DMDC TECHNICAL REPORT 93-007

MILITARY APTITUDE TESTING: THE PAST FIFTY YEARS

Milton H. Maier

JUNE 1993

Personnel Testing Division
DEFENSE MANPOWER DATA CENTER

This report was prepared for the Directorate for Accession Policy, Office of the Assistant Secretary of Defense (Personnel and Readiness). The views, opinions, and findings contained in this report are those of the author and should not be construed as an official Department of Defense position, policy, or decision, unless so designated by other official documentation.

DEFENSE MANPOWER DATA CENTER
99 Pacific Street, Suite 155-A • Monterey, CA 93940-2453
Telephone: (408) 655-0400 Telefax: (408) 656-2087

MILITARY APTITUDE TESTING:

THE PAST FIFTY YEARS

Milton H. Maier

JUNE 1993

Personnel Testing Division

DEFENSE MANPOWER DATA CENTER

ACRONYMS USED IN THIS REPORT

ACB	Army Classification Battery
AFES	Armed Forces Examining Stations
AFHRL	Air Force Human Resources Laboratory (now Armstrong Laboratory)
AFQT	Armed Forces Qualification Test
AFVTG	Armed Forces Vocational Testing Group
AFWST	Armed Forces Women's Selection Test
AGCT	Army General Classification Test
AL	Armstrong Laboratory
AQB	Army Qualification Battery
AQE	Airman Qualifying Examination
ARI	Army Research Institute for the Behavioral and Social Sciences
ASP	Adaptability Screening Profile
ASVAB	Armed Services Vocational Aptitude Battery
AVF	All Volunteer Force
BUPERS	Bureau of Personnel (Navy)
CAST	Computer Adaptive Screening Test
CAT	Computer-adaptive testing
CEP	Career Exploration Program
CNA	Center for Naval Analyses
DAC	Defense Advisory Committee on Military Personnel Testing
DFK	Deliberate Failure Key
DMDC	Defense Manpower Data Center
ECAT	Enhanced Computer-Administered Testing
ECFA	Examen Calificacion de Feurzas Armadas
EST	Enlistment Screening Test
ETP	Enlistment Testing Program
GAO	General Accounting Office
GATB	General Aptitude Test Battery
GED	General Education Development Program
IOT&E	Initial Operational Test and Evaluation
JPM	Job Performance Measurement Project
MAP	Military Applicant Profile
MAPWG	Manpower Accession Policy Working Group
MARDAC	Manpower Research and Data Analysis Center
MEPCOM	Military Entrance Processing Command
MEPS	Military Entrance Processing Stations
METS	Mobile Examining Team Sites
MTP	Military Testing Program
NAS	National Academy of Sciences
NGCT	Navy General Classification Test
NORC	National Opinion Research Center
NPRDC	Navy Personnel Research and Development Center
OASD-FM&P	Office of the Assistant Secretary of Defense (Force Management and Personnel)
OASD-FM&P-AP	Directorate for Accession Policy in OASD-FM&P
PTD	Personnel Testing Division of Defense Manpower Data Center (The Testing Center)
STP	Student Testing Program
USAREC	US Army Recruiting Command

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The purpose of the military selection and classification testing program is to improve the quality of personnel decisions in all the Military Services. In the process of accessioning military recruits, personnel decisions are made at three different times:

- The first is *selection* in or out of the Service, depending on whether or not a person meets the minimum qualification standards.
- The second is *classification*, in which the occupational specialties for which a person meets the qualifying standards are determined.
- The third stage is *assignment* to a specific occupational specialty for which a person is qualified.

Assignment to a specific occupational specialty is based in part on qualification standards and in part on the needs of the Service. Qualification standards are more complex than just mental standards; they also include medical and moral standards. Mental standards include educational level as well as aptitude test scores.

The military testing program (MTP) began with the aptitude tests used during World War II, the Army General Classification Test (AGCT) and the Navy Basic Test Battery which included the Navy General Classification Test (NGCT). The AGCT was taken by over nine million recruits who entered the Army, Air Force, and Marine Corps during World War II, and the NGCT was taken by over three million sailors. During World War II, the Army, Navy, and Air Force each set up research facilities to develop military aptitude tests. These research centers remain in existence, and their functions have expanded to include other personnel areas in addition to aptitude testing; research on training, recruiting, human factors, and team performance currently are other areas studied.

From the time the peacetime draft was initiated in 1948, until the end of the Vietnam War in 1973, military accessions were obtained through a combination of the draft and voluntary enlistment. Beginning in 1973 with the All Volunteer Force (AVF), the accessioning process was restructured to obtain a sufficient number of voluntary enlistees, and the MTP likewise was radically restructured to facilitate recruiting.

