P&P-ASVAB is administered under standardized conditions typical for a group-administered multiple aptitude battery. Administration is in a lock-step fashion with the instructions read by trained TAs. The P&P-ASVAB takes about three hours to complete and administration of each subtest is precisely timed, as specified using the time limits set forth in Table A.2 in Appendix A. The testing sessions are proctored, and the number of proctors is determined by the number of applicants being tested. Emphasis is placed on having a testing environment that is secure, comfortable, and free of distractions to the greatest extent possible. Detailed rules and procedures for administering and proctoring the ASVAB are set forth in a test administration manual produced by MEPCOM.

1.4.3. STP/ASVAB CEP

The ASVAB CEP is provided by the DoD free of charge to high schools and postsecondary schools nationwide, and is intended for use by students in grades 10–12, as well as students in postsecondary schools. The CEP is designed to help students learn more about themselves and the world of work, explore occupations in line with their interests and skills, and develop an effective strategy to realize their career goals. The cognitive testing component is the ASVAB, which is administered in the P&P version only. Test administration takes place in participating schools with DoD employees serving as TAs and proctors. The program also provides an interest inventory and a wealth of supporting materials. During the 2010–11 school year, 658,000 high school students were tested under the program (USMEPCOM, n.d., para. 5).

If CEP participants wish to consider enlisting in the military, their ASVAB scores are valid for two years from the date of testing. However, participation in the CEP carries no obligation on the part of students, nor is there a requirement that participants' scores be released to military recruiters. A more detailed description of the program may be found in the *ASVAB Career Exploration Program Counselor Manual* (U. S. Department of Defense, 2005).

In addition to the in-school program, DoD maintains a website[4] that allows students to use most of the materials (although ASVAB testing is not available on the site) and another website[5] with an abundance of material about military careers. The websites have links to various other career/vocation-related sites that are useful to students.

1.5. Retesting

Applicants who wish to improve their scores are allowed to retake the ASVAB, but time intervals between test sessions are governed by policy. The current retest policy allows the first retest as early as 30 days following initial testing, and a second retest after another 30-day wait; additional retests can occur at six month intervals. The controlled time intervals between test sessions are designed to minimize score gains that may result from familiarity with subtest

[3] More details about the ETP are available at http://official-asvab.com/.
[4] http://asvabprogram.com.
[5] http://www.careersinthemilitary.com.

content and to minimize the risk of compromise. Applicants who retest are given a different form than the one(s) taken previously. Frequently, applicants for enlistment who took the ASVAB in the STP will choose to retest in an effort to improve their scores.

1.6. Advisory Groups

1.6.1. Defense Advisory Committee (DAC) on Military Personnel Testing

The DAC is an independent advisory group, composed of volunteer experts in psychometrics, statistics, and survey work. The Committee is charged with reviewing the calibration of personnel selection and classification tests, reviewing relevant validation studies, reviewing ongoing testing research and development, and recommending improvements in the testing process. The DAC was established in 1981 in response to the earlier miscalibration of the ASVAB (Maier, 1993). The DAC meets two or three times per year.

1.6.2. Manpower Accession Policy Working Group (MAPWG)

The MAPWG is composed of representatives from each branch of the military, the Defense Manpower Data Center (DMDC), MEPCOM, and Accession Policy. MAPWG responsibilities include resolving issues related to test development, implementation, and maintenance; and making policy recommendations. The MAPWG was organized in 1974 (Maier, 1993), and it also meets two to three times per year.

1.7 Program Oversight

The Director of Accession Policy, Office of the Under Secretary of Defense (Personnel and Readiness), has primary responsibility for enlistment testing policy, and has a representative who sits on the MAPWG and serves as the DAC's Executive Secretary.

2. Contents and Scoring of the ASVAB Battery

2.1. Subtests

The subtests in the P&P-ASVAB are General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Auto & Shop Information (AS),[6] Mathematical Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), and Assembling Objects (AO). ASVAB subtests are designed to measure aptitudes in four domains: Verbal (V), Math (M), Science and Technical (S&T), and Spatial (Sp). Table 2.1 provides subtest names, number of items, test time limits, broad content description, and the domain tested.

[6] AS is administered as two separate tests in CAT-ASVAB, Auto Information and Shop Information, but reported as one single score (labeled AS).

Table 2.1. P&P-ASVAB Content Summary

Subtest[a]	Number of Items	Testing Time[b]	Subtest Description	Domain[c]
General Science	25	11	Knowledge of physical and biological sciences	S&T
Arithmetic Reasoning	30	36	Ability to solve arithmetic word problems	M
Word Knowledge	35	11	Ability to select the correct meaning of words presented in context and to identify the best synonym for a given word	V
Paragraph Comprehension	15	13	Ability to obtain information from written passages	V
Math Knowledge	25	24	Knowledge of high school mathematics principles	M
Electronics Information	20	9	Knowledge of electricity and electronics	S&T
Auto & Shop Information	25	11	Knowledge of automobile technology, tools, and shop terminology and practices	S&T
Mechanical Comprehension	25	19	Knowledge of mechanical and physical principles	S&T
Assembling Objects[d]	25	15	Ability to determine how an object will look when its parts are put together	Sp

Note. Table adapted from Sands and Waters (1997).
[a] The subtests are listed in the order in which they are currently administered.
[b] Testing time is given in minutes.
[c] V = Verbal; M = Mathematics; S&T = Science and Technical; Sp = Spatial.
[d] Assembling Objects is not administered in the STP.

The ASVAB subtests are designed as power subtests, as opposed to speeded subtests; as such, the objective is to measure maximum, or best, performance, not speed of cognitive processing. Thus, time limits for both P&P-ASVAB and CAT-ASVAB are set with the goal that virtually all examinees are able to complete the subtests, in accordance with the philosophy of power subtests. Analyses for CAT-ASVAB support the conclusion that almost all examinees do, in fact, have adequate time (DMDC, 2008).

2.2. Composites

Three main types of composite scores are calculated from ASVAB subtest standard scores: the AFQT score, Service composites, and STP/CEP composites. Subtest scores are not used in any official manner apart from composites, although individual subtest standard scores are reported.

2.2.1. AFQT Scores

By law, DoD is not allowed to accept into active duty an applicant who scores below the 10th percentile on the AFQT. Furthermore, the military branches are severely limited in the number

that may be inducted from among those who score from the 10th to the 30th percentiles, inclusive. Beyond these restrictions, the individual Services are free to impose other limitations.

The composition of the AFQT has shifted over the decades, as illustrated in Table A.1 in Appendix A. Since January 1989, the AFQT has been composed of four subtests AR, MK, WK, and PC. A verbal composite (VE) score is formed from an optimally weighted composite of unrounded WK and PC standard scores (Segall, 2004). VE, in turn, is double-weighted in the computation of AFQT scores:

$$AFQT = AR + MK + 2(VE). \quad (1)$$

Prior to January 2002, AFQT scores for P&P-ASVAB were calculated as the sum of weighted standard scores based on number right (NR) scoring. In 2002, IRT scoring (discussed in Section 2.3.1) replaced NR scoring, and transformation equations operating on IRT ability estimates for examinee j ($\hat{\theta}_j$) were adopted to compute AFQT scores (Segall, 2004):

$$AFQT_s = (t_{AR}\hat{\theta}_j^{AR} + u_{AR}) + (t_{MK}\hat{\theta}_j^{MK} + u_{MK}) + 2(VE_s), \quad (2)$$

where t_k and u_k, are weights that convert score k to a standard score (representing the slope and intercept parameters, respectively), and VE_s is the standard score verbal composite:

$$VE_s = t_{WK}\hat{\theta}_j^{WK} + t_{PC}\hat{\theta}_j^{PC} + u_{VE}. \quad (3)$$

The transformation equation weights t_k and u_k were developed from the most recent norming effort, PAY97, which is discussed in Section 5. The standard scores are scaled to have a mean of 50 and a standard deviation of 10.

The AFQT standard score is then converted to a percentile, which is used for enlistment qualification. Further, the percentiles are grouped into five categories (I–V) for reporting purposes, with Category I being the highest; the levels can be thought of as representing "trainability" (Sands & Waters, 1997). Table 2.2 provides the breakdown of the categories and their corresponding percentile score ranges.

Table 2.2. AFQT Categories

AFQT Category	Percentile Score Range
I	93 - 99
II	65 - 92
III[a]	31 – 64
IV	10 - 30
V	1 - 9

[a]Category III is typically divided at the 50th percentile into Category IIIA (50th – 64th percentile) and Category IIIB (31st – 49th percentile). The modern "quality" benchmark is Category IIIA and above.

2.2.2. Service Classification Composites

Each Service develops and validates its own composites using ASVAB standard scores. The Service classification composites are used to qualify applicants for positions in training schools and for assignment to specific military occupations. In general, the validation criterion is a measure of success in entry-level training or performance on the job. The number of composites used by the Services varies. As of this writing, the Army and Navy each have ten, while the Air Force and Marine Corps each have four. Table C.1 in Appendix C lists the current Service composites and the subtests used in the computations.

2.2.3. STP Composites

Each student participant is given composite scores representing his/her Verbal, Mathematical, and Science and Technical aptitudes. The Verbal score is a composite of WK and PC; the Mathematics score is a composite of MK and AR; and the Science and Technical score is a composite of GS, EI, and MC. Student participants also receive a Military Entrance Score (which is their AFQT score) that is valid for enlistment (U. S. Department of Defense, 2005).

2.3. Major Changes to the Battery

Several major changes were introduced simultaneously with the implementation of Forms 23–27. The changes included upgrading from NR scoring to IRT scoring, dropping two speeded tests, Numerical Operations (NO) and Coding Speed (CS), adding AO, and changing the order of subtest administration.[7] Whenever subtests are candidates for addition to the battery, careful consideration of test administration time is necessary. Because P&P-ASVAB and CAT-ASVAB must provide parallel assessment of aptitudes, any change to one mode requires an identical change to the other mode. The implication is that adding tests to the battery would necessarily increase CAT-ASVAB testing time at the MEPS, and processing applicants through the MEPS is intensive and occurs on a very tight schedule. As a result, any changes to the composition of the battery that increase testing time must be balanced against competing demands of the situation. Dropping NO and CS and adding AO resulted in an acceptable testing-time increase of five minutes for ETP examinees.

Extending ASVAB testing time in the STP may jeopardize the participation of some schools because testing time is a major issue for schools. While AO is not currently administered in the schools, consideration is being given to changing that policy.

2.3.1. IRT Scoring for P&P-ASVAB

Prior to CAT-ASVAB development and implementation, all ASVAB testing was P&P-based. The reference form for equating and linking studies (Form 8A) was also P&P-based, so it was logical to use NR scores transformed to standard scores ($\overline{X} = 50$; standard deviation = 10) for each subtest. When data for the new norms were gathered in 1997, the designated reference

[7] In addition, the Services made some changes in their composites; those changes are not documented in this report.

form, CAT-ASVAB Form 04D, utilized IRT methodology for item calibration and examinee scoring, and it became logical to use IRT scoring for both P&P-ASVAB and CAT-ASVAB.

The conversion to IRT scoring in P&P-ASVAB was a significant improvement. IRT scoring promised more precise measurement; aligned P&P scoring with CAT-ASVAB scoring; produced a more nearly-normal AFQT score distribution with fewer score gaps; and alleviated, to some extent, the problem of ceiling effects that occurred in some P&P subtests (for example, PC) when using NR scoring. Details of the supporting research studies are reported in Section 4.

Item responses are assumed to follow the 3-PL model,

$$P_i(\theta) \equiv p(u_i = 1 \mid \theta) = c + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \quad (4)$$

where $P_i(\theta)$ is the probability of a correct response given θ; a_i is the item discrimination parameter; b_i is the item difficulty parameter; and c_i is the item lower asymptote, or "guessing" parameter. It is assumed that there is one underlying trait that accounts for θ, i.e., that unidimensionality holds.

2.3.2. Deletion of NO and CS from the Battery

The two speeded tests, NO and CS, were the subject of careful study and years of discussion before the decision was finalized to drop them from the battery. While analyses showed that the power tests were relatively robust to changes in administration, the same could not be said about the speeded tests. The nature of speeded tests, which contain easy items all examinees could answer correctly if given enough time (Allen and Yen, 1979), created special problems with the maintenance of the battery. The two tests were very sensitive to any change, however slight, in administration. For example, test performance was found to be affected by answer sheet format and textual differences in P&P-ASVAB booklets. In the CAT-ASVAB, the impact on NO and CS scores had been documented with changes in hardware or software. Consequently, every time a change of any sort occurred, equating studies had to be undertaken. Furthermore, the rapid obsolescence of computer equipment would require either stockpiling hardware to be used as replacements or upgrading equipment and conducting frequent equatings. Equating studies were not only expensive, but access to examinees at recruit depots was becoming increasingly difficult.

In addition, CS was susceptible to coaching, and, historically, NO and CS were among the least reliable (Palmer, Hartke, Ree, Welsh, & Valentine, 1988) and least valid (Welsh, Kucinkas, & Curran, 1990) tests in the battery. It was unclear what construct was being measured by CS, whether the same construct was being measured by both CAT-ASVAB and P&P-ASVAB, and what job-relevant construct(s) was (were) being tapped in the criterion space. Finally, retaining NO and CS would have implications for the plan to convert to IRT scoring because the use of IRT is not appropriate for speeded tests.

The Services were tasked to evaluate the usefulness/validity/fairness of the two speeded tests and explore alternatives for the makeup of the composites in which the tests appeared. All Services

except the Navy found satisfactory substitutes for NO and CS in their composites, although there was some hesitation on the part of the Army. The Navy produced evidence that NO and/or CS were of value (for a subset of ratings) in terms of validity, classification efficiency, and reduction of occupational barriers for women and blacks. Additional analyses were undertaken by DMDC to evaluate the use of alternative Navy composites as replacements for composites that included the speeded tests. Results suggested that the alternative composites could be used without compromising validity. In the DMDC analyses, using NO and/or CS did not provide greater average validity compared to an alternative composite, and the alternative composite was nearly neutral in terms of the impact on gender and ethnic subgroups. Attrition and cost data also were analyzed and, although there was a slight projected cost-savings with the speeded tests, no compelling case could be made for keeping them.

Ultimately, Accession Policy made the decision to terminate routine administration of the two speeded tests to all applicants, and to make CS available as a special test on the CAT-ASVAB platform for any Service that wished to have it administered.

2.3.3. Addition of AO to the Battery

AO was originally part of the Enhanced Computer Administered Test battery, a battery composed of nine tests measuring non-verbal reasoning, spatial ability, psychomotor skill, and perceptual speed (Wolfe, 1997; Wolfe, Alderton, Larson, Bloxom, & Wise, 1997). Although AO had been administered experimentally on the CAT-ASVAB platform[8] for many years to applicants for enlistment, it was not scored for operational use prior to the implementation of Forms 23–27. Thus, it required special attention as a candidate for a test that "counted."

AO's lengthy history provided solid evidence supporting the validity, reliability, and fairness of AO. Numerous studies using data from a nonadaptive computerized administration were reported in the literature, most of which evaluated AO in terms of incremental validity. Generally, the zero-order (uncorrected) validity was about $r = .47$ using final school grade as the criterion. AO was found to be lower on verbal demands, and thus had a lower correlation with years of education, as compared to other subtests (Wolfe, Alderton, Larson, & Held, 1995). Factor analytic work showed that AO also loaded highly on '*g*' measures, indicating that it was likely a better "knowledge-free" measure of reasoning (Alderton, Wolfe, & Larson, 1997; Wolfe, Alderton, Larson, & Held, 1995). Test-retest reliability estimates were $r = .83$ (Held & Wolfe, 1997; Larson & Alderton, 1992; Larson & Alderton, 1997). Also, gender and practice effects were both found to be nonsignificant (Larson & Alderton, 1997).

By June 1998, more than 300,000 cases from CAT-ASVAB were available for analysis. Maximum likelihood factor analysis followed by an oblique rotation yielded a three-factor solution. One interpretation of the result was that the construct measured by AO was similar to the constructs measured by AR, NO, CS, and MK. AO was hypothesized as a measure of deductive reasoning ability in that it requires the application of rules that are provided. In the case of AO, the application of the rules is in the spatial realm.

[8] AO was initially administered as a fixed form test in CAT-ASVAB, rather than as an adaptive test, until it was adopted operationally.

Data from the CAT-ASVAB administration of AO also were used to study effect size by Service, test, and gender, and mean score differences (*z*-scores) as compared to the same statistics for CS and MC. Across Services, the effect sizes were smaller for AO as compared to CS and MC. AO scores for males were very slightly higher than for females. CS favored females, while MC favored males. The *z*-score analyses indicated that there were advantages for females for CS and sizeable differences in favor of males for MC.

AO was subsequently added to P&P-ASVAB with the introduction of Forms 23–27 in 2002. Adaptive administration of AO was implemented in CAT-ASVAB in 1999 when Forms 3–4 were introduced.

2.3.4. Reordering the Subtests

Part of the rationale for changing the order of subtest administration was to place the AFQT subtests in proximity to one another while continuing to administer them early in the testing session. Thus, MK was moved to a location with the other AFQT subtests, from eighth position to fifth in order of administration. This placement of the AFQT subtests was expected to minimize the negative effects of examinee fatigue that may occur at a later point in the test session. GS remained as the initial subtest to serve as a "warm-up." EI was moved to an earlier position (from tenth to sixth) to help compensate for differential fatigue effects by gender. Additional details and research studies supporting reordering are discussed in Section 4.

3. Item Development and Form Assembly (Forms 23-26)

Research and development underlying the fielding of P&P-ASVAB forms is exhaustive in the sense that (a) the steps in test construction are specified in detail, (b) multiple checks are built into the system, (c) multiple research studies are conducted in support of implementation, (d) all work follows sound test construction principles, (e) all phases of test construction and all research studies are reviewed by outside experts serving on advisory committees, and (f) all work conforms to the Standards for Educational and Psychological Testing (American Educational Research Association, 1999).

This section describes the item writing and test construction procedures that were in effect when Forms 23–26 were constructed. Form 27 is a renamed, re-equated previously-operational (holdout) form that was equated alongside Forms 23–26, but developed independently of Forms 23–26. The development of Form 27 is not discussed here. Forms 23–24 are administered exclusively in the STP, while Forms 25–26 are designated for the ETP. Form 27 is reserved for use in the case of test compromise.

Two versions of each new form were created, labeled A and B. Thus, the following forms were created: 23A, 23B, 24A, 24B, 25A, 25B, 26A, and 26B. Forms 27A and 27B were derived by reordering items on Forms 15A and 15B.

The basic content structure of the current ASVAB was determined during the development of Forms 8–10 in the late 1970s. In 1980, Form 8A of the series was designated to be the norming reference form (U. S. Department of Defense, 1982) and all forms from the Forms 8–10 series

through the Forms 20–22 series were modeled on and equated to 8A. A slightly modified approach was used for development of Forms 23–26, as described later in this section.

3.1. Timeline for Fielding Forms 23–26

Table 3.1 summarizes the major events in the research, development, and fielding of the new forms. This section discusses events from item development through final form assembly, and Section 4 details the studies conducted in support of implementation, including an Operational Calibration (OPCAL), an initial operational test and evaluation (IOT&E), an Anchoring Study, and some special studies that were prompted by non-routine issues that arose.

Table 3.1. Timeline for Developing and Fielding Forms 23–26

Study	Dates	Location
Item development		
AFQT	Fall 1991	DMDC
Technical	Fall 1992	DMDC
Overlength form assembly	1992 – 1993	DMDC
Tryouts		
AFQT	June 1992 – November 1992	RTCs[a]
Technical	June 1993 – December 1993	RTCs[a]
Final form assembly	1994 – 1996	DMDC
OPCAL	August 1997 – March 1998	RTCs[a]
IOT&E		
Phase 1	November 2000 – March 2001	MET Sites
Phase 2	March 2001 – July 2001	MET Sites
Anchoring Study	October 2000 – April 2001	MEPS
Implementation		
ETP Forms 25 and 26	January 2002	Nationwide
STP Forms 23 and 24	July 2002	Nationwide

[a] RTC = Recruit Training Center.

3.2. Item Development

3.2.1. Sources of items

The three primary sources of items were (a) previously field-tested items with acceptable statistics, (b) previously field-tested items that had been edited (based on existing data), and (c) newly-written items.

3.2.1.1. Previously field-tested items with acceptable statistics.

In any large-scale item tryout, there are statistically acceptable items that are not selected for inclusion in the forms under construction. Although these items have valid data and may be used as is, they are tried out again in order to gather more current data.

3.2.1.2. Previously field-tested items that had been edited.

Item analysis data frequently provide diagnostic information useful for correcting items with statistical shortcomings. These edited items can be tried out again.

3.2.1.3. Newly-written items.

The large majority of try-out items were new items developed using the ASVAB's test development specifications and editorial procedures. This process is described in the next section.

3.2.2. Item Writing Procedures for New Items

Items for P&P-ASVAB Forms 23–26 were the first pools of new items developed by the staff of the then newly-formed Personnel Testing Division (PTD) of DMDC. The PTD editors also had free-lance (contractor) personnel to assist them with item-writing responsibilities. The contract item writers, who were recruited by advertising and word-of-mouth, spent the fall months of 1991 and 1992 writing items at DMDC. Although the item writers were experienced and were subject matter experts (SMEs), they were required to take a screening test to demonstrate their proficiency in writing items. The number of item writers varied between 10 and 20. They were trained by the PTD staff, and all were part-time employees.

Prior to item writing, the item writers attended a one-day training workshop. In addition, each one received an item writer's manual that covered the following topics:

- *General guidelines*. This section covered commonly accepted rules and practices in writing high-quality test items. It provided examples of well-written items and items demonstrating commonly made errors.
- *Specific guidelines*. This section dealt specifically with the content area assigned to the writer. Different content areas (for example, paragraph comprehension, mathematics, science) presented different problems and pitfalls, and each had specific guidelines for writing high-quality items.
- *Format and editorial guidelines*. This section included guidelines on grammar, punctuation, and usage, as well as instructions on how to format items according to PTD's editorial style. It also included information on copy editing.
- *Sensitivity guidelines*. This section helped writers to avoid content bias and stereotypes of a number of minority groups. Particularly emphasized were biases against gender, ethnicity, age, and special needs groups.
- *Writing assignments*. This section included specific writing assignments that addressed content taxonomy, difficulty level, and the number of items in each category.
- *Checklists*. This section included review checklists for use by writers at each stage of item development to assess both their own writing and the work of others.

Item development was done separately for the AFQT and non-AFQT tests. Writers worked in a two-person buddy system for the generation of new items. The pairs of item writers reviewed each other's items, and after editing, sent the items to PTD editors for editing. Altogether, they produced well in excess of 4,000 items that qualified for tryout procedures. The item writers

employed for writing the technical subtest items all had backgrounds as teachers. The technical items were more difficult to write because they involved specialized subject matter and contained many more illustrations. Some technical-item writers also had been employed as writers for the AFQT subtests.

Some "new" items were "cloned" versions of existing items. Cloning involved rewriting items with changes that rendered them different enough that they were unlikely to be affected by operational use of the items from which they were cloned. Caution must be exercised in using clones, however, because the similarity of cloned items may violate the local independence assumption of IRT in the event that an original item and its clone were in the same test. The IRT local independence assumption "requires that any two items be uncorrelated when θ is fixed" (Lord, 1980, p. 19).

Graphics were used when appropriate for an item; decorative art was not used. The ASVAB subtests that included graphics were MK, GS, AS, MC, EI and AO. Text and artwork were carefully specified as to style and font size, line weight, fills, and shading.

All subtests had the typical multiple choice one-stem, four-response options format. In Forms 23–26, PC had five items associated with each reading passage.

Each item was assigned a unique identification number. Detailed, precise records were kept on the source of each item and its history; for example, whether it had been used before and, if so, the identification code or codes previously associated with it.

An integral part of item development was the documentation/verification of content from authoritative sources, such as current textbooks or SMEs. This procedure ensured that items were relevant in content, technically correct, and appropriately targeted to the specified grade levels. Vocabulary level (or reading grade level) was controlled to ensure examinees were likely to understand what they read. The exception was when a term was necessary in a math, science, or technical subtest item to measure knowledge of the content area.

3.2.3. Taxonomies

All P&P-ASVAB subtests are constructed to content specifications. Form 8A was used as the model for developing all subsequent P&P-ASVAB forms through Forms 20–22. This entailed matching Form 8A on an item-by-item basis with respect to content, difficulty, and discrimination. While this procedure helped ensure a high level of equatability of each new set of forms, it did not take into account, or adjust for, the fact that the content of Form 8A could become obsolete or that content emphases might change over the next several decades. One particularly vulnerable subtest is EI because the field of electronics develops and changes so rapidly.

The procedure was changed with the development of Forms 23–26. Instead of item-by-item matching to Form 8A, preliminary content taxonomies were established for each subtest. The preliminary taxonomies were based on content taxonomies from similar testing programs, selected state and large-city curriculum guides, and the contents of Form 8A. The subtest contents of Form

8A were then mapped onto the preliminary content taxonomies, and those portions of the preliminary taxonomies that bore no resemblance to the content of the ASVAB were deleted; to include them would have required renorming the battery. Areas of the ASVAB content taxonomies were identified in which updating of contents or shifts in emphasis could be made without jeopardizing the equatability of the new forms with the reference form. "Evolution" rather than "revolution" was the approach taken to updating the content for new forms.

In addition to content specifications, difficulty levels of items in the target form also guided items writers. Difficulty, in classical testing theory, is defined by the item's *p* value, the proportion of examinees giving the correct response to the item. Although the definitions of easy, medium, and hard vary by subtest, in general, the easy ASVAB items had a *p* value greater than .75, medium difficulty items had a *p* value between .75 and .45, and hard items had a *p* value less than .45.

3.3. Item Review

3.3.1. Editorial Review

Provisions for editorial review were built into each phase of test development. The first level was peer review: each item writer had a partner who reviewed the newly-written items. Following the review checklists in the item writer's manual, they critiqued and edited each other's work for content accuracy and match to content taxonomy, estimated difficulty, format and style, sensitivity, and other features that characterize a technically well-written item. They also reviewed the item's documentation. The partner approach also proved helpful when a writer needed help during item development, such as ideas for an additional valid distractor. Suggested edits were returned to the "owner" editor, who decided which edits to accept and which to reject. After any additional work was done as a result of the peer review, and when an item writer deemed the items ready for submission, they were sent to the appropriate content editor at PTD for the next level of editorial review. After completing the review (which covered the same areas as the peer review), the content editor either returned the items to the writer or met with the writer to discuss necessary changes. When the items were approved by the content editor, they went to a senior editor for the next level of review.

3.3.2. Sensitivity Review

Because an important objective is fair measurement, ASVAB subtests were subjected to reviews designed to identify any items that may place subgroups at a disadvantage. In addition to the item-writing training outlined above, a two-hour sensitivity training workshop was conducted for the new item writers, and sensitivity reviews (using checklists) were conducted at each stage of the item-writing and review process. For subtests in which language bias may be introduced, for example, WK, native Spanish speakers reviewed the items.

Ideally, all content bias should be eliminated by the time the items are ready for tryout. PTD conducted an informal study to address whether additional sensitivity reviews needed to be done by outside experts (Harris & Weger-Montano, 2000). Results suggested that inexperienced writers who completed the sensitivity training were as skilled as experienced, paid reviewers in

finding content bias. Nevertheless, an additional sensitivity review was done under contract to outside experts before items were assigned to forms.

3.3.3. Differential Item Functioning (DIF) Analyses

DIF exists when "examinees of equal ability differ, on average, according to their group membership in their response to a particular item" (American Educational Research Association, 1999, p. 81). DIF is a necessary (but not sufficient) condition for the occurrence of bias. Test items are biased if "they contain sources of difficulty that are irrelevant or extraneous to the construct being measured, and these extraneous or irrelevant factors affect performance (Zumbo, 1999).

The subgroups for which ASVAB DIF analyses were conducted were females, blacks, and Hispanics. In each case, the subgroup (minority or "focal" group) performance was contrasted with the non-minority ("reference") group performance.

Items flagged for DIF underwent an additional review to assess the likelihood of the occurrence of bias.

3.4. Item Tryouts

Eligible items from the identified sources were assembled into booklets and tried out to evaluate their performance. Administration of tryout booklets continued until sufficient response data was obtained for item analyses.

3.4.1. Development of Tryout Forms

Overlength forms were developed for use in the item tryouts. Namely, the number of items in each form exceeded the number of items that would be included in the final forms, in order to allow for selection of the best items. The overlength forms were constructed in accordance with the newly revised taxonomies. All items that qualified for tryout were assigned to the overlength forms such that, to the greatest extent possible, the distribution of items in each subtest would approximate (proportionally) the content specifications for the final subtest.

A final review of the camera-ready P&P test booklets was conducted by the entire editorial staff prior to printing to check for key accuracy and to review (and revise, if needed) the distribution of response options corresponding to the item keys. Each item was carefully checked to ensure that it conformed to all details of specifications and standards.

3.4.2. Administration of Tryout Forms

The final goal, of course, is operational use of the items, and the applicant population is the population of interest. Enlistment decisions, however, cannot be based on untried items and tests with unknown psychometric properties. Thus, preliminary item statistics must be gathered before the final forms can be assembled, and the recruit population was the most feasible choice for

tryouts. [9] The tryout forms were administered at geographically-dispersed Recruit Training Centers (RTCs) in 1992 and 1993. Item tryouts were done separately for the AFQT and non-AFQT tests.

3.4.2.1. Item Tryouts for AFQT Tests

AFQT overlength-form item tryouts began in June, 1992 at nine RTCs (See Table D.1 in Appendix D) and ended in November, 1992 (N = 23,748). There were 24 tryout test booklets, four sets of six booklets each. Each set included the reference form and five booklets with new items. The total number of AFQT items tried out was nearly 2,400; a total of 2,215 items remained after the initial tryout. By design there was some overlap of items, as the intention was to study item position effects.

3.4.2.2. Item Tryouts for Non-AFQT Tests

Seven RTCs (see Table D.1 in Appendix D) were designated as test sites for item tryouts for the non-AFQT subtests. Testing began in June, 1993 and ended in December, 1993. The total number of new items was approximately 2,200, contained in five booklets. The target number of examinees was N = 20,200 (1,000 for each booklet); the actual number tested was N = 20,953.

3.4.3. Item Analysis

Item parameter estimation was conducted using BILOG (Mislevy & Bock, 1990). IRT parameter estimates were a_i, b_i, and c_i for the 3-PL model (see Equation 4). Classical statistics were also computed, including the biserial correlation coefficient (the correlation between the artificially dichotomized correct/incorrect item score and the total score) and the $\underline{p}$ value (proportion correct). Both classical and IRT statistics were placed on the 1980 youth population scale.

In addition to the IRT parameters for each item, classical statistics for both the correct response and for distractors were analyzed. Items were excluded if the biserial of the correct answer was not significantly positive, a distractor with a significant $\underline{p}$ value was selected by examinees with higher subtest scores, a distractor with a significant $\underline{p}$ value was selected by examinees with subtest scores not significantly lower than examinees with high subtest scores, examinees with high subtest scores had less than a $\underline{p}$ =.5 probability of answering the item correctly, or if $\underline{p}$ values for the highest and lowest distractors were too close. DIF statistics were also calculated. A few item position effects were found and adjustments were made. Survival rates for the tryout items are given in Table 3.2.

[9] Due to a number of limitations associated with using recruits (including restriction of range for aptitude due to the exclusion of low scoring applicants, low motivation, and cost), this was the final generation of items tried out in this way. New items are now pretested with applicants during CAT-ASVAB administration (DMDC, 2008).

Table 3.2. Item Survival Rates for Forms 23–26 Tryout Pool

Subtest	# of Items Written	# of Items Retained	Percent Survival[a]
AFQT			
Arithmetic Reasoning	626	585	93
Word Knowledge	660	605	92
Paragraph Comprehension	595	554	93
Mathematics Knowledge	505	452	90
Total (mean percent)	2,386	2,196	(92)
Science/Technical			
General Science	675	631	93
Auto & Shop Information	555	458	83
Mechanical Comprehension	555	460	83
Electronics Information	423	344	81
Total (mean percent)	2,208	1,893	(86)

[a] Some items were screened out previously due to out-of-range statistics.

3.5. Final Form Assembly

Final forms were built based on the revised and expanded content taxonomy discussed in Section 3.2.3. The objective in form assembly was to create the required number of forms such that each form met content area specifications, the forms were as similar to each other as possible, and each form was representative of the domain being tested.

The form assembly procedure utilized IRT test information functions, where test information is the sum of item information conditional on theta (Lord, 1980; Hulin, Drasgow, & Parsons, 1983). Summation of item information to yield test information rests on the assumption of local independence, where local independence requires that responses to any two items are uncorrelated, given θ. When the ability being measured by a test is unidimensional, the occurrence of local independence follows automatically (Lord, 1980, p. 19). Test information functions, which are inversely related to error variance, have been widely studied and accepted, and are easily understood. This approach provides ease of computation and optimization because locally-independent items contribute to the total test information in an additive fashion, and linear programming can be used to obtain the optimal solution for the same reason. A maximum likelihood procedure was used to estimate item information.

The main goal was to assemble parallel forms for each subtest that would match or exceed the information functions of the subtests in Form 8A. Indeed, if the subtests could be improved by attaining higher information functions throughout the θ scale (-2.25 to +2.25), that would be a desirable outcome. The exception was AO, which was considered an experimental test at the time; instead, the goal was to match the new AO information function to that of the AO form that was in the field.

In the first phase, items were selected that produced an initial solution characterized by test forms with maximal weighted information. The second phase consisted of "shaping" – swapping items to obtain the best fit to the desired shape of the information function in the range of -2.25 through +2.25 on the θ scale. Swapping could occur between forms, or between forms and the pool of remaining items; any swapping that was done ensured that the resulting forms conformed to taxonomy requirements. Third, the forms were made as similar (i.e., parallel) as possible by minimizing the distance between form information functions. Fourth, an editorial review followed, to look for shortcomings such as items cueing one another.

The original plan was to construct Forms 23–26 entirely with new items from what is called the 6000 series, and to construct all forms at the same time. However, the CAT-ASVAB 3–4 pools that were being simultaneously developed were under populated, so the P&P items were "donated" to the CAT-ASVAB pools. After the CAT-ASVAB pools were constructed, unused items were returned to the P&P pool for use in Forms 23–26. In addition, items from STP Forms 18–19 were re-evaluated and added to the P&P pool. The complexity of drawing items from several sources on several occasions, replacing unused items, and constructing multiple P&P forms and CAT item pools (nearly) simultaneously led to an item overlap complication. A special study, described in Section 4, was conducted to alleviate concerns about item substitutions made to correct the problem.

Ultimately, more acceptable items were developed and tested than were used; the excess items were placed in the item bank.

4. Equating and Associated Studies (Forms 23-27)

This section provides details of equating and linking studies and other studies conducted in support of implementation of Forms 23–27. Two primary data collections, an OPCAL and an IOT&E, were required for the scoring, equating, and scaling of Forms 23–27. Additional studies were conducted to answer other important, but non-routine, questions that arose during late stages of preparing to launch the new tests. Taken together, the overall research program was complex, involving numerous studies with multiple ASVAB forms, but ultimately provided the data needed for equating/linking all forms and versions in the new and old subtest orders.

4.1. ASVAB Equating/Scaling Overview

New forms of the ASVAB are introduced at regular intervals to decrease the likelihood of test compromise and because some items may become dated. At any one time there are multiple forms of ASVAB being administered operationally across different modes of administration (P&P-ASVAB or CAT-ASVAB). These circumstances motivate the equating/scaling of all forms such that an examinee may take any one of them and be assured that there will be no disadvantage due to form or mode. All new ASVAB forms (P&P and CAT) are equated to a common reference form. Thus, all reported scores are on the same scale and can be treated interchangeably, regardless of form or mode taken.

Two major steps were followed in the equating of Forms 23–27: (1) An OPCAL study was conducted whereby Forms 23–27 were administered to a sample of recruits. The resulting data

were used to develop interim equating transformations, for use during data collection (with applicants) for a final equating. (2) An IOT&E study was conducted whereby Forms 23–27 were administered to applicants in an operational setting. The resulting data were used to develop final equating transformations. The interim equating results from the OPCAL were used to assign operational scores during the study.

In both studies, test form booklets were "spiraled" to ensure the different forms were administered to approximately equal numbers of randomly equivalent groups of examinees. The reference form was administered along with Forms 23–27. Both the interim and final equatings used a program for equipercentile equating of NR scores that featured log-linear smoothing with a power function fitted at the lower tail of the distribution. The equating program employed the means and standard deviations from the 1980 Youth Population norms (U.S. Department of Defense, 1982), thereby placing scores from Forms 23–27 on the same scale of measurement as other operational forms. Generally, there were only very small differences between the interim and final equating transformations based on NR scores. An IRT scaling approach to equating was also employed in the IOT&E and compared to the NR equating approach.

All equatings were conducted at the subtest level. Officially, none of the subtest scores are used individually; all subtest scores are used in the formation of composite scores only. There was no equating on the composite level, although the composites were evaluated to confirm that there was no undesirable impact on enlistment qualification rates, training school qualification rates, or job assignment.

4.2. OPCAL

The OPCAL was conducted at RTCs and provided data for the interim equating that served temporarily for qualifying applicants for enlistment. The OPCAL data collection took place from August, 1997–March, 1998.

Table 4.1 displays the forms administered, the subtest order, and the original form names for forms that had been operational previously. Examinees were randomly assigned to one of 12 forms. Forms 23–27, and 28C utilized the new subtest order (see Section 2.3.4), while Form 18H utilized the old order. Thus, separate testing rooms were required due to the incompatibility of instructions and timing. The number of examinees was evenly distributed across forms ($N \approx$ 3,000 per form).

Seven sites (see Table D. 1 in Appendix D) were used for the OPCAL, with most of the data collected at the Great Lakes Naval Training Center. Group equivalence across sites, gender, race, and education was confirmed; after data editing, N = 35,692 cases were suitable for analysis. Data distributions were smoothed using the best-fitting log linear polynomials (Hanson, 1991, as cited in Thomasson, Bloxom, & Wise, 1994). Equating was carried out with an equipercentile equating of NR score distributions using linear interpolation, the same methodology used in earlier equatings. (The methodology is described in Thomasson, Bloxom, & Wise, 1994). After equating, the composites were checked to confirm that the forms had similar score distributions. The differences found were within acceptable limits and were similar to those found in previous equatings.

Table 4.1. Forms Administered in the OPCAL

Form (Use)	Original Form Name	Subtest Order
23A and B (STP)	–	New
24A and B (STP)	–	New
25A and B (ETP)	–	New
26A and B (ETP)	–	New
27A (Holdout)	15A	New
27B (Holdout)	15B	New
28C (1980 Reference)	8A	New
18H (1980 Reference)	8A	Old

4.2.1. Item Overlap

As mentioned at the end of Section 3.5, an unintended item overlap occurred during form construction. After the OPCAL, it was discovered that 14 shop items on the AS subtest overlapped between the new P&P-ASVAB forms and operational CAT-ASVAB Form 03D, and 16 shop items from AS overlapped between the new P&P-ASVAB forms and the CAT-ASVAB reference form, 04D. By policy, item overlap between CAT-ASVAB item pools and P&P-ASVAB forms is not allowed due to the increased risk of compromise. A further check revealed that one P&P-ASVAB WK item and one P&P-ASVAB GS item overlapped with CAT-ASVAB Form 03D and one WK item overlapped within P&P-ASVAB forms. As it happened, the forms had not yet been administered in the IOT&E, so there was no operational impact.

The shop item overlap situation was remedied by revising the AS subtest in the P&P-ASVAB forms. Replacement items were obtained from the unused P&P-ASVAB tryout pool and from CAT-ASVAB pretest administrations (n = 121 items were available for use). Initial analyses indicated that the newly-reconstructed, content-balanced forms were psychometrically sound in terms of reliability and precision.