CHAPTER 1
OVERVIEW OF MILITARY SELECTION AND CLASSIFICATION TESTING

Classification tests are designed to measure aptitudes related to performance in different occupational areas, notably maintenance and repair occupations, clerical and administrative occupations, and other occupations, such as medical, combat arms, intelligence, and operators of equipment. Aptitude composites, made up of scores received from combinations of tests in the classification batteries, are used to help determine qualification of recruits for assignments to military occupational specialties. Procedures for developing these military aptitude tests typically include the following steps:

- Identify the skills and knowledge that underlie performance in an occupational area.
- Develop experimental tests that may predict performance in the area.
- Administer the new tests to people assigned to military occupational specialties in that area.
- Follow the examinees to evaluate how well the test scores predict subsequent performance in the occupational area; usually final grades in training courses for the occupational specialties in the area are used as the criterion measure of performance.
- Evaluate the predictive validity of the new tests.
- Retain tests that improve the prediction of performance in one or more occupational areas.
- Prepare a test battery for administration to new or potential recruits for use in making personnel decisions.

Procedures for systematically assigning recruits to occupations did not become widespread until the 1950s, when improved technology enabled more timely matching of the aptitudes of recruits and the needs of the Services. An elaborate communication system was established to project the personnel vacancies in field units in the various specialties, and from these projected vacancies to determine the number of people to be trained in each specialty. Accessioning goals were set to help ensure an adequate supply of qualified people to fill the training vacancies.

In the 1960s, computer programs were developed to assign batches of recruits to occupational specialties in such a way as to minimize transportation costs from the sites of basic training to the sites of occupational training courses, and simultaneously maximize the mean expected performance of all recruits for all occupational specialties. The assignments were optimal in these respects. Variations of these assignment procedures have been used by all Services since then for matching people and occupational specialties.

Currently, each Service derives a set of aptitude composites to help set classification standards and make classification decisions. In the 1980s, the Air Force and Marine Corps each had four aptitude composites, the Army had nine, and the Navy had twelve. The number of aptitude composites each Service uses has been based to some extent on empirical data and to a large extent on traditional practices. The Services continue to compute their own sets of aptitude composites from the ASVAB, each tending to use different combinations of tests in the various composites.

Joint-service Testing

With the inception of the peacetime draft in 1948, the need for a joint-service selection test to test potential inductees became apparent. The Armed Forces Qualification Test (AFQT), modeled after the AGCT, was introduced on January 1, 1950, and taken by millions of registrants for the draft and applicants for enlistment from 1950 until 1973.

Beginning in 1973, when use of the separate AFQT was made optional by the Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P), the Services obtained an AFQT score from their classification batteries. The Army used a version of the Army

Classification Battery introduced in 1973. The Navy used its Basic Test Battery, introduced some years earlier. The Air Force and Marine Corps used a version of the Armed Services Vocational Aptitude Battery (ASVAB) that was parallel to the one used in the Student Testing Program (STP). Thus, in the years 1973, 1974, and 1975, the examining stations had to administer three separate classification batteries, each of which required over three hours of testing time and separate testing facilities. In addition to the strain of the transition from the draft to the AVF environment, the examining stations had to cope with the burden of administering three separate test batteries.

On January 1, 1976, the Services all stopped using their own classification batteries and started using the joint-service ASVAB, which was introduced to facilitate the accessioning of recruits by (a) permitting applicants to shop among the Services without taking several batteries, and (b) reducing the testing burden on examining stations. Since 1976, all applicants for all Services take the same battery of tests, and the separate Service batteries are no longer used. The AFQT score is derived from the ASVAB and is used to help set selection standards, help determine eligibility for special treatment (i.e., enlistment bonuses), facilitate manpower management, and report the quality of accessions to the Congress.

Qualification Standards

Qualification standards are set so that qualified people have a sufficiently high probability of being satisfactory performers, while unqualified people are likely to be unsatisfactory performers, consistent with the needs of the Services for an adequate number of recruits. Test scores are indicators of performance, and score scales are used to express the scores in terms that imply levels of expected or predicted performance.

During World War II, the aptitude tests had limited use to help make selection decisions; that is, few people were excluded from serving solely because of low test scores. The widespread use of aptitude tests to set military selection standards began after World War II; since the Korean War, the bottom ten percent of the mobilization population have been excluded from serving.

CHAPTER 2
MANAGING THE MILITARY SELECTION AND CLASSIFICATION TESTING PROGRAM

The first AFQT, conceived in 1948 to meet the needs of the draft, was developed by a joint-service committee that was supplemented with outside statistical and testing experts. The Army served as the executive agent for developing and administering the AFQT. Beginning in 1973, testing with the Service classification batteries was moved from recruit centers to the examining stations located throughout the country.