The adequacy of the substitution was examined by administering both the original and revised P&P-ASVAB forms for AS and computing new equatings of both forms. The CAT-ASVAB platform was used to administer the P&P-ASVAB forms for AS, but each P&P form was administered as a fixed form, rather than using an adaptive administration. The P&P AS forms were administered in place of AO at the end of the operational CAT-ASVAB battery. New equatings of the original AS forms were conducted using data from the special administration and compared with the OPCAL P&P-based equatings for the same original AS forms. It was found that generally the difference in equated standard scores fell within a 95 percent confidence interval except at low score values, which did not represent a significant problem because low-scoring applicants do not qualify to enlist. Thus, the transformations computed from the computerized administration of the revised AS forms were used as the interim equating transformations for AS during the IOT&E, while the transformations computed from the P&P-ASVAB OPCAL were used for all other subtests.

The WK and GS overlaps were allowed to stand, having been judged to represent little risk of test compromise. Each overlap involved only one item, and, furthermore, revising WK and GS would have imposed a significant research and development burden.

4.2.2. Omega Issue

Another issue that surfaced during studies supporting implementation was dubbed the "Omega Issue." The OPCAL was complete, and the IOT&E had not yet begun when it was discovered that four EI items distributed across six forms were affected by a faulty diagram that used a lower case "w" in place of the proper symbol, Ω (*omega).* The error affected the interim (OPCAL) equating transformations, but not the final operational transformations; data for the final transformations had not yet been collected. A study was conducted to determine whether the equating transformations from the OPCAL could be "repaired" without an additional data collection. Substitute items, chosen to be similar to the flawed items with respect to difficulty and discrimination, were used as a basis for modeling responses to the flawed items. A data set was constructed that consisted of a blend of observed and modeled responses, where modeled responses were substituted for responses to the flawed items. Frequency distributions, equating transformations, and composite distributions based on observed data were compared with frequency distributions, equating transformations, and composite distributions based on the blended data set. Composite scores were found to be affected to a very small degree. Ultimately, the test booklets were corrected and the blended approach was used to specify the equating transformations to be used for EI in the IOT&E.

4.3. IOT&E

In contrast to the OPCAL, which was conducted with recruits, the IOT&E enabled the collection of response data from applicants in an operational setting. The OPCAL transformations (with the corrected EI transformations and the AS tables based on the non-adaptive computer-administered data) were used during the IOT&E to qualify applicants for enlistment.

The IOT&E yielded the data for the final operational equating. It was conducted at MET sites in two phases to reduce the burden of test-booklet spiraling; the first phase was from November, 2000 – March, 2001 and the second from March, 2001 – July, 2001. Table 4.2 displays the forms administered, the original form names, and the phase in which each was administered. The IOT&E was designed to gather data for equating P&P-ASVAB Forms 23–27 to one another and to the reference form, 28C (formerly named Form 8A). Data collection took place at MET sites because only one testing room was needed, as all forms administered were in the new subtest order. All forms utilized the new order of subtest administration.

Each form administered in Phase 1 was taken by more than $N = 10{,}700$ examinees for a total greater than $N = 64{,}000$; in Phase 2, more than $N = 11{,}700$ responded to each form, for a total greater than $N = 70{,}000$. A form-assignment system was used to ensure that random assignment to form would be achieved. Within each phase, the number of examinees taking each form was roughly equal, and χ^2 analyses indicated that the groups were equivalent by gender, race, and education.

Two sets of analyses were conducted during the IOT&E: (a) analyses of the scoring methods/score scale and (b) analyses of the equating results. The first set of analyses evaluated applicant qualification rates across the old (NR) and new (IRT) scales/scoring methods, and the numbers and characteristics of applicants differentially selected by the two scoring procedures. The second set of analyses evaluated qualification rates associated with the equated forms, in terms of their similarity to the reference form and to each other. Results were compared across both the NR and IRT approaches to equating.

Table 4.2. Forms Administered in the IOT&E

Form (Use)	Original Form Name	Administered in IOT&E Phase	Sample Size
23A (STP)	–	1	10,752
23B (STP)	–	1	10,728
24A (STP)	–	2	11,932
24B (STP)	–	2	11,900
25A (ETP)	–	2	11,835
25B (ETP)	–	1	10,737
26A (ETP)	–	1	10,717
26B (ETP)	–	1	10,754
27A (Holdout)	15A	2	11,884
27B (Holdout)	15B	2	11,797
28C (1980 Reference)	8A	1	10,735
28C (1980 Reference)	8A	2	11,734

4.3.1. Scoring Methods/Score Scale Evaluation

In January 2002, a conversion to IRT scoring for P&P-ASVAB took place concurrently with implementation of the new P&P-ASVAB forms (see Section 2.3.1). Whereas NR scores had previously been the basis from which standard scores were computed, IRT scores were now used instead. Prior to implementing this conversion, the effects of the change in scoring method were evaluated. Results were evaluated for relevant cutscores for both the AFQT composite and the Service composites in place at the time of the study.

4.3.1.1. AFQT Scores.

Replacing NR scoring with IRT scoring was projected to have a trivial effect on enlistee flow-rates. An initial comparison of the qualification rates for IRT scoring and NR scoring at critical points on the AFQT scale indicated that slightly fewer applicants were qualified under IRT scoring. Adjustment of the transformation equation parameters (using weights) helped to equalize the rates, but the adjustments were constrained such that the transformation from BMEs to standard scores remained linear.

After the adjustment, individual examinees' data records from the reference form (28C) were scored by both methods and the impact on qualification rates was evaluated. The McNemar test

for nonindependent proportions (.01 level) was used to test the statistical significance of the difference between qualification rates. Tables E.1–E.4 in Appendix E show the results for the total group, females, blacks, and Hispanics. For the total group, there were some small differences in qualification rates over critical ranges in the AFQT distribution, near the 31st and 50th percentiles (.3 and .1, respectively, which were not statistically significant). Differences in qualification rates across scoring methods for females and blacks also were small. The difference at the 50th percentile was slightly larger for Hispanics, with the qualification rate for NR scoring exceeding the qualification rate for IRT scoring (Segall & Thomasson, 2001).

The analyses and evaluations were repeated for each of the new P&P-ASVAB forms. Tables E.5–E.20 in Appendix E show the results for relevant AFQT cutscores for the total group, females, blacks, and Hispanics. There were a few isolated significant differences in qualification rates where IRT scoring disadvantaged Hispanics (on one STP form and on the holdout Forms 27A and 27B). However, for most of the significant differences, IRT scoring provided an advantage for subgroup members. Researchers concluded that with respect to the AFQT, qualification rates based on IRT scoring were not lower at key cutscores than rates based on NR scoring, and the use of IRT scoring qualified about the same or slightly higher percentages of subgroups as NR scoring. The conclusion was that neither the total group nor any of the subgroups were placed at significant disadvantage at any AFQT cutscore on any of the new ETP forms when IRT scoring was used (Segall & Thomasson, 2001).

4.3.1.2. Service Selector Composites.

Service composite qualification rates were subjected to similar analyses using Form 28C. Results for the total group for the 79 cutscores examined are summarized in Table 4.3. The differences between qualification rates were small across NR and IRT scoring: 29 cutscores showed no significant difference (mean difference = 0.2), 48 cutscores showed a significant increase in qualification rates for IRT scoring (mean difference = 0.9), and two cutscores showed a significant decrease in qualification rates for IRT scoring (mean difference = -0.7).

Table 4.3. Comparison of Service Composite Qualification Rates for the Total Group Across Scoring Methods at Relevant Cutscores on Form 28C

Significance	# of Cutscores	Mean Difference
( )	29	0.2
(+)	48	0.9
(–)	2	-0.7

Notes:
( ) = No significant difference between IRT and NR qualification rates
(+) = Significant difference, IRT qualification rate > NR qualification rate
(–) = Significant difference, NR qualification rate > IRT qualification rate

The same cutscores were evaluated for differences in subgroup qualification rates. Table 4.4 summarizes the results for females, blacks, and Hispanics. Most qualification rates showed no significant difference across NR and IRT scoring. Where differences were significant, more

showed increases in qualification rates for IRT scoring (24) than decreases (8). Complete results by Service composite, group, cutscore, and form are reported in Segall and Thomasson (2001).

All the evidence pointed toward the conclusion that the differences in qualification rates observed across method of scoring were small enough for both AFQT and Service composites that implementation of IRT scoring would have a negligible effect; moreover, the differences found were typical of those seen across operational forms.

Table 4.4. Comparison of Service Composite Qualification Rates by Subgroup Across Scoring Methods at Relevant Cutscores on Form 28C

	Females (N=2690)		Blacks (N=2877)		Hispanics (N=1293)	
Significance	# of Cutscores	Mean Difference	# of Cutscores	Mean Difference	# of Cutscores	Mean Difference
( )	58	0.2	73	0.3	74	-0.3
(+)	19	1.2	4	1.4	1	1.8
(–)	2	-0.8	2	-0.7	4	-1.9

Notes:
( ) = No significant difference between IRT and NR qualification rates
(+) = Significant difference, IRT qualification rate > NR qualification rate
(–) = Significant difference, NR qualification rate > IRT qualification rate

4.3.2. Equating/Scaling Evaluation

Two approaches to equating were used in the final equating of Forms 23–27. The traditional NR equating approach was employed (discussed in Sections 4.1 and 4.2), along with an IRT scaling approach, and results were compared across the two approaches. Given the prospective change from NR scoring to IRT scoring, it was necessary to evaluate the effect of the IRT-based scoring and scaling method to ensure there would be no adverse effect on qualification rates and that scores could be treated interchangeably across forms.

In IRT, the item response function, viewed as the regression of item score on ability, is invariant in the sense that the parameters defining it (a_i, b_i, and c_i) do not depend on the distribution of ability in the particular group of examinees responding to the item. The invariance of item parameters (after the origin and unit of measurement are set) is one of the most important characteristics of IRT. "Ability parameters θ are also invariant from one test to another except for choice of origin and scale, assuming that the tests both measure the same ability, skill, or trait" (Lord, 1980, p. 37). Kolen and Brennan (1995, p. 167) state that "In the random groups equating design, the IRT parameters for Form X can be estimated separately from the parameters for Form Y. If the same scaling convention (for example, mean of 0 and standard deviation of 1) for ability is used in the separate estimations, then the parameter estimates for the two forms are assumed to be on the same scale without further transformation . . because the groups are randomly equivalent and the abilities are scaled to have the same mean and standard deviation in both groups."

The forms to be equated (i.e., Forms 23–27) were administered to randomly equivalent groups and calibrated separately using BILOG (Mislevy & Bock, 1990) assuming the same mean and

standard deviation for the underlying ability distribution for each form. Thus, as per Kolen and Brennan (1995), an IRT scaling should be sufficient to yield interchangeable scores across forms. Comparisons of results for the NR equating approach (discussed in Sections 4.1 and 4.2) and the IRT scaling approach addressed the veracity of this assertion.

In the IRT scaling approach, data collected from applicants during the IOT&E were used to obtain 3-PL item parameter estimates (Equation 4). Parameters from Phase II of the IOT&E were placed on the metric of Phase I parameters using the Stocking and Lord transformation (Stocking & Lord, 1983). Ability estimates were computed and transformed to standard scores on the 1980 reference scale using a common transformation across Forms 23–27 for each subtest. [10]

4.3.2.1. Form Qualification Rate Agreement

Table 4.5 summarizes the qualification rate agreement across forms for NR equating and IRT scaling. The equating/scaling evaluation criteria were the root mean squared differences by scoring method and the standard deviations by scoring method for the AFQT and Service composites. The root mean squared difference analysis compared Form 28C qualification rates with qualification rates for Forms 23–27, and the standard deviation comparison was conducted using qualification rates for Forms 23–27. The differences between Forms 23–27 and reference-form qualification rates tended to be smaller for the IRT scaling than the NR equating, and the similarity among qualification rates across Forms 23–27 tended to be higher for the IRT scaling than the NR equating. The results supported the use of IRT scaling and a common θ-to-standard-score transformation across Forms 23–27 for each subtest (Segall & Thomasson, 2001).

4.3.2.2. Score Gaps and Score Ceilings

Simulation studies were conducted to evaluate the possibility of score gaps or score ceilings. The results indicated that the distribution of standard scores based on IRT scoring/equating would have no score gaps and would have higher score ceilings, as well as marginally higher reliability (relative to NR scoring/equating). Shortly after implementation, it became apparent that, as predicted, score gaps were far fewer for IRT scoring/scaling as compared to NR scoring/equating. The expected improvement in ceiling effects also was seen, with ceilings increasing for all subtests in the battery.

4.3.2.3. Construct Validity and Precision Using IRT and NR Scoring Procedures

Another issue to be resolved was whether the IRT and NR scoring/equating methods differed in terms of validity and/or precision. More specifically, the questions were whether IRT scoring/equating altered the constructs measured and whether it increased precision. Data collected previously from 687 Navy recruits were analyzed to answer the questions. Each examinee had taken three P&P-ASVAB forms, one pre-enlistment and two post-enlistment.

[10] The transformations were determined using the reference form (28C) and then applied to Forms 23–27. For each subtest, the IRT scores were transformed to have the same mean and variance as standard scores based on NR scoring for the reference form.

Two predictions were made: (a) If the proposed IRT scoring/scaling yielded reliable scores, then two alternate forms scored by IRT should correlate more highly than the same two forms scored/equated by NR; and (b) IRT scoring/scaling, as compared to NR scoring/equating, should show improved prediction of NR-based scores on an alternate form. The power subtests, the AFQT, and the Service composites were evaluated in this manner, and the results confirmed expectations: both predictions were upheld. The results thus supported the construct validity and precision of the IRT-based approach to scoring and equating.

Table 4.5. Form Qualification Rate Agreement for NR Equating and IRT Scaling

		RMSD		SD	
Service	Composite[a]	IRT	NR	IRT	NR
All	AFQT	0.6926	0.6297	0.2987	0.2420
Army	GT	0.4682	0.6852	0.2533	0.1759
	GM	0.7014	0.8761	0.2297	0.3680
	EL	0.7991	0.9252	0.2673	0.3309
	CL	0.5699	0.8473	0.2425	0.2889
	MM	1.0077	1.0066	0.2636	0.4162
	SC	0.6887	0.9146	0.2943	0.3240
	CO	0.6551	0.8479	0.2291	0.3843
	FA	0.7019	0.8399	0.2398	0.3642
	OF	0.8112	0.9081	0.2430	0.3805
	ST	0.5986	0.8814	0.2052	0.2978
Navy	GT	0.5837	0.6070	0.2892	0.2118
	EL	0.5490	0.6273	0.2926	0.3211
	BEE	0.3805	0.3472	0.2236	0.1970
	EMG	0.9263	0.8440	0.2169	0.5062
	MEC	0.4480	0.5771	0.2563	0.3077
	HM	0.4555	0.5316	0.2896	0.1623
Air Force	M	0.8858	0.9601	0.2357	0.3110
	A	0.6460	0.8361	0.2619	0.2670
	G	0.5126	0.6474	0.2616	0.2132
	E	0.5167	0.6281	0.2802	0.3144
Marines	MM	0.7469	0.9378	0.3218	0.3430
	GT	0.4279	0.7029	0.2799	0.3294
	EL	0.7938	0.5940	0.4007	0.3205
	Average	0.6563	0.7618	0.2746	0.3048

[a]Results are based on Service composite definitions that were in place at the time of the analysis. Some composites now differ for some services. Current composite definitions are given in Table C.1.

4.4. Anchoring Study

A separate study called the Anchoring Study was conducted independently of the equating studies. Data collection was conducted from October, 2000 through April, 2001. Analyses for the Anchoring Study were conducted in two phases. Phase I analyses examined whether reordering

the subtests would have an impact on P&P-ASVAB scores.[11] Phase II analyses were conducted to link Forms 23–27 to a new reference form (CAT-ASVAB 04D), as part of the work to develop new norms. Phase I will be discussed here, while Phase II will be discussed in Section 5.

Data collection for Phase I of the Anchoring Study occurred from October 2000 through January 2001. Testing was conducted at MEPS with applicants, because data for the IOT&E was being concurrently collected in MET sites. A total of 11 MEPS participated (identified in Table D.2 in Appendix D), resulting in a sample size of N = 25,189. The MEPS were selected because they had multiple testing rooms available for simultaneous administration of forms in the old and new subtest order. Table 4.6 specifies the old and new orders for the P&P-ASVAB subtests. Note that only intact subtests were re-ordered; the items within the subtests remained in the same order.

Table 4.6. P&P-ASVAB Subtests in Old and New Orders[a]

Old Order	New Order
General Science	General Science
Arithmetic Reasoning	Arithmetic Reasoning
Word Knowledge	Word Knowledge
Paragraph Comprehension	Paragraph Comprehension
Numerical Operations	Mathematics Knowledge
Coding Speed	Electronics Information
Auto & Shop Information	Auto & Shop Information
Mathematics Knowledge	Mechanical Comprehension
Mechanical Comprehension	Assembling Objects
Electronics Information	Coding Speed
	Numerical Operations

[a]Coding Speed and Numerical Operations were administered for this study, but were dropped from the battery with the implementation of the new P&P-ASVAB forms.

Table 4.7 displays the forms administered in Phase I of the Anchoring Study. The study design utilized random assignment to forms to obtain randomly equivalent groups taking each form. The 1980 reference form was administered simultaneously in both the old order and the new order, thus, separate testing rooms were required due to the incompatibility of instructions and timing.

Table 4.7. Forms Administered in Phase I of the Anchoring Study

Form (Use)	Subtest Order	Original Form Name
15H (1980 Reference)	Old	8A
25B (ETP)	New	
28C[a] (1980 Reference)	New	8A

[a] Form 28C did not include Coding Speed or Numerical Operations for this study.

[11] See ASVAB Technical Bulletin No. 2 (DMDC, 2009) for details of studies related to CAT-ASVAB reordering.

To assess order effects on the subtest level, scores were compared across Forms 15H (the 1980 reference form in the old order) and 28C (the 1980 reference form in the new order). Both NR and IRT scoring were used. Equating tables from the IOT&E were used to convert NR scores to standard scores, while the IRT ability estimates (BMEs) were transformed to standard scores using the linear transformations obtained in the IOT&E. Order effects were assessed in three ways: subtest standard score means were compared, standard score distributions were compared using the Kolmogorov-Smirnov (K-S) Test, and composite score distributions were compared using the K-S Test.

Table 4.8 summarizes the standard score mean differences and effect sizes across the old and new orders, for each method of scoring. Mean differences were significant for MC (lower score for new order), EI (higher score for new order), and AS (higher score for new order). Except for MC, all subtests had a magnitude difference less than or about ± 0.5 standard score units, which is similar to the magnitude of rounding error. Sampling error, as evident in the first four tests (because they did not change order), could be as much as ± -0.2 standard score units. Effect sizes were small for all subtests. For all subtests that changed positions, the order effects were in the direction predicted by a fatigue factor (the earlier the presentation, the higher the scores). The direction and magnitude of the order effects were similar for both NR and IRT scoring.

Table 4.9 summarizes the results of the K-S test comparing differences in cumulative distribution functions (CDF) for the standard scores across the old and new orders. A largest CDF difference (15H-28C) > 0 implies that the new order is favored for applicants at that hypothetical cutscore. A largest CDF difference (15H-28C) < 0 implies the old order is favored for applicants at the hypothetical cutscore. Only MC and EI showed significant K-S differences ($p < 0.05$) in cumulative distribution functions, with MC favoring the old order and EI favoring the new order. The largest CDF differences were less than 4% at the most problematic cutscores. Results were similar for both NR and IRT scoring.

Table 4.10 summarizes the results of the K-S test comparing differences in cumulative distribution functions (CDF) for the composite scores across the old and new orders. Results are presented only for IRT scoring. None of the composites showed significant K-S test statistics, indicating that the order effects seen in some individual subtests were greatly diluted in the composites. The results suggested that it was not necessary to adjust reported subtest scores on Forms 23–27 for order of presentation because none of the composite scores were significantly affected by subtest order.

4.5. Implementation of Forms 23–27

Forms 25–26 were implemented in the ETP on January 2, 2002 and Forms 23–24 were implemented in the STP six months later, on July 2, 2002. STP Forms 18–19 and ETP Forms 20–22 were retired upon implementation of the new forms. Form 27 was placed on reserve for use in the case of test compromise.

Table 4.8. Score Mean Differences and Effect Sizes Across Old and New Orders

Subtest	Scoring Method	Form	Subtest Order	N	Mean	SD	Mean Difference	Effect Size
GS	NR	15H	Old	5106	49.83	8.35	0.03	0.00
		28C	New	5056	49.86	8.47		
	IRT	15H	Old	5106	49.83	8.44	0.07	0.01
		28C	New	5056	49.90	8.52		
AR	NR	15H	Old	5106	49.81	8.24	0.05	0.01
		28C	New	5056	49.87	8.36		
	IRT	15H	Old	5106	50.06	7.98	0.00	0.00
		28C	New	5056	50.07	8.10		
WK	NR	15H	Old	5106	50.94	6.95	-0.21	-0.03
		28C	New	5056	50.73	7.08		
	IRT	15H	Old	5106	50.78	6.89	-0.18	-0.03
		28C	New	5056	50.60	6.98		
PC	NR	15H	Old	5106	50.99	7.85	-0.16	-0.02
		28C	New	5056	50.84	7.90		
	IRT	15H	Old	5106	51.54	7.82	-0.19	-0.02
		28C	New	5056	51.35	7.86		
AS	NR	15H	Old	5106	47.45	8.58	0.51*	0.06
		28C	New	5056	47.96	8.63		
	IRT	15H	Old	5106	47.14	8.30	0.43*	0.05
		28C	New	5056	47.57	8.28		
MK	NR	15H	Old	5106	51.99	8.18	0.28	0.03
		28C	New	5056	52.27	8.14		
	IRT	15H	Old	5106	51.98	8.16	0.31	0.04
		28C	New	5056	52.29	8.18		
MC	NR	15H	Old	5106	49.96	9.07	-0.78**	-0.09
		28C	New	5056	49.18	9.24		
	IRT	15H	Old	5106	50.56	8.40	-0.78**	-0.09
		28C	New	5056	49.79	8.66		
EI	NR	15H	Old	5106	48.59	8.05	0.46*	0.06
		28C	New	5056	49.05	8.18		
	IRT	15H	Old	5106	47.76	8.10	0.57**	0.07
		28C	New	5056	48.32	8.21		

Note. *p < 0.01 **p<.001

Table 4.9. Differences in CDFs for Standard Scores Across Old and New Orders.

Subtest	Scoring Method	Largest CDF Difference	Kolmogorov-Smirnov Z	Asymptotic Significance
GS	NR	0.010	0.514	0.955
	IRT	0.007	0.339	1.000
AR	NR	0.014	0.713	0.690
	IRT	0.014	0.726	0.667
WK	NR	-0.018	0.883	0.417
	IRT	-0.013	0.656	0.782
PC	NR	-0.010	0.521	0.949
	IRT	-0.015	0.767	0.599
AS	NR	0.024	1.203	0.111
	IRT	0.022	1.131	0.155
MK	NR	0.017	0.872	0.432
	IRT	0.021	1.051	0.219
MC	NR	-0.038	1.937	0.001
	IRT	-0.039	1.944	0.001
EI	NR	0.033	1.676	0.007
	IRT	0.030	1.499	0.022

Table 4.10. Differences in CDFs for Composite Scores Across Old and New Orders.

Service	Composite [a]	Largest CDF Difference	Kolmogorov-Smirnov Z	Asymptotic Significance
All	AFQT	-0.013	0.643	0.803
Army	CL	-0.012	0.600	0.865
	CO	0.017	0.871	0.434
	EL	0.017	0.841	0.479
	FA	0.014	0.694	0.721
	GM	0.021	1.063	0.209
	MM	0.014	0.707	0.700
	OF	0.016	0.817	0.517
	SC	0.015	0.738	0.647
	ST	0.013	0.647	0.796
Navy	GT	-0.017	0.849	0.467
	MEC	-0.020	0.998	0.273
	EL	0.023	1.172	0.128
	BEE	0.015	0.754	0.621
	ENG	0.026	1.333	0.057
	HM	0.010	0.511	0.957
	MR	-0.017	0.855	0.457
	ADM	0.009	0.436	0.991
	NUC	-0.020	0.984	0.287
Air Force	M	-0.016	0.821	0.511
	A	0.008	0.396	0.998
	G	-0.017	0.843	0.475
	E	0.023	1.153	0.140
Marine Corps	MM	0.015	0.732	0.658
	CL	-0.012	0.609	0.852
	GT	-0.026	1.314	0.063
	EL	0.023	1.153	0.140
Coast Guard	AT	-0.017	0.843	0.475
	BT	-0.020	0.994	0.276
	CT	0.010	0.494	0.967
	DT	-0.018	0.897	0.397
	ET	0.023	1.153	0.140
	FT	0.026	1.330	0.058

[a] Results are based on Service composite definitions that were in place at the time of the analysis. Some composites now differ for some services. Current composite definitions are given in Table C.1 in Appendix C.

5. Norms for Forms 23–27

Norms provide a summary of test performance for a group of examinees. The group (referred to here as the reference group) is typically a large sample of examinees that is representative of the examinee population of interest. Norms are developed using the reference group, and then applied to the scores of individual examinees to summarize their performance relative to the performance of the reference group. National norms are norms that are developed using a nationally representative sample of examinees that are at the age or educational level for which the test is developed. National norms provide a basis for evaluating performance for all examinees nationwide. National norms are created by conducting national norming studies. Because the accuracy and usefulness of norms may diminish over time due to changes in demographics and population educational achievement, it is necessary to evaluate and perhaps renorm at regular intervals (American Educational Research Association, 1999).

When Forms 23–27 were implemented in 2002, the 1980 norms were in effect. The 1980 standard score scale was developed using data collected by administering P&P-ASVAB Form 8A to a nationally representative sample of American youth (U. S. Department of Defense, 1982). The study was called the PAY80. In July, 2004, new national norms for the ASVAB were implemented, based on data collected as part of the Profile of American Youth 1997 (PAY97) study. This section provides a summary of the 1997 ASVAB norming study, PAY97, and describes an equating study linking P&P-ASVAB Forms 23–27 to the new 1997 score scale. More details are provided in (Segall, 2004). The PAY97 norms are the norms currently in effect for ASVAB Forms 23–27.

5.1. 1997 Norming Study

PAY97 (Segall, 2004) was done in conjunction with a study being conducted by the DoL called the National Longitudinal Survey of Youth 1997 (NLSY97). Since 1980, there had been changes in performance on standardized tests and substantial changes in demographics (MaCurdy & Vytlacil, 2003; Wise & Curran, 1995; Welsh, 2003), and norms were needed for two new measures that had been developed, one of which was AO. Also, the ASVAB was being administered adaptively via computers to an ever-increasing percentage of applicants, while the 1980 norms were based on P&P administration and classical NR scoring methods. Kolen and Brennan (1995) caution against assuming that measures of the same construct administered in different modes are comparable.

Five preliminary studies, conducted from May, 1995 to November, 1996, preceded the main study. The purposes of the studies were to (a) determine the appropriateness of financial incentives and method of payment for minors, (b) test procedures and incentive amounts, (c) determine the appropriateness of administering CAT-ASVAB to the 12–14 year-olds in the NLSY97 sample, (d) gather additional data on a participation incentive, and (e) conduct a “dress rehearsal” for the main study.

In early 1997, housing units were listed in the selected primary sampling units. A multi-stage probability sampling plan was employed wherein over 90,000 housing units were selected for

screening. Hispanic and non-Hispanic black youths were oversampled to ensure adequate subgroup sample size. Eligible participants were assigned to one or more of three age-group samples, the ETP sample (ages 18–23), the STP sample (persons in grades 10–12 or in a post-secondary school), and the NLSY97 sample (ages 12–16). Testing followed from June 1997 until April 1998. Participants were administered CAT-ASVAB Form 04D under standardized conditions. Test administration was done under contract at a chain of commercial testing centers and at a few temporary centers set up specifically for the study. The final sample size for the ETP was N = 5,997 and the STP sample was composed of N = 4,655 students. The samples were weighted to ensure that they were nationally representative of the respective populations. The ETP norms apply to the military-eligible population of youth and the STP norms are grade-by-gender norms. See Segall (2004) for details of the PAY97 ETP score scale development process. Segall also describes analyses of the impact of the new scale on qualification rates for the AFQT, Service composites, and subgroups. Information on the STP score scale development and STP norms may be found Hiatt & Sims (2003).

5.2. Test Form Equating

Before the 1997 score scale could be applied to P&P-ASVAB Forms 23–27, it was necessary to equate them to the new reference form administered in the PAY97 norming study (CAT-ASVAB Form 04D). Data collected in Phase II of the Anchoring study (see Section 4.4) were used to conduct a direct linear equating between P&P-ASVAB Form 25B and CAT-ASVAB Form 04D. The obtained subtest equating transformations were then applied to the other P&P-ASVAB forms (possible to do because ability estimates among the P&P-ASVAB forms are treated as interchangeable). The equating helped to ensure that the P&P ability estimates were placed on a metric comparable to that of the CAT reference form (04D) and that applicants could be indifferent with regard to mode of administration (P&P versus CAT). More details about the equating and form equivalence analyses are reported in Segall (2004).

6. Statistical and Psychometric Properties of Forms 23–27

This section provides details of the statistical and psychometric properties of Forms 23–27.

6.1. Subtest Moments

Means and standard deviations (in parentheses) for Forms 23–27 are displayed in Table 6.1. The moments of the distributions were calculated using subtest NR scores. Table 6.2 displays the Forms 23–27 means and standard deviations (in parentheses) for the IRT BMEs.

6.2. Subtest Intercorrelations

Data from the IOT&E Phase 1 forms were combined to calculate subtest correlations, as displayed in Table 6.3. Likewise, Phase 2 subtest correlations (reported in Table 6.4) are based on data across Phase 2 forms. The pattern of correlations was fairly similar across phases; where there were differences between phases, the magnitude was mostly .04 or less. The greatest difference between phases was the correlation between EI and PC, where they differed by .07.

The lowest correlation between subtests in Phase 1 was $r = .21$ between AS and MK, while the highest was $r = .73$ between AS and AR. In Phase 2, again AS and MK had the lowest correlation at $r = .22$, and WK and GS were correlated most highly at $r = .71$. Tables F.1 through F.10 in Appendix F display IRT score (BME) correlations among subtests for each form separately, and Tables G.1 through G.10 in Appendix G display the corresponding statistics for responses scored as number right.

Table 6.1. Forms 23–27 Subtest NR Score Means and Standard Deviations[a,b]

	GS	AR	WK	PC	MK	EI	AS	MC	AO[c]
23A	14.57	14.23	23.30	9.99	13.27	10.87	12.78	13.92	18.07
N = 10,752	(4.40)	(6.08)	(6.76)	(3.13)	(5.40)	(3.49)	(5.33)	(4.04)	(5.24)
23B	14.57	14.27	23.37	9.87	13.29	10.90	12.75	13.80	18.21
N = 10,728	(4.47)	(6.10)	(6.82)	(3.14)	(5.48)	(3.53)	(5.35)	(4.06)	(5.28)
24A	14.27	15.09	21.33	10.62	13.61	9.41	11.44	13.04	18.44
N = 11,932	(4.77)	(5.88)	(6.38)	(2.73)	(4.97)	(3.39)	(5.03)	(4.29)	(5.69)
24B	14.30	15.06	21.51	10.64	13.42	9.53	11.56	13.16	18.45
N = 11,900	(4.76)	(5.90)	(6.41)	(2.71)	(5.04)	(3.40)	(5.11)	(4.26)	(5.65)
25A	13.66	15.32	22.42	10.29	13.58	10.89	11.22	13.86	18.76
N = 11,835	(4.10)	(6.34)	(5.82)	(2.81)	(5.19)	(3.47)	(5.28)	(4.61)	(5.25)
25B	13.90	15.02	22.79	10.90	13.77	10.96	11.52	13.93	19.01
N = 10,737	(4.10)	(6.12)	(5.84)	(3.15)	(5.32)	(3.48)	(5.37)	(4.73)	(5.14)
26A	12.99	16.89	22.92	10.36	13.79	11.14	11.06	13.39	18.22
N = 10,717	(4.36)	(5.75)	(6.46)	(3.17)	(5.44)	(3.48)	(5.65)	(4.11)	(5.56)
26B	13.16	15.42	22.70	9.88	13.89	11.14	11.22	13.42	18.11
N = 10,754	(4.37)	(6.24)	(5.95)	(3.07)	(5.54)	(3.43)	(5.69)	(4.11)	(5.56)
27A	15.03	16.76	24.90	12.18	13.83	10.47	12.65	14.01	18.25
N = 11,884	(4.42)	(5.96)	(6.70)	(2.64)	(5.00)	(3.24)	(4.90)	(4.65)	(5.46)
27B	15.03	16.68	25.11	11.35	13.99	10.35	12.42	13.91	18.09
N = 11,797	(4.36)	(6.15)	(5.96)	(2.82)	(5.14)	(3.33)	(4.87)	(4.65)	(5.47)

[a] Data source is the IOT&E.
[b] Standard deviations are in parentheses.
[c] Although AO is contained in the test booklets for Forms 23–24, it is not administered during CEP/STP testing.

Table 6.2. Forms 23–27 Subtest IRT Score (BME) Means and Standard Deviations[ab]

Form	Phase	GS	AR	WK	PC	MK	EI	AS	MC	AO[c]
23A	1	0.05	0.08	0.02	0.01	0.07	0.06	0.10	0.05	-0.02
		(.86)	(.86)	(.91)	(.83)	(.85)	(.83)	(.82)	(.86)	(.86)
23B	1	0.05	0.08	0.02	0.02	0.07	0.06	0.10	0.06	-0.02
		(.85)	(.87)	(.91)	(.82)	(.85)	(.83)	(.82)	(.87)	(.85)
24A	2	0.04	0.07	0.03	-0.01	0.06	0.09	0.13	0.06	-0.03
		(.84)	(.87)	(.90)	(.81)	(.87)	(.82)	(.80)	(.88)	(.85)
24B	2	0.05	0.08	0.03	0.01	0.07	0.08	0.12	0.06	-0.03
		(.83)	(.86)	(.90)	(.82)	(.86)	(.82)	(.82)	(.87)	(.85)
25A	2	0.06	0.05	0.02	0.04	0.09	0.07	0.10	0.08	-0.03
		(.86)	(.88)	(.91)	(.82)	(.82)	(.84)	(.83)	(.84)	(.86)
25B	1	0.06	0.06	0.02	-0.02	0.06	0.07	0.09	0.07	-0.03
		(.86)	(.88)	(.92)	(.81)	(.85)	(.83)	(.83)	(.85)	(.86)
26A	1	0.07	0.05	0.02	0.01	0.06	0.07	0.11	0.07	-0.02
		(.84)	(.89)	(.91)	(.81)	(.85)	(.82)	(.81)	(.85)	(.85)
26B	1	0.07	0.05	0.02	0.02	0.06	0.07	0.10	0.07	-0.02
		(.84)	(.88)	(.91)	(.81)	(.86)	(.83)	(.82)	(.85)	(.85)
27A	2	0.06	0.05	0.01	-0.04	0.06	0.14	0.11	0.06	-0.02
		(.86)	(.90)	(.91)	(.80)	(.88)	(.80)	(.82)	(.86)	(.85)
27B	2	0.07	0.06	0.01	-0.02	0.06	0.14	0.11	0.06	-0.02
		(.86)	(.87)	(.92)	(.81)	(.87)	(.79)	(.82)	(.85)	(.85)

[a] Data source is IOT&E.

[b] Standard deviations are in parentheses.

[c] Although AO is contained in the test booklets for Forms 23–24, it is not administered during CEP/STP testing.

Table 6.3. IOT&E Phase 1[a] Subtest NR Mean[b] Intercorrelations[c]

	GS	AR	WK	PC	MK	AS	MC	EI	AO
GS	1.00								
AR	0.56	1.00							
WK	0.71	0.58	1.00						
PC	0.59	0.58	0.70	1.00					
MK	0.52	0.37	0.47	0.34	1.00				
AS	0.53	0.73	0.48	0.52	0.21	1.00			
MC	0.62	0.56	0.55	0.48	0.62	0.47	1.00		
EI	0.62	0.53	0.60	0.51	0.61	0.44	0.65	1.00	
AO	0.43	0.49	0.38	0.41	0.32	0.48	0.53	0.41	1.00

[a] Includes Forms 23A, 23B, 25B, 26A, 26B, and 28C.

[b] Data from all Phase 1 forms were combined to calculate mean correlations; $N = 64{,}423$.

[c] All correlations were significant at $p < .0001$.

Table 6.4. IOT&E Phase 2[a] Subtest NR Mean[b] Intercorrelations[c]

	GS	AR	WK	PC	MK	AS	MC	EI	AO
GS	1.00								
AR	0.59	1.00							
WK	0.71	0.59	1.00						
PC	0.59	0.54	0.68	1.00					
MK	0.52	0.42	0.48	0.35	1.00				
AS	0.53	0.70	0.47	0.47	0.22	1.00			
MC	0.63	0.59	0.55	0.47	0.63	0.47	1.00		
EI	0.61	0.51	0.56	0.44	0.62	0.38	0.63	1.00	
AO	0.42	0.48	0.36	0.36	0.31	0.46	0.52	0.38	1.00

[a] Includes Forms 24A, 24B, 25A, 27A, 27B, and 28C.
[b] Data from all Phase 2 forms were combined to calculate mean correlations; N = 71,082.
[c] All correlations were significant at $p < .0001$.

6.3 Item Parameters

Tables 6.5–6.13 summarize the item parameters for Forms 23–27 for each P&P-ASVAB subtest. The tables present the average, standard deviation, minimum, and maximum values observed over all items in a form.