The first ASVAB Working Group, composed of technical and policy representatives from each Service and initially chaired by a representative from the OASD-FM&P, started meeting in 1974 to develop the first joint-service ASVAB for use in testing applicants for enlistment. The ASVAB Working Group later was expanded to include representatives from the Military Entrance Processing Command (MEPCOM).

Responsibility for research on measurement issues related to the accessioning process has been carried out by various executive agents. The Air Force served as the executive agent for development of the ASVAB from 1972 until 1989, and the Navy has served as the executive agent for related research efforts, such as developing computer-based testing. In 1989, management of the MTP took a new direction with formation of the Personnel Testing Division of the Defense Manpower Data Center. This Testing Center currently serves as the executive agent for the ASVAB, and as other research efforts reach the stage of operational use, they will become the responsibility of the Testing Center.

The Testing Center works closely with the Manpower Accession Policy Working Group (MAPWG), the title assumed by the ASVAB Working Group in the mid 1980s. Issues that affect the MTP are discussed and resolved if possible at the working-group level. The MAPWG and the Testing Center keep the Defense Advisory Committee for Military Personnel Testing (DAC), composed of testing experts from academia and industry, informed of plans and results, and this Committee provides input as it sees fit.

The Manpower Accession Policy Steering Committee is composed of flag officers responsible for the Services' military manpower policies, plus the Commander of MEPCOM; it is chaired by the Director for Accession Policy in OASD-FM&P (OASD-FM&P-AP). The Steering Committee reviews and approves larger MTP issues, such as introducing new forms of the ASVAB and providing troop support for research efforts. Final authority for the MTP resides in the ASD-FM&P.

In the 1980s, research efforts were begun to change (a) the testing medium of the ASVAB from the paper-and-pencil mode to computer-based testing, and (b) the battery content by substituting and/or adding new kinds of tests. These efforts have received the careful attention of the entire military manpower management community. The principals are aware of the need to maintain the technical rigor of the MTP, and they help provide the resources needed to maintain its quality.

CHAPTER 3
THE DEPARTMENT OF DEFENSE STUDENT TESTING PROGRAM

Military aptitude tests are made available to students in high schools and post-secondary institutions through the Department of Defense Student Testing Program (STP). The purposes of the STP are (a) to provide recruiting leads, (b) to help determine mental qualifications for enlistment, and (c) to provide aptitude test scores and other information useful in vocational counseling and career exploration.

The STP had its origins in 1958 when the Air Force offered to test students with a version of the Airman Classification Battery. Other Services quickly followed suit, and soon schools were visited by recruiters from each of the Services, all offering to test students with a military classification battery. In 1966, the OASD-FM&P directed the Services to develop a single test battery that could be used for vocational counseling of students and that would provide information useful in the accessioning process, such as accurate estimates of AFQT scores. With the Army serving as executive agent, the Services worked together to develop the first ASVAB, and it was introduced in 1968 for testing students.

The first editions of the ASVAB had modest materials for use by counselors and schools on how

to use the scores for vocational counseling. A counselor manual simply listed military occupations and contained a brief description of the aptitude composite scores reported to schools. However, the STP gained prominence with the advent of the AVF; the usefulness of the STP to obtain recruiting leads was apparent, and the STP grew to meet the need. A separate organization, the Armed Forces Vocational Testing Group, was created in 1972 to administer the ASVAB and develop materials for interpreting the test scores. Participation grew to include about 1,000,000 students in about 15,000 schools, and it has remained at about that level into the 1990s.

One problem with the STP surfaced in 1974 through a criticism made by Congressman Charles Mosher from Ohio. He was critical because the role of the military services in the STP was obfuscated in that the Services' use of the scores to obtain recruiting leads was barely, if at all, mentioned. In 1977, the Department of Defense developed a set of principles and actions that made the role of recruiting apparent to schools and students and protected the privacy of students. The resolution became known as the Mosher Agreement, and the STP continues to be governed by these rules.

In 1977 and 1978, the STP once again came under attack, this time by Professor Lee J. Cronbach, an expert in psychological testing. He was especially critical of the aptitude composites reported in the STP and the association of aptitude composite scores and civilian occupations; he believed that the composites were too highly intercorrelated and that the association of composites and civilian occupations was often implausible. The military testing community responded quickly by changing the aptitude composites and their association with civilian occupations.

In the 1980s, the OASD-FM&P-AP assumed greater responsibility for the STP, and the quality of the STP increased through the publication of counselor manuals, student workbooks, and other supporting materials.