Table 6.5. Summary of Item Parameters for GS Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
a	Ave	0.82	0.85	0.77	0.77	0.90	0.92	0.95	0.95	0.92	0.92
	SD	0.30	0.27	0.15	0.20	0.36	0.37	0.33	0.26	0.34	0.36
	Min	0.48	0.52	0.45	0.41	0.28	0.25	0.50	0.55	0.52	0.51
	Max	1.70	1.67	1.03	1.18	1.63	1.70	1.86	1.58	1.65	1.61
b	Ave	0.12	0.15	0.19	0.19	0.26	0.27	0.70	0.66	-0.16	0.12
	SD	1.03	1.03	0.87	0.82	2.02	2.08	1.26	1.25	1.30	1.32
	Min	-1.63	-1.72	-1.53	-1.34	-6.94	-7.42	-1.76	-1.50	-2.92	-3.03
	Max	2.06	2.08	2.08	1.92	3.34	3.13	3.63	3.46	1.69	1.63
c	Ave	0.22	0.23	0.21	0.22	0.22	0.24	0.25	0.25	0.20	0.22
	SD	0.11	0.10	0.08	0.10	0.10	0.10	0.10	0.10	0.08	0.09
	Min	0.10	0.10	0.09	0.10	0.09	0.08	0.08	0.08	0.10	0.11
	Max	0.44	0.46	0.46	0.48	0.41	0.50	0.48	0.45	0.41	0.42

Table 6.6. Summary of Item Parameters for AR Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	1.18	1.21	1.05	1.06	1.14	1.14	1.20	1.11	1.22	1.09
a	SD	0.30	0.30	0.28	0.31	0.35	0.26	0.38	0.37	0.45	0.33
	Min	0.64	0.66	0.62	0.58	0.48	0.60	0.50	0.47	0.49	0.60
	Max	1.89	1.78	1.92	2.06	1.70	1.79	2.18	1.99	2.29	1.73
	Ave	0.58	0.58	0.49	0.48	0.32	0.47	0.21	0.25	0.17	0.11
b	SD	0.76	0.74	0.80	0.81	0.84	0.90	1.05	1.04	0.94	1.00
	Min	-1.33	-1.20	-1.51	-1.57	-1.55	-1.99	-2.75	-3.18	-1.57	-2.79
	Max	1.84	1.74	1.53	1.59	1.90	1.80	2.58	1.83	1.59	1.44
	Ave	0.20	0.21	0.21	0.21	0.17	0.20	0.23	0.17	0.21	0.21
c	SD	0.11	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.10	0.10
	Min	0.07	0.07	0.06	0.05	0.05	0.05	0.06	0.07	0.06	0.09
	Max	0.42	0.38	0.49	0.46	0.43	0.43	0.47	0.33	0.44	0.50

Table 6.7. Summary of Item Parameters for WK Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	1.03	1.08	0.97	0.98	1.02	1.08	1.15	1.05	1.24	1.25
a	SD	0.25	0.27	0.29	0.34	0.33	0.46	0.39	0.37	0.42	0.51
	Min	0.43	0.40	0.51	0.53	0.29	0.30	0.17	0.20	0.41	0.47
	Max	1.59	1.81	1.75	1.95	1.88	2.43	2.14	1.64	2.37	2.86
	Ave	-0.36	-0.34	-0.03	-0.03	-0.09	-0.24	-0.13	-0.15	-0.56	-0.49
b	SD	0.97	0.96	1.13	1.16	1.77	1.34	1.29	1.35	1.05	1.26
	Min	-2.38	-2.28	-2.34	-2.46	-2.87	-2.39	-2.70	-2.69	-3.67	-3.82
	Max	1.77	1.82	2.66	3.08	6.77	3.03	3.74	3.18	1.69	1.90
	Ave	0.21	0.22	0.22	0.22	0.22	0.21	0.23	0.23	0.22	0.25
c	SD	0.09	0.09	0.12	0.12	0.11	0.12	0.11	0.11	0.11	0.14
	Min	0.08	0.10	0.05	0.06	0.07	0.07	0.07	0.08	0.07	0.03
	Max	0.49	0.50	0.49	0.50	0.50	0.49	0.50	0.50	0.45	0.50

Table 6.8. Summary of Item Parameters for PC Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	0.90	0.88	0.82	0.86	0.88	0.93	0.99	0.88	0.93	1.04
a	SD	0.20	0.20	0.24	0.23	0.38	0.22	0.25	0.31	0.20	0.49
	Min	0.61	0.51	0.45	0.56	0.32	0.71	0.56	0.40	0.50	0.41
	Max	1.26	1.25	1.21	1.23	1.78	1.57	1.43	1.52	1.38	1.98
	Ave	-0.36	-0.33	-0.58	-0.59	-0.76	-0.71	-0.38	-0.40	-1.16	-0.94
b	SD	0.85	0.82	1.31	1.26	1.29	0.74	0.82	0.95	0.69	1.23
	Min	-1.51	-1.53	-2.44	-2.46	-2.60	-1.88	-1.50	-1.82	-2.00	-4.45
	Max	0.90	0.95	2.46	2.13	0.90	0.67	1.60	1.26	0.34	0.91
	Ave	0.19	0.19	0.20	0.22	0.17	0.18	0.22	0.18	0.21	0.22
c	SD	0.09	0.08	0.10	0.10	0.06	0.08	0.08	0.08	0.11	0.10
	Min	0.06	0.07	0.06	0.08	0.12	0.08	0.08	0.07	0.09	0.04
	Max	0.32	0.33	0.41	0.41	0.28	0.36	0.36	0.43	0.43	0.36

Table 6.9. Summary of Item Parameters for MK Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	1.08	1.11	1.03	1.03	1.08	1.04	1.06	1.26	1.17	1.22
a	SD	0.38	0.36	0.42	0.42	0.37	0.39	0.33	0.62	0.48	0.48
	Min	0.24	0.20	0.55	0.36	0.33	0.42	0.47	0.54	0.68	0.57
	Max	1.73	1.74	2.05	1.94	1.87	2.22	1.67	2.58	2.65	2.21
	Ave	0.22	0.21	0.36	0.39	0.23	0.16	0.19	0.20	0.27	0.17
b	SD	0.91	0.96	0.84	0.36	0.93	0.86	0.64	0.65	0.95	0.98
	Min	-2.91	-3.09	-1.36	-0.90	-3.22	-1.90	-0.99	-1.26	-1.88	-2.08
	Max	1.57	1.58	1.97	1.73	1.43	1.81	1.42	1.59	2.16	1.67
	Ave	0.20	0.20	0.22	0.22	0.23	0.20	0.20	0.21	0.22	0.21
c	SD	0.10	0.10	0.13	0.13	0.09	0.09	0.10	0.12	0.11	0.10
	Min	0.04	0.05	0.05	0.03	0.08	0.04	0.03	0.03	0.03	0.05
	Max	0.45	0.50	0.50	0.50	0.48	0.33	0.41	0.50	0.44	0.42

Table 6.10. Summary of Item Parameters for AS Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	1.14	1.16	1.14	1.16	1.14	1.10	1.17	1.16	1.13	1.11
a	SD	0.38	0.36	0.36	0.36	0.38	0.35	0.35	0.37	0.48	0.43
	Min	0.35	0.37	0.46	0.51	0.59	0.61	0.58	0.64	0.47	0.45
	Max	1.88	1.81	1.77	1.77	2.12	1.89	1.97	2.18	2.14	1.98
	Ave	0.49	0.52	0.69	0.64	0.70	0.70	0.68	0.65	0.56	0.60
b	SD	0.59	0.56	0.69	0.72	0.69	0.70	0.53	0.52	0.86	0.85
	Min	-0.71	-0.49	-1.09	-1.13	-0.85	-0.88	-0.18	-0.23	-1.50	-1.47
	Max	1.64	1.61	1.72	1.75	2.60	2.77	1.64	1.54	2.28	1.95
	Ave	0.23	0.24	0.22	0.22	0.20	0.19	0.18	0.18	0.25	0.24
c	SD	0.09	0.10	0.09	0.08	0.12	0.12	0.11	0.10	0.10	0.11
	Min	0.06	0.07	0.05	0.07	0.06	0.06	0.04	0.06	0.07	0.06
	Max	0.39	0.41	0.44	0.34	0.48	0.47	0.49	0.45	0.49	0.50

Table 6.11. Summary of Item Parameters for MC Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	0.81	0.89	0.98	0.93	0.91	0.92	0.94	0.92	0.92	0.91
a	SD	0.26	0.30	0.23	0.21	0.31	0.29	0.31	0.30	0.36	0.38
	Min	0.39	0.39	0.62	0.60	0.41	0.41	0.55	0.48	0.40	0.41
	Max	1.27	1.38	1.36	1.28	1.63	1.65	1.60	1.73	1.88	2.00
	Ave	0.33	0.37	0.44	0.40	0.19	0.23	0.48	0.55	0.34	0.39
b	SD	1.32	1.27	1.15	1.15	1.14	1.04	1.31	1.26	1.00	0.95
	Min	-2.09	-1.96	-1.65	-1.74	-2.56	-2.03	-2.91	-2.83	-1.29	-1.27
	Max	2.34	2.35	2.11	2.12	1.65	1.79	2.36	2.47	2.57	2.52
	Ave	0.22	0.23	0.22	0.22	0.23	0.22	0.25	0.25	0.23	0.24
c	SD	0.09	0.10	0.10	0.08	0.08	0.08	0.12	0.13	0.12	0.12
	Min	0.09	0.10	0.06	0.06	0.09	0.09	0.05	0.06	0.08	0.06
	Max	0.46	0.50	0.42	0.35	0.47	0.40	0.49	0.50	0.50	0.50

Table 6.12. Summary of Item Parameters for EI Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	0.81	0.81	0.91	0.85	0.84	0.84	0.81	0.82	1.04	1.00
a	SD	0.30	0.30	0.34	0.33	0.35	0.31	0.36	0.33	0.58	0.44
	Min	0.36	0.37	0.38	0.32	0.24	0.25	0.31	0.31	0.36	0.42
	Max	1.35	1.37	1.65	1.61	1.74	1.42	1.74	1.66	2.85	2.30
	Ave	0.46	0.47	0.88	0.86	0.30	0.41	0.23	0.23	0.61	0.71
b	SD	1.25	1.21	0.99	1.09	1.12	1.10	1.14	1.20	1.15	0.98
	Min	-1.42	-1.41	-0.56	-0.71	-1.62	-1.77	-1.91	-1.77	-2.47	-1.89
	Max	2.86	2.90	2.37	2.62	1.71	1.93	2.01	1.98	2.12	2.15
	Ave	0.21	0.22	0.23	0.22	1.22	0.24	0.22	0.22	0.29	0.30
c	SD	0.09	0.09	0.11	0.11	0.09	0.10	0.08	0.07	0.11	0.13
	Min	0.09	0.10	0.08	0.09	0.08	0.10	0.09	0.09	0.11	0.12
	Max	0.43	0.44	0.50	0.50	0.40	0.44	0.42	0.35	0.50	0.50

Table 6.13. Summary of Item Parameters for AO Forms 23–27

Parameter	Statistic	Form 23A	Form 23B	Form 24A	Form 24B	Form 25A	Form 25B	Form 26A	Form 26B	Form 27A	Form 27B
	Ave	1.10	1.05	1.09	1.07	1.20	1.17	1.11	1.10	1.06	1.06
a	SD	0.32	0.30	0.22	0.25	0.43	0.40	0.28	0.25	0.28	0.28
	Min	0.48	0.52	0.75	0.65	0.60	0.62	0.68	0.69	0.65	0.64
	Max	1.64	1.68	1.60	1.64	2.03	2.21	1.80	1.68	1.61	1.60
	Ave	-0.44	-0.53	-0.68	-0.70	-0.67	-0.70	-0.60	-0.57	-0.67	-0.61
b	SD	0.62	0.55	0.43	0.42	0.57	0.58	0.49	0.49	0.51	0.49
	Min	-1.50	-1.53	-1.47	-1.17	-1.57	-1.62	-1.62	-1.61	-1.69	-1.61
	Max	1.26	1.06	0.66	0.58	0.33	0.34	0.39	0.44	0.24	0.27
	Ave	0.23	0.21	0.18	0.17	0.22	0.21	0.19	0.20	0.19	0.20
c	SD	0.13	0.11	0.07	0.07	0.15	0.15	0.08	0.08	0.08	0.08
	Min	0.03	0.04	0.07	0.09	0.04	0.04	0.07	0.08	0.05	0.09
	Max	0.43	0.40	0.31	0.30	0.49	0.50	0.40	0.39	0.39	0.41

6.4. Test Information Functions

Figures 6.1–6.9 show the test information functions for Forms 23–27 for each P&P-ASVAB subtest. The test information functions are an upper bound to the amount of information that can be obtained by any method of scoring the test (Lord, 1980).

Figure 6.1. Test Information Functions for GS Forms 23–27.

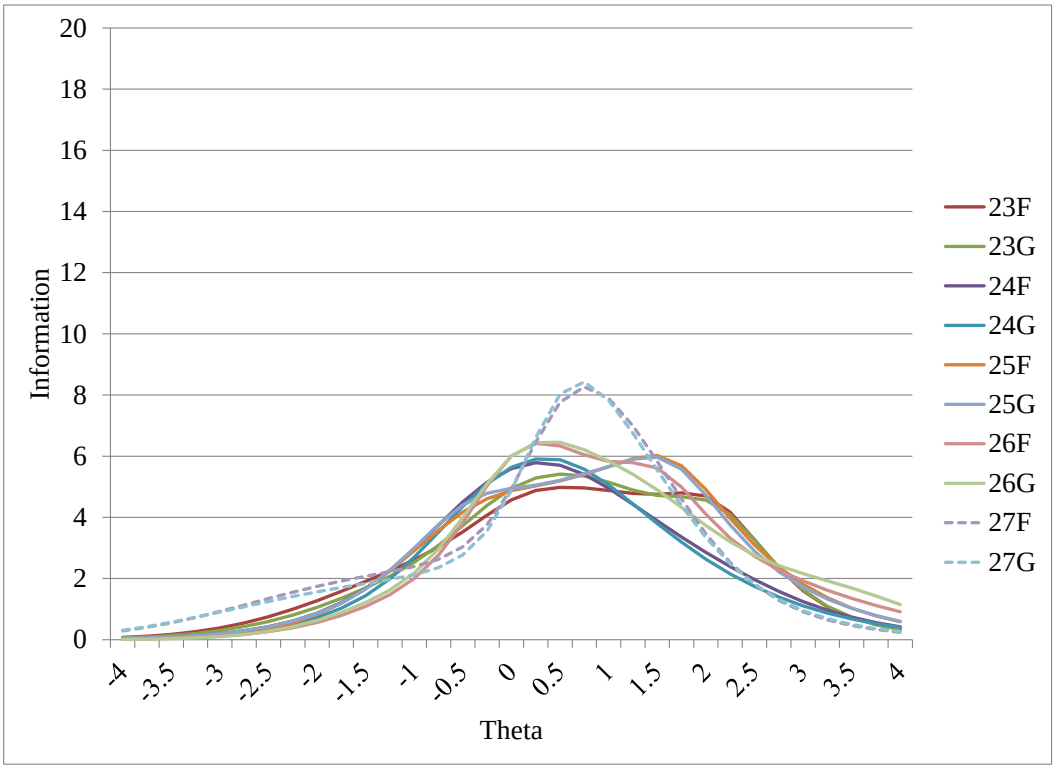


Figure 6.2. Test Information Functions for AR Forms 23–27.

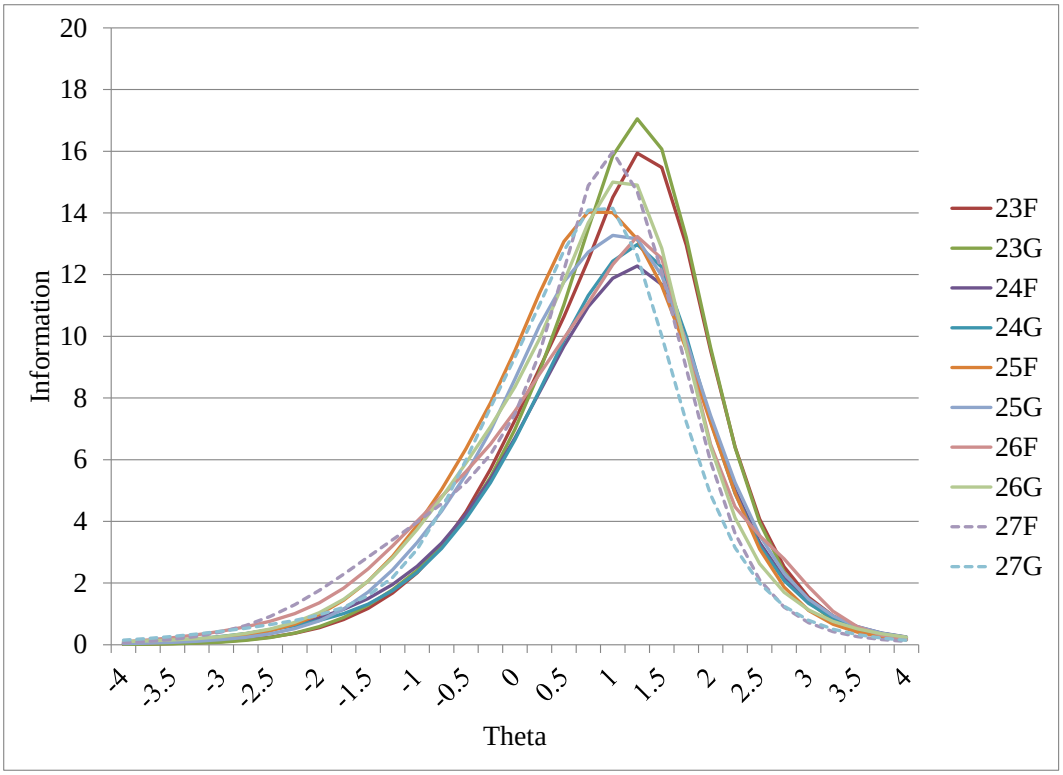


Figure 6.3. Test Information Functions for WK Forms 23–27.

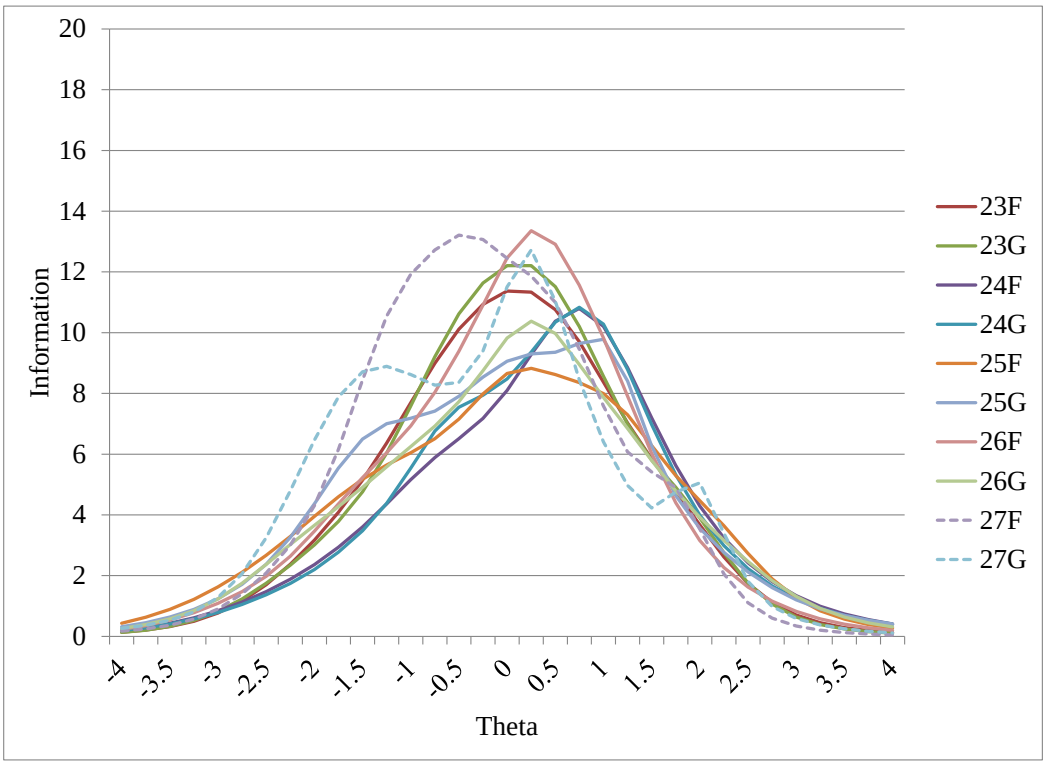


Figure 6.4. Test Information Functions for PC Forms 23–27.

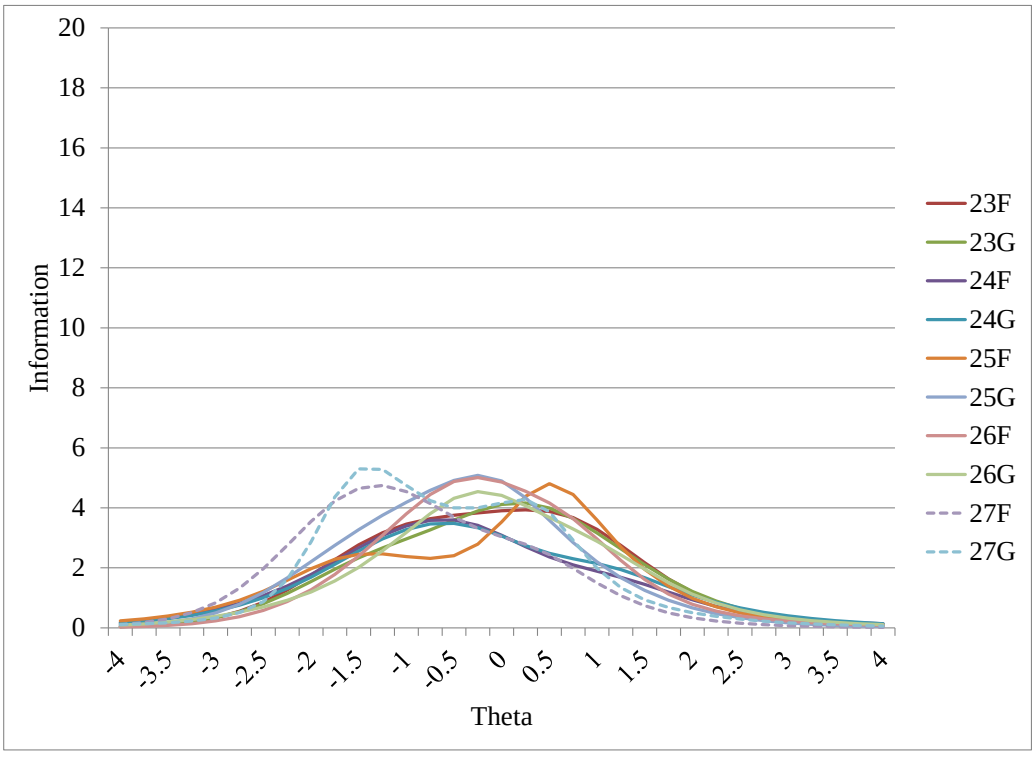


Figure 6.5. Test Information Functions for MK Forms 23–27.

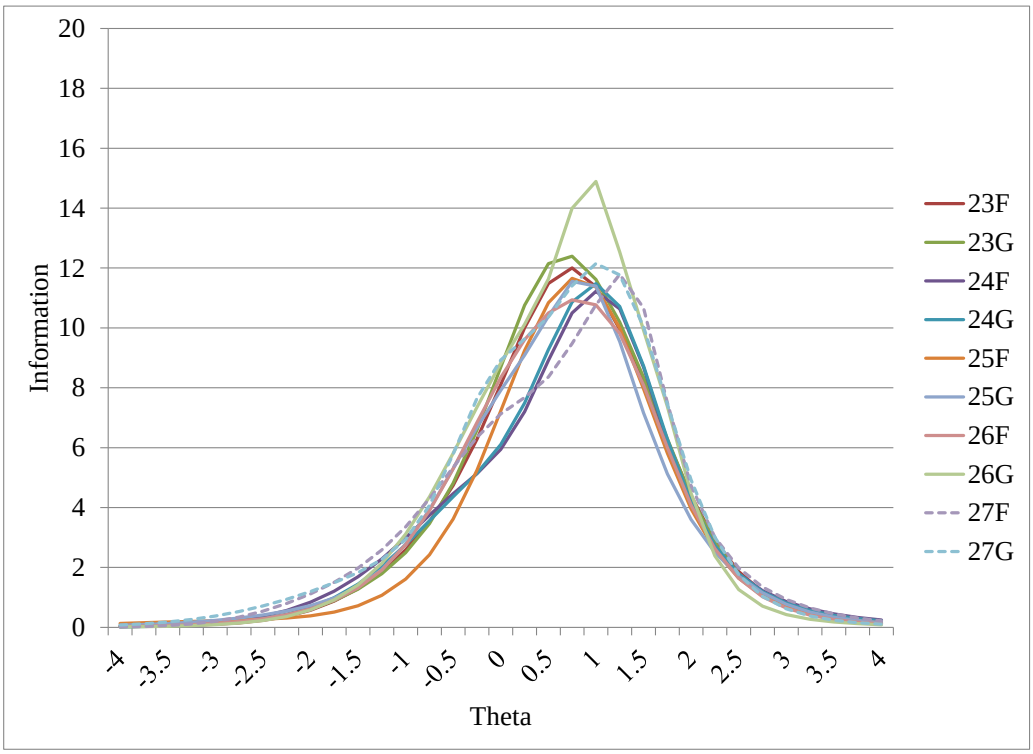


Figure 6.6. Test Information Functions for AS Forms 23–27.

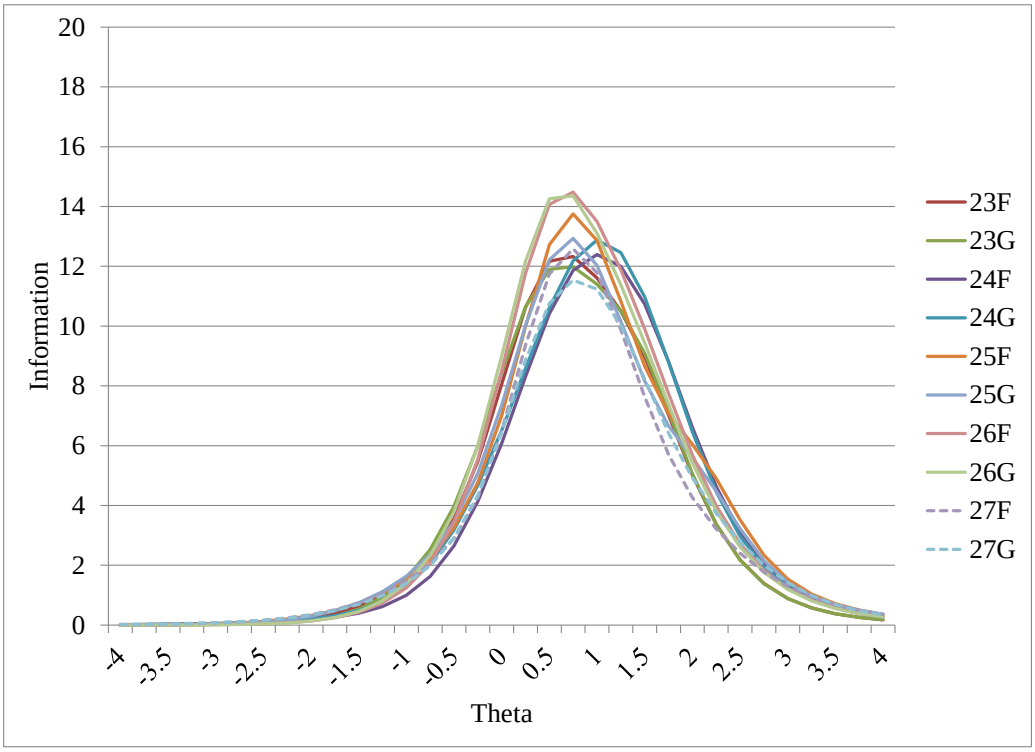


Figure 6.7. Test Information Functions for MC Forms 23–27.

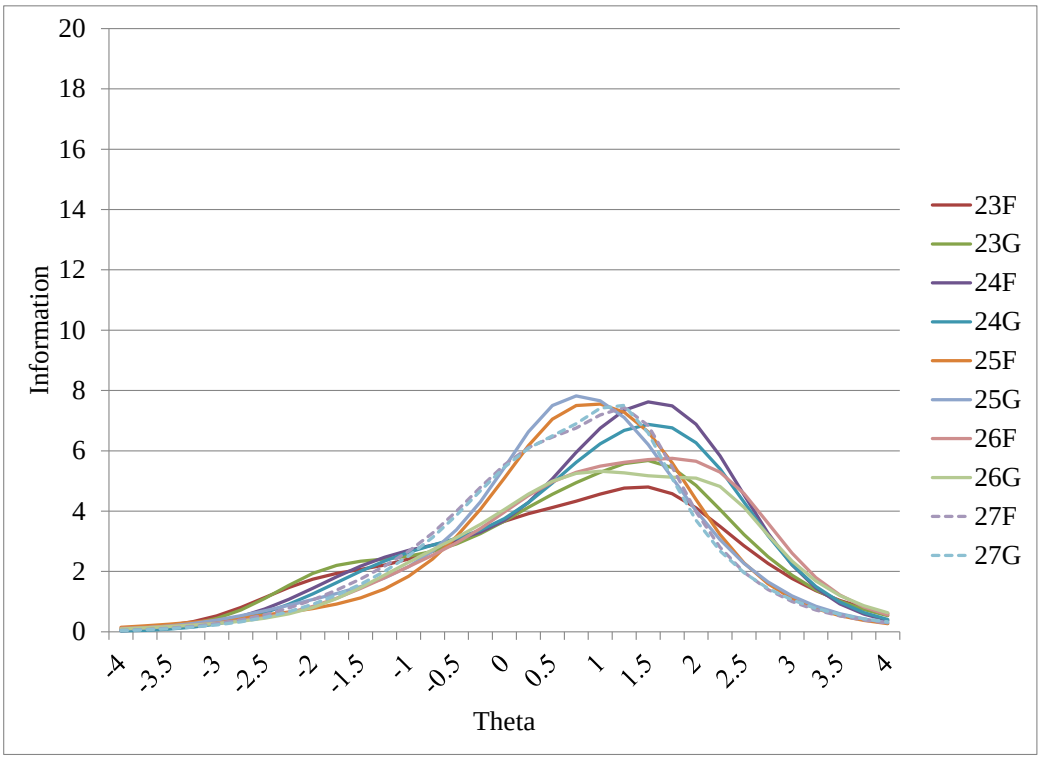


Figure 6.8. Test Information Functions for EI Forms 23–27.

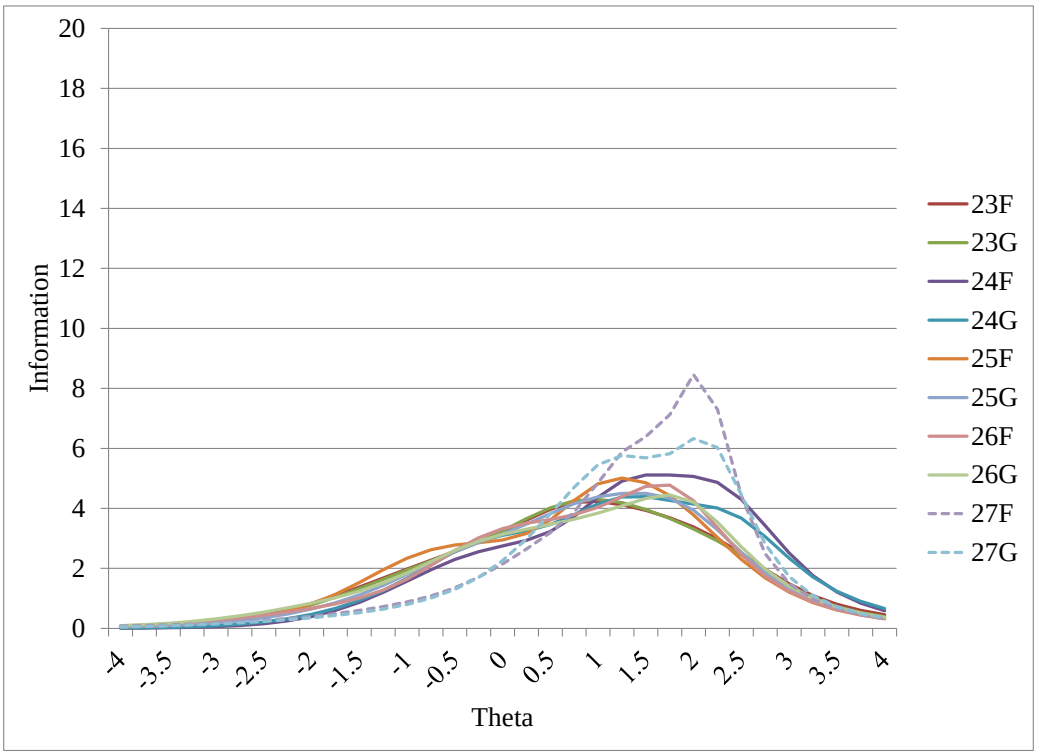
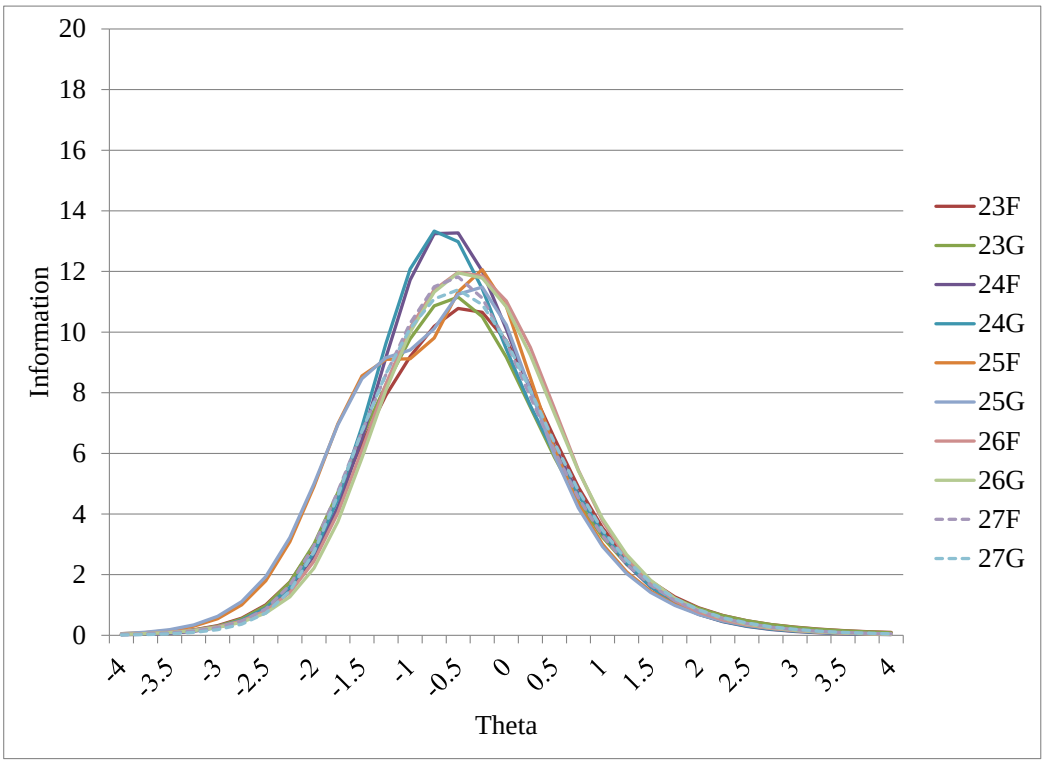


Figure 6.9. Test Information Functions for AO Forms 23–27.



6.5. Test-Retest Reliabilities

Test-retest reliabilities for Forms 23–27 were estimated by correlating IRT scores (i.e., BMEs) across two simulated P&P administrations for 10,000 examinees sampled from a N(0,1) distribution. Table 6.14 summarizes the test-retest reliability estimates for each form.

Table 6.14. Test-Retest Reliability Estimates

Test	23A	23B	24A	24B	25A	25B	26A	26B	27A	27B
GS	0.79	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.80	0.80
AR	0.86	0.86	0.84	0.84	0.88	0.87	0.87	0.88	0.87	0.87
WK	0.89	0.90	0.87	0.87	0.87	0.88	0.89	0.88	0.90	0.89
PC	0.77	0.75	0.71	0.72	0.74	0.78	0.78	0.76	0.75	0.77
MK	0.85	0.86	0.84	0.84	0.84	0.85	0.86	0.87	0.86	0.87
AS	0.84	0.85	0.83	0.83	0.83	0.85	0.85	0.86	0.83	0.83
MC	0.77	0.78	0.80	0.80	0.80	0.81	0.78	0.78	0.81	0.82
EI	0.74	0.74	0.72	0.74	0.75	0.75	0.74	0.75	0.70	0.69
AO	0.87	0.87	0.88	0.87	0.87	0.87	0.88	0.88	0.87	0.87

References

Alderton, D. L., Wolfe, J. H., & Larson, G. E. (1997). The ECAT battery. *Military Psychology, 9,* 5-37.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey CA: Brooks/Cole.

American Educational Research Association (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

DMDC (2006). *CAT-ASVAB Forms 1–2* (Technical Bulletin No. 1). Seaside, CA: Defense Manpower Data Center.

DMDC (2009). *CAT-ASVAB Forms 3–4* (Technical Bulletin No. 2). Seaside, CA: Defense Manpower Data Center.

DMDC (2008). *CAT–ASVAB Forms 5–9* (Technical Bulletin No. 3). Seaside, CA: Defense Manpower Data Center.

Harris, J. A., & Weger-Montano, G. (2000, November). *Results of an informal study of sensitivity reviews of test items*. Paper presented at the 42nd conference of the International Military Testing Association, Edinburgh, Scotland.

Held, J. D., & Wolfe, J. H. (1997). Validities of unit-weighted composites of the ASVAB and the ECAT battery. *Military Psychology, 9*, 77-84.

Hiatt, C. M., & Sims, W. H. (2003). *Norming tables for the student testing program* (*STP97*) (CAB D0009233.A2). Alexandria VA: Center for Naval Analyses.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory*: *Application to Psychological Measurement*. Homewood IL: Down Jones-Irwin.

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating: Methods and Practices*. New York: Springer.

Larson, G. E., & Alderton, D. L. (1992). *Reliabilities and practice effects for the ECAT battery*. In Proceedings of the 34th Annual Conference of the Military Testing Association. (Vol. 2, 33-38). San Diego, CA: Navy Personnel Research and Development Center.

Larson, G. E., & Alderton, D. L. (1997). Test-retest reliabilities of the ECAT batteries. *Military Psychology, 9,* 39-47.

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

MaCurdy, T., & Vytlacil, E. (2003). *Establishing new norms for the AFQT using data from PAY97.* Unpublished manuscript, Stanford University.

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

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software International, Inc.

Palmer, P., Hartke, D. D., Ree, M. I., Welsh, J. R., & Valentine, L. D., Jr. (1988). *Armed Services Vocational Aptitude Battery* (*ASVAB*)*: Alternate forms reliability* (*Forms 8. 9, 10, and 11*) (AFHRL-TP-87-48), Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sands, W. A., & Waters, B. K. (1997). Introduction to ASVAB and CAT. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (*p 6*). Washington, DC: American Psychological Association.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Segall, D. O. & Thomasson, G. L. (2001, October). *Scoring, Equating, and Scaling Analyses: P&P-ASVAB Forms 23 – 27*. Briefing presented to the Defense Advisory Committee on Military Personnel Testing, Monterey CA.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. (DMDC No. 2004-002), Seaside CA: Defense Manpower Data Center.

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

Thomasson, G. L., Bloxom, B., & Wise, L. (1994). *Initial operational test and evaluation of Forms 20, 21, and 22 of the Armed Services Vocational Aptitude Battery* (*ASVAB*) (DMDC No. 94-001). Monterey, CA: Defense Manpower Data Center.

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

U. S. Department of Defense (2005). *ASVAB Career Exploration Program Counselor Manual*. North Chicago, IL: U. S. Military Entrance Processing Command.

USMEPCOM (n.d.). Retrieved from http://www.mepcom.army.mil/hq.html.

Welsh, J. R. (2003). *Profiles of American Youth* (*1980 and 1997 Norming Sample*). Briefing presented at OSD norming workshop. Alexandria VA, October 30-31, 2003.

Welsh, J. R., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Aptitude Battery* (*ASVAB*)*: Integrative review of validity studies* (AFHRL-TR-90-22). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Wise, L. & Curran, L. (1995). Fact Sheet on ASVAB Norms - The 1996 Profile of American Youth Study. Unpublished manuscript, Defense Manpower Data Center.

Wolfe, J. H. (Ed.). (1997). Enhanced computer-administered test (ECAT) battery [Special Issue]. *Military Psychology*, *9(1)*.

Wolfe, J. H., Alderton, D. L., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (*p. 240*). Washington, DC: American Psychological Association.