In 1992, a new set of materials, called the ASVAB 18/19 Career Exploration Program (CEP), was introduced. A prominent feature of the new program is a student interest inventory, the results of which are used to supplement the aptitude scores provided by the ASVAB. The new materials were developed through close coordination with panels of experts in vocational counseling and aptitude testing. Responses to the CEP have been favorable, and resources are available to review and revise the materials on a regular basis.

Another major accomplishment in 1992 was the validation of the ASVAB for eleven civilian occupations that represent a variety of occupations high school students are likely to enter. The results support the usefulness of the ASVAB for use by students in exploring civilian occupations.

The CEP conforms to high professional standards; the ASVAB, the interest inventory, and the wealth of interpretive materials comprise a useful package that can be used by school counselors or by students themselves if no counseling assistance is available. The continuing attention to the STP helps ensure that future editions of the materials will also meet high professional standards.

CHAPTER 4
NORMING AND SCALING MILITARY SELECTION AND CLASSIFICATION TESTS

Before military aptitude tests are introduced to help make personnel decisions, their scores are calibrated to an existing score scale that can be interpreted in terms of expected performance. Qualification standards require a score scale that indicates levels of expected performance and that does not change meaning when new forms of a test are introduced.

The AFQT and Service classification batteries were placed on the World War II score scale that was computed from the distribution of AGCT and NGCT scores of men who served during World War II. Qualification standards, therefore, were tied to the abilities, and expected performance, of the World War II mobilization population. As new forms of the AFQT and Service classification batteries were introduced, the scores were calibrated to the World War II population, and the meaning of qualifying standards remained relatively constant.

The ASVAB, too, was scaled to the World War II population from its first use in 1968 in the STP. When the ASVAB was designated as the joint-service aptitude battery and introduced in 1976 as the replacement for the Service classification batteries, the intent was that it be on the World War II score scale. Qualifying standards for enlistment and assignment to occupational specialties were not changed in 1976 because both researchers and managers expected that the ASVAB scores had the same meaning as the scores from the tests it replaced.

The facts, however, proved otherwise. Problems with the score scale surfaced within two months after the ASVAB was put into use, and, based on analyses conducted by the Services, the AFQT scores derived from the ASVAB were adjusted in the top half of the score scale. Allegations that the scores in the bottom half of the scale were also in error persisted, and in 1979 data were collected that confirmed serious errors did exist in the bottom half of the score scale.

The ASVAB miscalibration lingered for close to five years before it was fully documented and fixed. Not until 1979 were serious joint-service efforts initiated to collect data on the magnitude of the problem, and in 1980 the error in the score scale was fully documented. The ASVAB score scale was fixed in October 1980, when new forms of the battery were introduced, and the traditional meaning of the scores, and qualifying standards, was restored.

Briefly, the problems arose because the researchers attempted to take shortcuts in scaling the new tests; the shortcuts were in response to the tight schedule for developing the new tests and to a perceived requirement that extra testing of applicants for enlistment be kept to an absolute minimum for fear that some of them might become disgruntled by the extra testing and decide not to enlist. Also, clerical errors were made that resulted in inflated scores on the new tests.

Development of the first joint-service ASVAB was accomplished in less than two years, beginning in May 1974 when the Services were directed by OASD-FM&P to develop a joint-service classification battery, and ending on January 1, 1976, when the ASVAB was introduced for testing all applicants. (For perspective, in the 1980s over four years are required to develop and introduce new forms of the ASVAB.)

The errors in the ASVAB score scale resulted in inflated scores reported for examinees. The effect of the inflated scores was to permit hundreds of thousands of people to qualify for enlistment

and assignment to technically demanding occupational specialties who would not have qualified if the scores had been accurately scaled to have their traditional meaning.

Prior to the ASVAB miscalibration, manpower managers in the OASD-FM&P little knew or cared about the details of the MTP. Their primary concern was that enough qualified people were recruited and retained. As the ASVAB miscalibration unfolded, they learned more about the meaning of the test scores, and they did not like what they found. Two facts were especially troublesome: one was that the then current scores on the ASVAB were calibrated to the World War II mobilization population, and no one knew how the population of potential recruits in 1979 would have scored on the ASVAB; the second fact was that no one had empirical data on how well the ASVAB predicted job performance in the occupational specialties. The ASVAB was known to be a valid predictor of performance in training courses for the occupational specialties, but the evidence for predicting performance on the job was sparse at best.