Wolfe, J. H., Alderton, D. L., Larson, G. E., & Held, J. D (1995). *Incremental validity of Enhanced Computer Administered Testing* (*ECAT*) (Tech Note TN-96-6). San Diego, CA: Navy Personnel Research and Development Center.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning* (*DIF*)*: Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type* (*Ordinal*) *Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

History of the AFQT and the P&P-ASVAB

Table A.1. AFQT History 1950–Present[1]

Forms	Dates	Uses	Subtests	Number of Items	Admin Time
1 and 2	Jan. 1950 – Dec. 1952	Induction and enlistment	Verbal Arithmetic Reasoning Space Perception	30 30 30	45 min.
3 and 4	Jan. 1953 – Jul. 1956	Same	Verbal Arithmetic Reasoning Space Perception Tool Knowledge	25 25 25 25	50 min.
5 and 6	Aug. 1956 – Jun. 1960	Same	Same as Forms 3 and 4		
7 and 8	Jul. 1960 – May 1973	Same	Same as Forms 3 and 4		
	May 1973 – Jan. 1976		Each Service had its own AFQT		
	Jan. 1976	Joint-Service ASVAB implemented; AFQT obtained from ASVAB			
5	Jul. 1976 – Jun. 1984	STP[a] and joint-Service enlistment	Word Knowledge Arithmetic Reasoning Space Perception	30 20 20	10 20 12
	AFQT = Sum of raw scores converted to percentile scale				
6 and 7	Jan. 1976 – Sep. 1980	Joint-Service enlistment	Same as Form 5		
8, 9, and 10	Oct. 1980 – Sep. 1984	Joint-Service enlistment	Word Knowledge Paragraph Comprehension Arithmetic Reasoning Numerical Operations	35 15 30 50	11 13 36 3
	AFQT = Sum of raw scores converted to percentile scale				
11, 12, and 13	Oct. 1984 – Dec. 1988	Joint-Service enlistment	Same as Forms 8, 9, and 10		
14	Jan. 1989 – Jun. 1992	Joint-Service enlistment	Word Knowledge Paragraph Comprehension Arithmetic Reasoning Mathematics Knowledge	35 15 30 25	11 13 36 24
	AFQT = Sum of standard scores (Word Knowledge and Paragraph Comprehension scores doubled) converted to percentile scale				
15, 16 , and 17	Jan. 1989	Joint-Service enlistment	Same as Form 14		
18 and 19	Jul. 1992 – Jul. 2002	STP and Joint-Service enlistment	Same as Form 14		
20, 21, and 22	Oct. 1993 – Dec. 2001	Joint-Service enlistment	Same as Form 14		

Table A.1 (cont.) AFQT History 1950–Present

Forms	Dates	Uses	Subtests	Number of Items	Admin Time
23 and 24	Jul. 2002 - present	STP	Same as Form 14		
25 and 26	Jan. 2002 - present	Joint-Service enlistment	Same as Form 14		
Item response theory scoring applied to P&P forms 23 through 26 in 2002; new norms implemented in 2004					

Note. Table adapted from Maier (1993).
[a] STP = Student Testing Program.

Table A.2. P&P-ASVAB Forms History 1968–Present

Form	Dates	Uses	Subtests	Number of Items	Admin Time
1	Sep. 1968 – Dec. 1972	STP only	Word Knowledge Arithmetic Reasoning Tool Knowledge Space Perception Mechanical Comprehension Shop Information Automotive Information Electronics Information Coding Speed	25 25 25 25 25 25 25 25 100	10 25 10 15 15 10 10 10 7
2	Jan. 1973 – Jun. 1976	STP[a] and enlistment	Same as Form 1		
3	Sep. 1973 – Dec. 1975	Air Force enlistees	Same as Form 1		
	Jul. 1974 – Dec. 1975	Marine Corps enlistees	Same as Form 1		
4	Not used				
5	Jul. 1976 – Jun. 1984	STP and enlistment	Word Knowledge Arithmetic Reasoning Mathematics Knowledge Numerical Operations Attention to Detail General Science General Information Space Perception Mechanical Comprehension Shop Information Automotive Information Electronics Information	30 20 20 50 30 20 15 20 20 20 20 30	10 20 20 3 5 10 7 12 15 8 10 15
6 and 7	Jan. 1976 – Sep. 1980	Joint-Service enlistment	Same as Form 5 plus: Army Classification Inventory[b]	107	
8, 9, and 10	Oct. 1980 – Sep. 1984	Joint-Service enlistment	Word Knowledge Paragraph Comprehension Arithmetic Reasoning Mathematics Knowledge Auto & Shop Information Mechanical Comprehension Electronics Information Numerical Operations Coding Speed General Science	35 15 30 25 25 25 20 50 84 25	11 13 36 24 11 19 9 3 7 11
11, 12, and 13	Oct. 1984 – Dec. 1988	Joint-Service enlistment	Same as Forms 8, 9, and 10		
14	Jul. 1984 – Jun. 1992	STP and Joint-Service enlistment	Same as Forms 8, 9, and 10		

Table A.2 (cont.) P&P-ASVAB Forms History 1968–Present

Form	Dates	Uses	Subtests	Number of Items	Admin Time
15, 16, and 17	Jan. 1989 – Sep. 1993	Joint-Service enlistment	Same as Forms 8, 9, and 10		
18 and 19	Jul. 1992 – Jul. 2002	STP and Joint-Service enlistment	Same as Forms 8, 9, and 10		
20, 21, and 22	Oct. 1993 – Dec. 2001	Joint-Service enlistment	General Science Arithmetic Reasoning Word Knowledge Paragraph Comprehension Numerical Operations Coding Speed Auto & Shop Information Mathematics Knowledge Mechanical Comprehension Electronics Information	25 30 35 15 50 84 25 25 25 20	11 36 11 13 3 7 11 24 19 9
23 and 24	Jul. 2002 - present	STP	General Science Arithmetic Reasoning Word Knowledge Paragraph Comprehension Mathematics Knowledge Electronics Information Auto & Shop Information Mechanical Comprehension	25 30 35 15 25 20 25 25	11 36 11 13 24 9 11 19
25 and 26	Jan. 2002 - present	Joint-Service enlistment	General Science Arithmetic Reasoning Word Knowledge Paragraph Comprehension Mathematics Knowledge Electronics Information Auto & Shop Information Mechanical Comprehension Assembling Objects[c]	25 30 35 15 25 20 25 25 25	11 36 11 13 24 9 11 19 15
	Subtest scores based on IRT started Jan. 2002; new norms implemented 1 July 2004				

Note. Table adapted from Maier (1993).
[a] STP = Student Testing Program.
[b] Composed of Mechanical Interest, Electronics Interest, Combat Interest, and Attentiveness Interest.
[c] Assembling Objects is not administered in the STP.

Appendix B

Timeline of Major Events in Recent ASVAB History

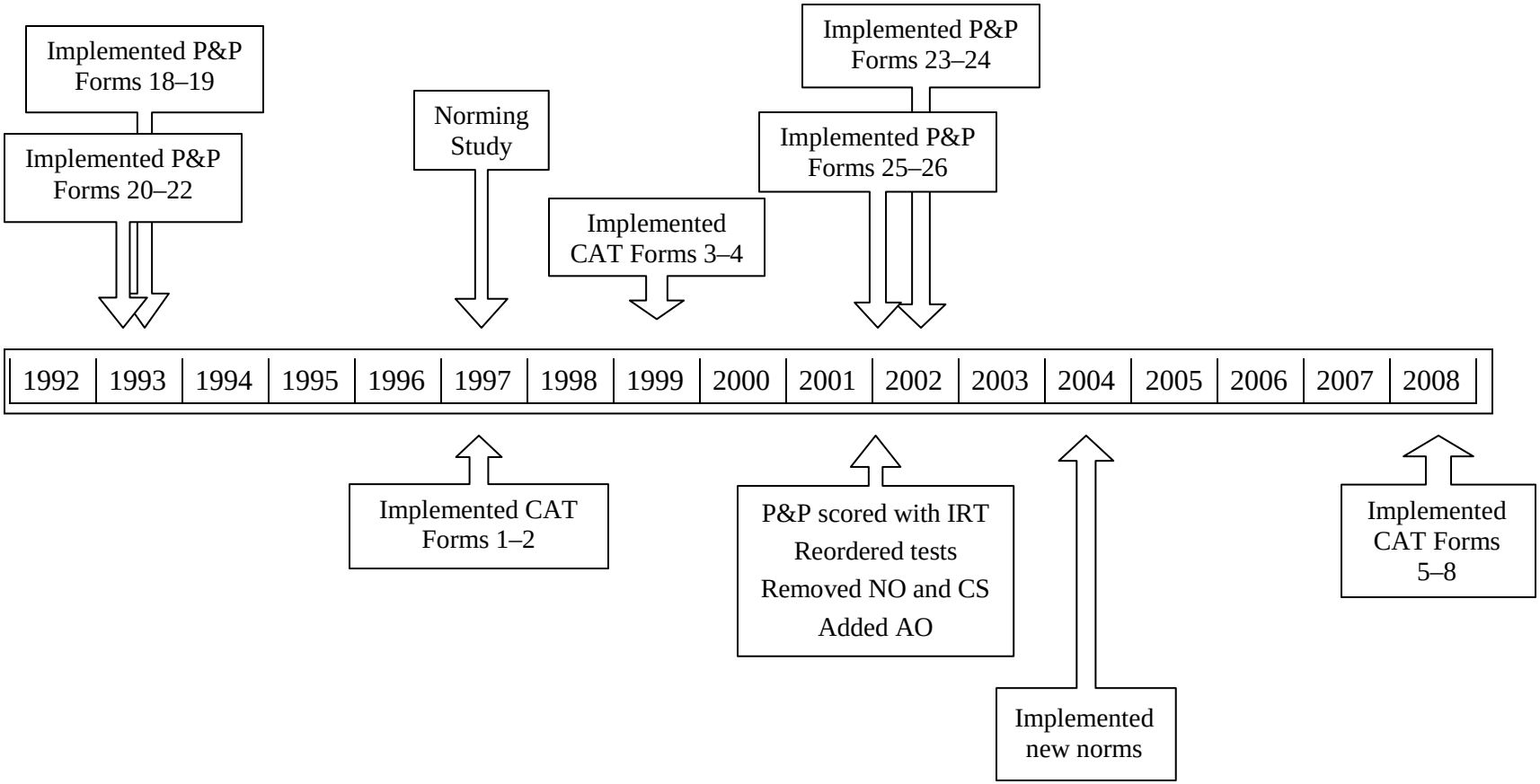


Figure B.1. Timeline of Major Events in Recent ASVAB History.

Appendix C

Service-Specific Composites

Table C.1. Service-Specific Composites

Service	Composite	Computational Formula
Army	General Technical (GT)	AR + VE
	Clerical (CL)	*
	Combat (CO)	*
	Electronics Repair (EL)	*
	Field Artillery (FA)	*
	General Maintenance (GM)	*
	Mechanical Maintenance (MM)	*
	Operators/Food (OF)	*
	Surveillance/Communication (SC)	*
	Skilled Technician (ST)	*
Navy	General Technician (GT)	VE + AR
	Electronics (EL)	GS + AR + MK + EI
	Basic Electricity and Electronics (BEE)	GS + AR + 2MK
	Engineering (ENG)	AS + MK
	Mechanical1 (MEC)	AR + AS + MC
	Mechanical2 (MEC2)	AR + MC + AO
	Nuclear (NUC)	VE + AR + MK + MC
	Operations (OPS)	VE + AR + MK + AO
	Hospitalman (HM)	VE + GS + MK
	Administrative (ADM)	VE + MK
Air Force (AF)	Mechanical (M)	AR + 2VE + MC + AS
	Administrative (A)	VE + MK
	General (G)	VE + AR
	Electronic (E)	AR + MK + EI + GS
Marine Corps (MC)	Mechanical (MM)	AR + MC + AS + EI
	Clerical (CL)	VE + MK
	General Technician (GT)	VE + AR + MC
	Electrical (EL)	AR + MK + EI + GS
All	AFQT	2(VE) + AR + MK

* Computed as a non-integer weighted linear combination of all ASVAB subtests GS, AR, WK, PC, MK, EI, AS, MC, and AO.

Appendix D

Military Installations used for Studies

Table D.1. Recruit Training Centers Used in Item Tryout and OPCAL Studies

			Item Tryout		
Site	State	Service	AFQT	Non-AFQT	OPCAL
Ft. Benning	GA	Army	√	√	√
Great Lakes	IL	Navy		√	√
Ft. Knox	KY	Army	√		√
Ft. McClellan	AL	Army			√
Lackland	TX	Air Force	√	√	√
Ft. Leonard Wood	MO	Army	√	√	
Orlando	FL	Navy	√		
Parris Island	SC	Marine Corps	√	√	√
San Diego	CA	Navy	√	√	
San Diego	CA	Marine Corps	√	√	
Ft. Sill	OK	Army	√	√	√

Table D.2. MEPS Participating in Anchoring Study

MEPS
Boston, MA
New York, NY
Pittsburgh, PA
Tampa, FL
Atlanta, GA
Amarillo, TX
Dallas, TX
Little Rock, AR
Des Moines, IA
San Diego, CA
Sacramento, CA

Appendix E

AFQT Qualification Rates by Scoring Procedure

Table E.1. Form 28C Qualification Rates by Scoring Procedure for Total Group (N=10735)

Cutscore	IRT	NR	Difference	AGR
31	68.8	68.5	0.3 ( )	96
50	40.3	40.1	0.1 ( )	97
65	22.8	22.6	0.3 ( )	97
93	3.3	2.1	1.2 (+)	99

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate
- AGR = percentage of group members that received the same classifications on both IRT and NR scoring approaches.

Table E.2. Form 28C Qualification Rates by Scoring Procedure for Females (N=2690)

Cutscore	IRT	NR	Difference	AGR
31	68.1	67.1	1.0 ( )	95
50	37.3	37.0	0.3 ( )	96
65	20.0	19.1	0.9 (+)	98
93	2.2	1.4	0.7 (+)	99

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate
- AGR = percentage of group members that received the same classifications on both IRT and NR scoring approaches.

Table E.3. Form 28C Qualification Rates by Scoring Procedure for Blacks (N=2877)

Cutscore	IRT	NR	Difference	AGR
31	53.6	52.9	0.7 ( )	95
50	21.7	22.0	-0.3 ( )	97
65	8.7	8.8	-0.1 ( )	99
93	0.3	0.2	0.1 ( )	100

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate
- AGR = percentage of group members that received the same classifications on both IRT and NR scoring approaches.

Table E.4. Form 28C Qualification Rates by Scoring Procedure for Hispanics (N=1293)

Cutscore	IRT	NR	Difference	AGR
31	50.0	50.3	-0.3 ( )	95
50	20.6	22.3	-1.6 (–)	97
65	8.9	9.3	-0.4 ( )	98
93	0.9	0.7	0.2 ( )	100

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate
- AGR = percentage of group members that received the same classifications on both IRT and NR scoring approaches.

Table E.5. New Form Qualification Rates by Scoring Procedure for Total Group (AFQT Cutscore = 31)

Cutscore	Phase	N	Form	IRT	NR	Difference
31	1	10752	23A	69.7	69.0	0.7 (+)
31	1	10728	23B	68.9	68.3	0.7 (+)
31	1	10737	25B	69.1	68.2	0.9 (+)
31	1	10717	26A	70.0	69.0	1.0 (+)
31	1	10754	26B	69.3	67.9	1.4 (+)
31	1	10735	28C	68.8	68.5	0.3 ( )
31	2	11932	24A	70.7	69.5	1.2 (+)
31	2	11900	24B	70.3	69.1	1.2 (+)
31	2	11835	25A	69.8	68.1	1.7 (+)
31	2	11884	27A	69.3	69.3	0.0 ( )
31	2	11797	27B	70.0	69.5	0.4 ( )
31	2	11734	28C	69.3	68.9	0.5 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.6. New Form Qualification Rates by Scoring Procedure for Total Group (AFQT Cutscore = 50)

Cutscore	Phase	N	Form	IRT	NR	Difference
50	1	10752	23A	39.4	38.8	0.5 (+)
50	1	10728	23B	39.2	38.9	0.2 ( )
50	1	10737	25B	40.1	39.3	0.7 (+)
50	1	10717	26A	39.5	39.2	0.3 ( )
50	1	10754	26B	40.5	39.3	1.2 (+)
50	1	10735	28C	40.3	40.1	0.1 ( )
50	2	11932	24A	38.8	38.3	0.5 (+)
50	2	11900	24B	38.8	38.3	0.6 (+)
50	2	11835	25A	40.0	38.9	1.1 (+)
50	2	11884	27A	39.2	39.2	0.1 ( )
50	2	11797	27B	38.5	38.3	0.2 ( )
50	2	11734	28C	39.6	39.5	0.1 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.7. New Form Qualification Rates by Scoring Procedure for Total Group (AFQT Cutscore = 65)

Cutscore	Phase	N	Form	IRT	NR	Difference
65	1	10752	23A	23.4	22.8	0.7 (+)
65	1	10728	23B	23.3	22.7	0.6 (+)
65	1	10737	25B	22.9	22.9	0.0 ( )
65	1	10717	26A	22.4	21.9	0.4 (+)
65	1	10754	26B	23.3	22.3	1.0 (+)
65	1	10735	28C	22.8	22.6	0.3 ( )
65	2	11932	24A	21.7	21.4	0.3 ( )
65	2	11900	24B	22.1	21.7	0.4 (+)
65	2	11835	25A	22.7	21.8	0.8 (+)
65	2	11884	27A	22.3	22.0	0.3 ( )
65	2	11797	27B	21.8	21.3	0.5 (+)
65	2	11734	28C	22.7	22.0	0.6 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.8. New Form Qualification Rates by Scoring Procedure for Total Group (AFQT Cutscore = 93)

Cutscore	Phase	N	Form	IRT	NR	Difference
93	1	10752	23A	4.3	2.5	1.8 (+)
93	1	10728	23B	4.4	2.4	2.1 (+)
93	1	10737	25B	3.7	2.5	1.2 (+)
93	1	10717	26A	3.7	2.0	1.7 (+)
93	1	10754	26B	3.6	2.3	1.3 (+)
93	1	10735	28C	3.3	2.1	1.2 (+)
93	2	11932	24A	4.0	2.4	1.6 (+)
93	2	11900	24B	4.1	2.5	1.6 (+)
93	2	11835	25A	4.1	2.2	2.0 (+)
93	2	11884	27A	3.1	2.1	1.1 (+)
93	2	11797	27B	3.4	2.3	1.1 (+)
93	2	11734	28C	3.3	2.1	1.2 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.9. New Form Qualification Rates by Scoring Procedure for Females (AFQT Cutscore = 31)

Cutscore	Phase	N	Form	IRT	NR	Difference
31	1	2670	23A	71.2	70.3	0.9 ( )
31	1	2779	23B	69.5	68.7	0.8 ( )
31	1	2821	25B	70.3	68.8	1.5 (+)
31	1	2770	26A	69.0	68.4	0.6 ( )
31	1	2702	26B	68.6	67.1	1.5 (+)
31	1	2690	28C	68.1	67.1	1.0 ( )
31	2	3219	24A	71.6	69.6	2.0 (+)
31	2	3225	24B	69.4	67.5	1.9 (+)
31	2	3204	25A	68.0	66.1	1.8 (+)
31	2	3273	27A	68.4	68.1	0.3 ( )
31	2	3276	27B	70.9	69.8	1.0 ( )
31	2	3208	28C	66.8	66.1	0.8 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.10. New Form Qualification Rates by Scoring Procedure for Females (AFQT Cutscore = 50)

Cutscore	Phase	N	Form	IRT	NR	Difference
50	1	2670	23A	37.5	36.9	0.6 ( )
50	1	2779	23B	38.1	37.5	0.6 ( )
50	1	2821	25B	37.4	36.0	1.4 (+)
50	1	2770	26A	35.4	35.4	0.0 ( )
50	1	2702	26B	37.1	36.2	0.9 (+)
50	1	2690	28C	37.3	37.0	0.3 ( )
50	2	3219	24A	36.8	35.5	1.2 (+)
50	2	3225	24B	36.4	35.3	1.1 (+)
50	2	3204	25A	35.4	34.1	1.3 (+)
50	2	3273	27A	35.0	35.0	0.1 ( )
50	2	3276	27B	35.9	35.7	0.2 ( )
50	2	3208	28C	36.3	36.0	0.3 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.11. New Form Qualification Rates by Scoring Procedure for Females (AFQT Cutscore = 65)

Cutscore	Phase	N	Form	IRT	NR	Difference
65	1	2670	23A	20.4	19.5	0.9 (+)
65	1	2779	23B	20.8	19.9	0.9 (+)
65	1	2821	25B	20.1	19.5	0.6 ( )
65	1	2770	26A	18.2	17.8	0.3 ( )
65	1	2702	26B	19.1	17.7	1.4 (+)
65	1	2690	28C	20.0	19.1	0.9 (+)
65	2	3219	24A	19.1	18.5	0.6 ( )
65	2	3225	24B	19.3	18.7	0.5 ( )
65	2	3204	25A	18.9	18.0	0.9 (+)
65	2	3273	27A	17.2	16.3	0.9 (+)
65	2	3276	27B	18.7	18.4	0.3 ( )
65	2	3208	28C	19.0	17.9	1.1 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.12. New Form Qualification Rates by Scoring Procedure for Females (AFQT Cutscore = 93)

Cutscore	Phase	N	Form	IRT	NR	Difference
93	1	2670	23A	3.3	1.8	1.5 (+)
93	1	2779	23B	3.1	1.6	1.4 (+)
93	1	2821	25B	2.4	1.6	0.9 (+)
93	1	2770	26A	2.8	1.5	1.3 (+)
93	1	2702	26B	2.2	1.3	0.9 (+)
93	1	2690	28C	2.2	1.4	0.7 (+)
93	2	3219	24A	3.1	1.7	1.4 (+)
93	2	3225	24B	2.8	1.6	1.2 (+)
93	2	3204	25A	2.6	1.5	1.1 (+)
93	2	3273	27A	1.9	1.3	0.6 (+)
93	2	3276	27B	2.3	1.5	0.8 (+)
93	2	3208	28C	1.6	1.1	0.5 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.13. New Form Qualification Rates by Scoring Procedure for Blacks (AFQT Cutscore = 31)

Cutscore	Phase	N	Form	IRT	NR	Difference
31	1	2871	23A	58.5	56.7	1.8 (+)
31	1	2808	23B	56.8	54.8	2.1 (+)
31	1	2977	25B	58.0	55.9	2.1 (+)
31	1	2879	26A	58.5	56.3	2.2 (+)
31	1	2843	26B	56.0	53.2	2.8 (+)
31	1	2877	28C	53.6	52.9	0.7 ( )
31	2	3492	24A	58.1	55.5	2.6 (+)
31	2	3506	24B	57.9	55.3	2.6 (+)
31	2	3562	25A	56.0	53.4	2.6 (+)
31	2	3534	27A	56.5	55.8	0.7 ( )
31	2	3531	27B	57.8	56.2	1.6 (+)
31	2	3361	28C	54.0	52.9	1.1 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.14. New Form Qualification Rates by Scoring Procedure for Blacks (AFQT Cutscore = 50)

Cutscore	Phase	N	Form	IRT	NR	Difference
50	1	2871	23A	22.9	22.2	0.7 ( )
50	1	2808	23B	22.9	22.6	0.4 ( )
50	1	2977	25B	23.4	22.2	1.2 (+)
50	1	2879	26A	22.2	22.0	0.3 ( )
50	1	2843	26B	21.9	21.1	0.7 ( )
50	1	2877	28C	21.7	22.0	-0.3 ( )
50	2	3492	24A	21.0	20.4	0.6 ( )
50	2	3506	24B	21.3	20.5	0.8 ( )
50	2	3562	25A	21.4	21.0	0.4 ( )
50	2	3534	27A	22.3	22.0	0.3 ( )
50	2	3531	27B	21.5	21.3	0.2 ( )
50	2	3361	28C	20.1	20.3	-0.2 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.15. New Form Qualification Rates by Scoring Procedure for Blacks (AFQT Cutscore = 65)

Cutscore	Phase	N	Form	IRT	NR	Difference
65	1	2871	23A	9.7	9.1	0.7 (+)
65	1	2808	23B	9.6	9.3	0.2 ( )
65	1	2977	25B	9.9	9.6	0.3 ( )
65	1	2879	26A	8.7	8.9	-0.2 ( )
65	1	2843	26B	9.2	8.6	0.6 (+)
65	1	2877	28C	8.7	8.8	-0.1 ( )
65	2	3492	24A	7.4	7.0	0.4 ( )
65	2	3506	24B	8.2	8.1	0.1 ( )
65	2	3562	25A	8.3	7.7	0.5 ( )
65	2	3534	27A	8.9	8.7	0.2 ( )
65	2	3531	27B	9.0	8.5	0.5 ( )
65	2	3361	28C	8.0	8.3	-0.3 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.16. New Form Qualification Rates by Scoring Procedure for Blacks (AFQT Cutscore = 93)

Cutscore	Phase	N	Form	IRT	NR	Difference
93	1	2871	23A	0.8	0.3	0.5 (+)
93	1	2808	23B	0.8	0.3	0.5 (+)
93	1	2977	25B	0.7	0.4	0.3 ( )
93	1	2879	26A	0.8	0.3	0.4 (+)
93	1	2843	26B	0.8	0.5	0.2 (+)
93	1	2877	28C	0.3	0.2	0.1 ( )
93	2	3492	24A	0.3	0.1	0.2 ( )
93	2	3506	24B	0.4	0.4	0.0 ( )
93	2	3562	25A	0.5	0.3	0.3 (+)
93	2	3534	27A	0.5	0.2	0.2 (+)
93	2	3531	27B	0.6	0.3	0.3 (+)
93	2	3361	28C	0.5	0.3	0.2 (+)

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.17. New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT Cutscore = 31)

Cutscore	Phase	N	Form	IRT	NR	Difference
31	1	1316	23A	51.5	50.9	0.6 ( )
31	1	1288	23B	46.7	49.1	-2.4 (–)
31	1	1302	25B	47.8	48.0	-0.2 ( )
31	1	1352	26A	47.3	47.6	-0.2 ( )
31	1	1319	26B	49.9	48.6	1.3 ( )
31	1	1293	28C	50.0	50.3	-0.3 ( )
31	2	1447	24A	54.2	54.2	0.0 ( )
31	2	1416	24B	53.0	53.5	-0.6 ( )
31	2	1430	25A	52.1	51.6	0.5 ( )
31	2	1429	27A	51.4	53.3	-2.0 (–)
31	2	1425	27B	50.0	52.6	-2.5 (–)
31	2	1454	28C	53.9	54.3	-0.3 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.18. New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT Cutscore = 50)

Cutscore	Phase	N	Form	IRT	NR	Difference
50	1	1316	23A	21.4	21.9	-0.5 ( )
50	1	1288	23B	18.3	18.6	-0.3 ( )
50	1	1302	25B	20.2	20.7	-0.5 ( )
50	1	1352	26A	17.7	18.3	-0.6 ( )
50	1	1319	26B	23.4	23.0	0.4 ( )
50	1	1293	28C	20.6	22.3	-1.6 (–)
50	2	1447	24A	23.5	23.9	-0.4 ( )
50	2	1416	24B	22.9	22.4	0.5 ( )
50	2	1430	25A	22.4	22.4	0.0 ( )
50	2	1429	27A	21.0	22.0	-1.0 ( )
50	2	1425	27B	20.5	20.8	-0.3 ( )
50	2	1454	28C	24.1	25.4	-1.3 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.19. New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT Cutscore = 65)

Cutscore	Phase	N	Form	IRT	NR	Difference
65	1	1316	23A	10.3	10.0	0.3 ( )
65	1	1288	23B	8.2	7.9	0.2 ( )
65	1	1302	25B	8.7	9.1	-0.5 ( )
65	1	1352	26A	7.5	7.9	-0.4 ( )
65	1	1319	26B	10.8	10.2	0.5 ( )
65	1	1293	28C	8.9	9.3	-0.4 ( )
65	2	1447	24A	9.1	9.1	0.1 ( )
65	2	1416	24B	9.7	9.9	-0.1 ( )
65	2	1430	25A	10.2	9.5	0.7 ( )
65	2	1429	27A	10.1	9.5	0.6 ( )
65	2	1425	27B	9.1	8.4	0.7 ( )
65	2	1454	28C	10.7	10.9	-0.2 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Table E.20. New Form Qualification Rates by Scoring Procedure for Hispanics (AFQT Cutscore = 93)

Cutscore	Phase	N	Form	IRT	NR	Difference
93	1	1316	23A	0.5	0.2	0.4 ( )
93	1	1288	23B	0.7	0.4	0.3 ( )
93	1	1302	25B	0.5	0.2	0.2 ( )
93	1	1352	26A	0.4	0.3	0.1 ( )
93	1	1319	26B	0.8	0.6	0.2 ( )
93	1	1293	28C	0.9	0.7	0.2 ( )
93	2	1447	24A	1.0	0.6	0.3 ( )
93	2	1416	24B	1.3	0.8	0.5 (+)
93	2	1430	25A	0.9	0.3	0.6 (+)
93	2	1429	27A	1.0	0.5	0.6 (+)
93	2	1425	27B	0.6	0.2	0.4 ( )
93	2	1454	28C	0.8	0.5	0.3 ( )

Notes:

- Test used is the McNemar test for nonindependent proportions (.01 level)
- ( ) = No significant difference between IRT and NR qualification rates
- (+) = Significant difference, IRT qualification rate > NR qualification rate
- (–) = Significant difference, NR qualification rate > IRT qualification rate

Appendix F

Subtest Theta Score (BME) Correlations

Table F.1. Correlations Among Subtest Theta Score Estimates: Form 23A (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.58	1.00							
WK	0.72	0.60	1.00						
PC	0.64	0.62	0.74	1.00					
AS	0.47	0.34	0.44	0.35	1.00				
MK	0.56	0.74	0.52	0.56	0.19	1.00			
MC	0.65	0.55	0.63	0.57	0.64	0.47	1.00		
EI	0.69	0.55	0.68	0.60	0.61	0.47	0.69	1.00	
AO	0.44	0.51	0.38	0.42	0.33	0.49	0.53	0.43	1.00

Table F.2. Correlations Among Subtest Theta Score Estimates: Form 23B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.58	1.00							
WK	0.73	0.61	1.00						
PC	0.65	0.62	0.75	1.00					
AS	0.48	0.35	0.45	0.35	1.00				
MK	0.57	0.73	0.52	0.56	0.21	1.00			
MC	0.66	0.56	0.64	0.58	0.66	0.48	1.00		
EI	0.69	0.54	0.69	0.60	0.61	0.47	0.70	1.00	
AO	0.44	0.50	0.40	0.44	0.32	0.51	0.54	0.43	1.00

Table F.3. Correlations Among Subtest Theta Score Estimates: Form 24A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.000								
AR	0.596	1.000							
WK	0.728	0.592	1.000						
PC	0.614	0.554	0.695	1.000					
AS	0.430	0.389	0.460	0.312	1.000				
MK	0.568	0.700	0.480	0.497	0.179	1.000			
MC	0.647	0.584	0.618	0.515	0.629	0.455	1.000		
EI	0.655	0.549	0.628	0.495	0.620	0.425	0.691	1.000	
AO	0.462	0.520	0.393	0.397	0.292	0.478	0.547	0.437	1.000

Table F.4. Correlations Among Subtest Theta Score Estimates: Form 24B (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.62	1.00							
WK	0.73	0.60	1.00						
PC	0.61	0.56	0.69	1.00					
AS	0.45	0.40	0.47	0.32	1.00				
MK	0.58	0.70	0.49	0.50	0.20	1.00			
MC	0.66	0.60	0.63	0.51	0.65	0.46	1.00		
EI	0.66	0.56	0.64	0.51	0.64	0.43	0.70	1.00	
AO	0.48	0.52	0.41	0.40	0.31	0.48	0.55	0.44	1.00

Table F.5. Correlations Among Subtest Theta Score Estimates: Form 25A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.62	1.00							
WK	0.76	0.62	1.00						
PC	0.65	0.60	0.72	1.00					
AS	0.56	0.44	0.51	0.39	1.00				
MK	0.55	0.73	0.51	0.51	0.25	1.00			
MC	0.71	0.64	0.66	0.58	0.69	0.50	1.00		
EI	0.71	0.59	0.68	0.58	0.65	0.48	0.72	1.00	
AO	0.43	0.51	0.37	0.38	0.34	0.47	0.52	0.42	1.00

Table F.6. Correlations Among Subtest Theta Score Estimates: Form 25B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.63	1.00							
WK	0.76	0.61	1.00						
PC	0.62	0.60	0.70	1.00					
AS	0.57	0.42	0.49	0.31	1.00				
MK	0.55	0.76	0.50	0.56	0.23	1.00			
MC	0.72	0.62	0.65	0.53	0.70	0.50	1.00		
EI	0.72	0.59	0.67	0.54	0.66	0.48	0.72	1.00	
AO	0.45	0.51	0.39	0.42	0.35	0.50	0.53	0.44	1.00

Table F.7. Correlations Among Subtest Theta Score Estimates: Form 26A (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.62	1.00							
WK	0.72	0.57	1.00						
PC	0.64	0.62	0.71	1.00					
AS	0.55	0.39	0.52	0.35	1.00				
MK	0.53	0.73	0.43	0.53	0.20	1.00			
MC	0.67	0.63	0.58	0.51	0.61	0.50	1.00		
EI	0.68	0.57	0.61	0.52	0.67	0.44	0.67	1.00	
AO	0.48	0.54	0.37	0.42	0.32	0.50	0.57	0.45	1.00

Table F.8. Correlations Among Subtest Theta Score Estimates: Form 26B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.64	1.00							
WK	0.73	0.62	1.00						
PC	0.61	0.62	0.71	1.00					
AS	0.55	0.43	0.47	0.37	1.00				
MK	0.51	0.71	0.46	0.50	0.16	1.00			
MC	0.65	0.64	0.57	0.52	0.60	0.48	1.00		
EI	0.68	0.60	0.61	0.54	0.65	0.42	0.67	1.00	
AO	0.48	0.56	0.42	0.43	0.34	0.49	0.60	0.47	1.00

Table F.9. Correlations Among Subtest Theta Score Estimates: Form 27A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.61	1.00							
WK	0.76	0.61	1.00						
PC	0.63	0.57	0.73	1.00					
AS	0.58	0.42	0.49	0.37	1.00				
MK	0.52	0.71	0.50	0.49	0.22	1.00			
MC	0.66	0.63	0.58	0.47	0.62	0.50	1.00		
EI	0.62	0.47	0.53	0.43	0.66	0.34	0.59	1.00	
AO	0.46	0.53	0.41	0.38	0.36	0.51	0.58	0.37	1.00

Table F.10. Correlations Among Subtest Theta Score Estimates: Form 27B (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.60	1.00							
WK	0.73	0.58	1.00						
PC	0.62	0.55	0.72	1.00					
AS	0.57	0.42	0.46	0.36	1.00				
MK	0.47	0.66	0.43	0.47	0.15	1.00			
MC	0.66	0.64	0.54	0.46	0.62	0.44	1.00		
EI	0.63	0.47	0.51	0.43	0.66	0.29	0.61	1.00	
AO	0.43	0.53	0.37	0.35	0.33	0.46	0.57	0.37	1.00

Appendix G

Subtest Number Right Score Correlations

Table G.1. Correlations Among Subtest Number Right Scores: Form 23A (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.57	1.00							
WK	0.70	0.59	1.00						
PC	0.63	0.59	0.73	1.00					
AS	0.44	0.34	0.43	0.34	1.00				
MK	0.57	0.74	0.51	0.54	0.21	1.00			
MC	0.61	0.53	0.58	0.53	0.62	0.47	1.00		
EI	0.64	0.52	0.64	0.56	0.60	0.45	0.64	1.00	
AO	0.42	0.48	0.38	0.41	0.32	0.49	0.51	0.40	1.00

Table G.2. Correlations Among Subtest Number Right Scores: Form 23B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.57	1.00							
WK	0.71	0.59	1.00						
PC	0.63	0.60	0.73	1.00					
AS	0.46	0.35	0.45	0.35	1.00				
MK	0.57	0.74	0.51	0.54	0.23	1.00			
MC	0.62	0.53	0.59	0.53	0.63	0.47	1.00		
EI	0.64	0.51	0.64	0.56	0.60	0.45	0.65	1.00	
AO	0.42	0.48	0.39	0.42	0.32	0.49	0.51	0.41	1.00

Table G.3. Correlations Among Subtest Number Right Scores: Form 24A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.58	1.00							
WK	0.70	0.58	1.00						
PC	0.59	0.51	0.67	1.00					
AS	0.43	0.39	0.45	0.31	1.00				
MK	0.56	0.69	0.48	0.47	0.21	1.00			
MC	0.62	0.56	0.59	0.48	0.64	0.45	1.00		
EI	0.60	0.51	0.57	0.45	0.61	0.41	0.65	1.00	
AO	0.44	0.48	0.37	0.38	0.29	0.46	0.50	0.39	1.00

Note: Subtest scores from Phase 2 were not transformed to the Phase 1 scale.

Table G.4. Correlations Among Subtest Number Right Scores: Form 24B (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.60	1.00							
WK	0.71	0.59	1.00						
PC	0.59	0.52	0.67	1.00					
AS	0.44	0.40	0.46	0.31	1.00				
MK	0.58	0.70	0.49	0.48	0.21	1.00			
MC	0.63	0.58	0.60	0.47	0.66	0.46	1.00		
EI	0.61	0.53	0.59	0.46	0.62	0.42	0.66	1.00	
AO	0.46	0.48	0.39	0.38	0.30	0.46	0.50	0.39	1.00

Note: Subtest scores from Phase 2 were not transformed to the Phase 1 scale.

Table G.5. Correlations Among Subtest Number Right Scores: Form 25A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.60	1.00							
WK	0.72	0.61	1.00						
PC	0.61	0.58	0.69	1.00					
AS	0.52	0.43	0.50	0.37	1.00				
MK	0.54	0.73	0.51	0.50	0.25	1.00			
MC	0.67	0.63	0.64	0.55	0.68	0.50	1.00		
EI	0.64	0.55	0.62	0.52	0.63	0.45	0.68	1.00	
AO	0.41	0.49	0.37	0.37	0.32	0.46	0.51	0.40	1.00

Note: Subtest scores from Phase 2 were not transformed to the Phase 1 scale.

Table G.6. Correlations Among Subtest Number Right Scores: Form 25B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.61	1.00							
WK	0.71	0.60	1.00						
PC	0.59	0.57	0.69	1.00					
AS	0.53	0.42	0.47	0.30	1.00				
MK	0.55	0.75	0.52	0.55	0.25	1.00			
MC	0.67	0.61	0.62	0.51	0.69	0.52	1.00		
EI	0.65	0.55	0.61	0.50	0.65	0.45	0.68	1.00	
AO	0.44	0.49	0.39	0.41	0.34	0.50	0.51	0.43	1.00

Table G.7. Correlations Among Subtest Number Right Scores: Form 26A (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.59	1.00							
WK	0.68	0.56	1.00						
PC	0.60	0.58	0.70	1.00					
AS	0.52	0.38	0.52	0.34	1.00				
MK	0.53	0.72	0.43	0.52	0.20	1.00			
MC	0.63	0.59	0.56	0.48	0.61	0.48	1.00		
EI	0.63	0.54	0.58	0.48	0.64	0.43	0.63	1.00	
AO	0.45	0.50	0.37	0.40	0.31	0.47	0.53	0.43	1.00

Table G.8. Correlations Among Subtest Number Right Scores: Form 26B (Phase 1)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.62	1.00							
WK	0.69	0.61	1.00						
PC	0.59	0.59	0.69	1.00					
AS	0.53	0.43	0.46	0.37	1.00				
MK	0.51	0.70	0.46	0.49	0.17	1.00			
MC	0.62	0.62	0.54	0.50	0.60	0.46	1.00		
EI	0.63	0.58	0.58	0.52	0.62	0.42	0.63	1.00	
AO	0.46	0.53	0.40	0.42	0.33	0.47	0.56	0.45	1.00

Table G.9. Correlations Among Subtest Number Right Scores: Form 27A (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.60	1.00							
WK	0.73	0.58	1.00						
PC	0.59	0.52	0.72	1.00					
AS	0.57	0.41	0.47	0.34	1.00				
MK	0.52	0.69	0.48	0.46	0.23	1.00			
MC	0.66	0.61	0.55	0.45	0.60	0.50	1.00		
EI	0.60	0.45	0.49	0.40	0.62	0.34	0.56	1.00	
AO	0.45	0.51	0.39	0.38	0.34	0.49	0.56	0.35	1.00

Note: Subtest scores from Phase 2 were not transformed to the Phase 1 scale.

Table G.10. Correlations Among Subtest Number Right Scores: Form 27B (Phase 2)

	GS	AR	WK	PC	AS	MK	MC	EI	AO
GS	1.00								
AR	0.60	1.00							
WK	0.70	0.57	1.00						
PC	0.57	0.53	0.72	1.00					
AS	0.56	0.41	0.45	0.34	1.00				
MK	0.49	0.67	0.44	0.45	0.18	1.00			
MC	0.65	0.61	0.53	0.44	0.59	0.45	1.00		
EI	0.61	0.45	0.49	0.40	0.62	0.31	0.58	1.00	
AO	0.42	0.50	0.36	0.34	0.32	0.45	0.54	0.35	1.00

Note: Subtest scores from Phase 2 were not transformed to the Phase 1 scale.