The managers responded quickly and decisively to these deficiencies. In the summer and fall of 1980, a new version of the ASVAB was administered to a nationally representative sample of American youth. A new score scale was developed for this population, called the 1980 Youth Population, and it was introduced in 1984 along with new forms of the ASVAB. The new score scale, called the 1980 score scale, is based on both females and males, whereas the World War II scale was based on only males. The 1980 score scale was a big technical improvement over the World War II scale; it was more rigorously defined, so it permitted more accurate scaling of the aptitude composites.

The response to the question of how well the ASVAB predicted job performance was to initiate a massive effort to develop measures of job performance and validate the ASVAB against these measures. The study involved thousands of enlisted people from each Service, and it took about ten years to complete the collection of data. Analyses of the data to evaluate enlistment standards is continuing well into the 1990s. The validation effort has shown that the ASVAB is a valid predictor of performance on the job, just as it is of performance in training courses.

A third outcome of the ASVAB miscalibration was to increase the technical rigor with which military aptitude tests are developed. During the 1980s, the MTP was restored to the high level of professional quality traditionally enjoyed by military aptitude tests. Both the STP and testing of applicants now meet high professional standards, and current efforts to further improve the MTP are receiving careful consideration and evaluation by interested agencies.

CHAPTER 1

OVERVIEW OF MILITARY SELECTION AND CLASSIFICATION TESTING

Since World War II, more young men and women have been employed by all the branches of the Military Services combined than by any other single employer. Some young people were inducted into service through the draft, but more served, and continue to serve, voluntarily through enlistment.

As employers, the Services must each make the following personnel decisions:

- Who is qualified to serve?
- What training and job assignments match the recruit's capabilities?
- Where are the needs of that Service?

The purpose of the military testing program (MTP) is to help improve the quality of military personnel decisions. Test scores are used to help make the selection decisions about who is qualified to serve, classification decisions about the occupational specialties for which people are qualified, and assignment decisions about the occupational specialties in which people are to be trained.[1]

Most recruits have limited job experience and training, or what they have is not relevant to military occupational specialties. Thousands of young people may be tested each day, which means that both effective and efficient procedures are needed to find out about their capabilities, or potential, to perform in the Military Services. Aptitude testing was identified as the procedure that would meet this need.

The Military Services were early leaders in developing and using aptitude tests to help make personnel decisions. During World War I, the Army tested recruits with a test of general mental ability, called the Army Alpha. Before World War II, the Army started developing the Army General Classification Test (AGCT), and during the war, the AGCT and the Navy General Classification Test (NGCT) were administered to all male recruits.

The development of the AGCT is described in two early reports published in the 1940s; for more information refer to *Psychological Bulletin*, Volume 42 (Staff, 1945), and *Journal of Educational Psychology*, Volume 38 (Staff, 1947). Development of the NGCT and the Navy Basic Test Battery is described by Stuit (1947).

Additional classification tests were developed early in World War II to supplement the AGCT and NGCT. These supplemental tests covered specialized aptitudes related to the technical fields

---

[1]Personnel decisions require other information besides test scores, such as medical and moral information, to determine qualification for serving; only aptitude tests and other testing instruments are addressed in this report.

(mechanical, electrical, and later electronics); there were also clerical and administrative tests, radio code operator tests, and later, language tests and driver selection tests. In June 1943, the Navy introduced an improved Basic Test Battery composed of the General Classification Test, a reading test, an arithmetic reasoning test, and three tests of mechanical aptitude and knowledge to replace an earlier version that had been in use since at least December 1941 (Stuit, 1947). In addition, clerical tests and a radio code aptitude test were available for use in the Navy. In 1947, the Army published a report, "Classification Tests for Reception Processing (Provisional Program)," (Army Research Institute for the Behavioral and Social Sciences [ARI], 1947, January), that set the stage for introduction of a formal classification system. By 1949, two more Army reports were published that covered the introduction of a classification system (ARI, 1949, September; ARI, 1949, October). The Air Force introduced the Airman Classification Battery in November 1948, for use in classifying and assigning recruits (Weeks, Mullin, & Vitola, 1975). Much of the theoretical foundation of classification testing was developed by Army testing psychologists (Brogden, 1955).

For manpower management purposes, the single most dominant score in the MTP has been the one from the Armed Forces Qualification Test (AFQT), most widely used in making selection decisions. All Services continue to have formal selection standards defined by the AFQT that each new recruit must meet. During World War II, aptitude tests were used to help attain an equitable distribution of ability across the branches of the Army, especially between the air and ground-combat forces. The AFQT was also used during times of mobilization to help attain an equitable distribution of ability across the Services.[2] This practice was used during both the Korean and Vietnam Wars. Also during times of mobilization, as in the Vietnam era, the AFQT standards played a vital role in helping determine mental qualifications of new recruits. Currently, in the All Volunteer Force (AVF) environment, the AFQT scores are used primarily to (a) help determine the eligibility of new recruits for special programs (such as enlistment bonuses); (b) develop reports for Congress about the quality, or aptitudes, of the enlisted force; and (c) help define manpower policies.[3]

Matching Abilities of Recruits to the Needs of the Services

The information that aptitude tests provide about the abilities of people to perform in the various occupational specialties open to recruits is, by itself, not sufficient to result in accurate assignment decisions. As employers, the Services must match the abilities of new people to the requirements of the specialties. In this section, procedures for accomplishing this matching are discussed.

The starting point in the matching is the force structure of each Service, which determines the number of people and grade levels in each occupational specialty. The Services project the losses of people from each specialty, and the accessioning process is tasked to provide an adequate number of qualified people to meet the projected needs. Intervening between accessioning and utilization of people on the job in field units are the training courses that prepare people to perform their jobs. The training, including basic training, may last from a few months to over a year. The extended training pipeline means that many slips can occur between the initial assignment to a specialty at the time of accessioning and subsequent utilization in a field billet.

In today's communication environment where the access to information is almost instantaneous,

---

[2]See Appendix A for information about use of the AFQT during times of mobilization.

[3]The term "currently" in this report refers generally to the early 1990s.

the matching process can allow a new recruit considerable choice about assignments and still enable the Services to meet their needs for enough qualified people in each specialty. In former times, when the flow of information was more cumbersome, a longer lead time was required to match the abilities of recruits to the needs of the Services.

The steps involved in carrying out the match are as follows.

- Determine the number of vacancies that must be filled in each specialty, based on force structure and projected losses.
- Factor in training times, schedules, and projected losses during training.
- Establish goals or quotas for the numbers of accessions in each specialty to meet the training schedule and the needs of the field units.
- Obtain sufficient numbers of qualified people.

With modern, centralized, manpower data files, the projection of losses from field units can be done with reasonable accuracy and speed. Because the training times and losses of people from training courses are well known, the goals or quotas fall out from the algorithms used by manpower managers to maintain the force structure.

In the AVF environment, where people must be actively recruited and persuaded to join the military, all Services allow recruits some choice in their assignments to occupational specialties. The degree of choice may range from a specific start date for training in a specific specialty, to choice of an occupational area such as electronics, with the Services retaining the right to pick the specific specialty in the area.

The offering of choices to enlistees was inaugurated by the Army during the Vietnam War to attract people to enlist. People who enlisted for a three-year tour were allowed to select an initial assignment from an appropriate listing, whereas people who were drafted had to take the specialty to which the Army assigned them. The advertising slogan used by the Army was "choice, not chance." After the Vietnam War when the AVF replaced the draft, the offering of choice of assignments was adopted by all Services.

The automated algorithms for assigning recruits were introduced in the early 1960s, first by the Marine Corps, and then adapted by the Army in time for the massive buildup during the Vietnam War (Boldt, 1964, April). The choice of assignments by enlistees in the late 1960s was feasible because of these algorithms. A computer program was available to assign up to 5,000 people at a time to specialties. The input to the algorithm was the number of vacancies in each specialty and the aptitudes of the recruits. The first step was to assign the enlistees who had a guaranteed assignment to the specialty of their choice, reducing quotas accordingly. The remaining recruits, mainly draftees, were then assigned to the remaining training vacancies.

The assignment of recruits who did not have guarantees was further controlled by policy decisions to (a) minimize the mean transportation cost from the site of basic training to the site of specialty training, and (b) maximize the mean aptitude score of all people across all specialties. When the constraints of quotas and transportation costs were met, the degree of freedom for capitalizing on aptitude scores were greatly diminished, and in effect, aptitude scores had relatively small influence

on assignments because the other factors were satisfied first.

Prior to computerized algorithms, assignments by the Army were made using punched cards and sorting machines. The matching process was essentially the same as is done by computer today, except that it was more cumbersome and did not attempt to maximize the mean aptitude score of the people assigned. Maximization requires numerous iterations of trial assignments, and iterations were not feasible with punched cards and sorters.

Before the days of punched cards, the matching was done by hand. Classification officers and their staffs would attempt to match aptitudes and needs of the Service, to the extent that they knew the needs. All too often, the matching was done simply by lining up the recruits and counting them off; one group would be sent to one specialty, another group to a different specialty, and so on. Aptitude test scores may have been used in some cases (e.g., encouraging people with high scores to apply for training to become commissioned officers), but when the flow of people was heavy and the time pressures great, assignments tended to be made in the most expeditious manner.

Built into the assignment process, both for enlistees with a guaranteed assignment and other people who were assigned by the Service, were classification standards. All recruits in all Services since the late 1940s have had to meet the minimum aptitude standards for the specialty to which they are assigned. However, the Services can, and did, waive the minimum classification standards when necessary. For example, during the Vietnam period, the minimum selection standards were so low that many recruits were not qualified for any specialty, or the specialties for which they were qualified had already been filled by people with higher aptitude scores. These people, called no-quals, were rejected by the algorithm and had to be assigned by hand. Typically they were assigned as infantrymen, cooks, or stevedores.

In the current person-job matching procedures, a list of openings in specialties is presented to prospective recruits. The list is determined by the aptitude scores of the person and the needs of the Service. People may choose from the list, or they may reject all offerings, in which case a new list is generated. Eventually the people choose an assignment, or they decide not to enlist. Some enlistees do not choose a specialty at the time of enlistment and simply accept the specialty to which the Service assigns them. The quotas for each specialty are changed as recruits are assigned to them and as the needs of the Service change.

The burden of obtaining enough qualified people falls on recruiters and counselors at processing stations. Enlistment bonuses, choice of location, and special college funds are devices commonly employed to induce accessions to sign up for specific assignments. Conversely, special restrictions may be placed on specialties that are especially desirable, such as an extended tour of duty of up to six years in the nuclear area. Recruiters and counselors must keep abreast of the needs of the Services and of special programs available to fill the gamut of openings.

Developing Military Selection and Classification Tests

The fundamental criteria for using aptitude tests to help make personnel decisions is that they have validity for predicting subsequent performance in training courses and on the job. Essentially, predictive validity means that test scores are related to performance, in that people with high scores are likely to perform well, whereas people with low test scores are likely to perform less well. Only to the extent that test scores are related to performance is their use in making personnel decisions

justified.

The great strength of military aptitude tests is the solid empirical foundation demonstrating their usefulness for making personnel decisions. All military aptitude tests dating back to World War II were carefully evaluated before being introduced for operational use, and they were, and continue to be, evaluated further during operational use.

The procedures for developing military selection and classification tests historically and currently involve the following steps.

- Formulate hypotheses about test content that may be related to performance in an occupational area. (Researchers typically observe and talk to workers in the area and visit training programs. Occupational areas that have received special attention for developing tests include electronics repair, aircrew members, combat arms, including Special Forces, and nuclear technicians.)

- Construct experimental tests that measure the skills, knowledge, and aptitudes needed for success in that occupational area.

- Administer the experimental tests to students in training courses for specialties in the area and perhaps to job incumbents.

- Analyze the results from the experimental tests to determine their psychometric properties and to estimate their predictive validity.

- Produce revised tests based on results of the analyses of the experimental tests.

- Administer the revised tests to samples of trainees in the occupational area.

- Analyze the results from the revised tests to evaluate the validity of the tests as predictors of performance in the job training courses.

- Incorporate promising new tests into an experimental battery, perhaps with tests that cover other occupational areas.

- Administer the experimental battery to samples from the major occupational areas.

- Analyze results from the experimental battery to evaluate the differential validity of the new tests relative to existing tests.

- Add the new tests that improve the accuracy of prediction or classification efficiency.

- Scale the new tests to make the scores comparable to existing tests for purposes of setting standards and making assignment decisions.

Historically, the rigor of following these procedures varied from time to time. Essentially, they were followed before the Service classification batteries were revised; empirical evaluation of their usefulness for making personnel decisions was a necessary prerequisite. The Services were justifiably proud of their classification batteries, and users in the Services were reasonably well satisfied with the

outcomes of personnel decisions based on aptitude tests. The Services continued to improve their classification tests throughout the 1950s and 1960s, and into the early 1970s. For example, the Army introduced a new battery and classification system in May 1973 to replace the previous version which had been introduced in 1958.

In the typical validation study of existing tests, aptitude scores of record are obtained from the files and their accuracy in predicting subsequent performance is evaluated. Typically, subsequent performance is measured by grades in training courses for occupational specialties. These grades have many advantages as measures of performance because virtually all recruits go through a formal training course before being assigned to a billet in a field unit, and grades are measures of performance in the first assignment. The main advantage of grades is that they are economical because they have been routinely available. Almost all validation studies in the military have included training grades as a measure of performance.

The validity of military aptitude tests for predicting performance in training courses across the range of specialties for all Services is on average about 0.6 (Welsh, Kucinkas, & Curran, 1990, July). Some occupational specialties which involve physical skills and endurance, such as infantry, traditionally have been less predictable than specialties that require conceptual knowledge and problem-solving ability, such as technical repair and medical specialties. But when hands-on job performance tests are used as the criterion measure to be predicted, aptitude tests can predict job performance of infantrymen and mechanics at 0.6 or better, which is comparable to the usual validity against training grades. The empirical base for predicting job performance is not yet fully developed, but indications are that a validity coefficient of 0.5 to 0.6 is a good summary value across a wide range of specialties (Wigdor & Green, 1991). A correlation of 0.6 is thus a good single value to describe the predictive accuracy of military aptitude tests.

Interpreting the Validity Coefficient

The validity coefficient by itself does not indicate the usefulness of tests for making personnel decisions. The magnitude needs to be translated into terms meaningful to manpower managers. An interpretation of test validity has been formulated by Brogden that shows the gain in performance of a group of performers selected on the basis of test scores, over that which would be obtained from a group selected at random (Brogden, 1946). Random selection has a validity of 0.0, and it defines one end of the scale. The other end of the scale is defined by perfect prediction, a value of 1.0. If people could be selected on the criterion of performance itself (i.e., if selection decisions were made after people performed in their duty assignments and the measure of performance were perfectly accurate), then the decisions too would be perfectly accurate, and they would have a validity of 1.0. The performance of this selected group would be the highest possible; no other selection procedure could result in a group of the same size selected from the same pool of people that had a higher level of performance.

With a validity of 0.6, the gain in performance of groups selected with aptitude tests is 60 percent of the difference in performance between a group selected at random and a group selected on the criterion itself. The relationship between the validity coefficient and gain in performance is linear: tests with a validity of 0.7 result in a gain of 70 percent, and tests with a validity of 0.5 have a gain of 50 percent.

The meaning of a gain in performance can be translated into reduced failure rates, especially in

training courses. The higher the validity coefficient and the more difficult the training course, the greater the reduction in failure rates when people are selected on the basis of their test scores instead of selected randomly. Training failures are costly to the Services, and reduction in failure rates can result in tangible dollar savings.

Brogden developed a formula for estimating classification efficiency resulting from the use of aptitude tests (Brogden, 1959). The formula is as follows.

$$CE = Rxy\ (1 - Rxx)^{1/2}$$

CE is classification efficiency, Rxy is the mean validity coefficient of the aptitude scores, and Rxx is the mean intercorrelation among the performance estimates. The formula shows that the effectiveness of aptitude tests for making assignment decisions is reduced by the intercorrelation among the performance estimates. At the extreme, if the intercorrelation among the estimates were 1.0, classification efficiency would be 0.0.

Aptitude composites are used by the Services as performance estimates, and they can reasonably be substituted into the formula as performance estimates. The intercorrelation among the aptitude composites used by the Services during the 1980s tended to be about 0.8 to 0.9. With the intercorrelation this high, classification efficiency was low, and the gain in performance was relatively small compared to the gain that could have been realized if the aptitude composites were less highly intercorrelated. Increases in classification efficiency would have resulted because a larger percentage of recruits would have had aptitude scores in the above-average range on at least one aptitude composite; if they were assigned to the occupation areas in which they scored higher, the mean level of performance for the recruits as a whole would have been increased. Brogden's formula for classification efficiency reflects the fundamental requirement that aptitude tests measure different abilities; the subsequent classification based on that differential measurement could increase the performance of the group as a whole.

The value of the gain in performance--or classification efficiency--from using tests to select and assign people is difficult to quantify. Better performance from workers is valuable, as is obvious to any employer, supervisor, or customer. Quantifying how much the gain is worth to employers is still a controversial area, especially for the Military Services that do not compete in the same economic marketplace as profit-making enterprises. In the past, the military testing community used intuitive arguments about the value of the gain in performance, and managers tended to be convinced. In the 1980s, the need to attach dollar values to the outcomes of decisions became increasingly prevalent. Any substantial changes to military testing procedures would be easier to introduce if the benefits could be described convincingly in dollar terms.

Content of the Current ASVAB

The Armed Services Vocational Aptitude Battery (ASVAB) has been in use since January 1976 as the joint-service selection and classification test battery.[4] It contains eight power tests (which have relatively long testing times) and two speeded tests (with short testing times) that measure abilities in the content areas as shown in Table 1.

---

[4]The lineage of the AFQT and the ASVAB is given in Appendix B.

Table 1

Content of the Current ASVAB

- Verbal (VE)

  Word Knowledge (WK) - vocabulary items
  Paragraph Comprehension (PC) - reading comprehension

- Mathematics

  Arithmetic Reasoning (AR) - word problems
  Mathematics Knowledge (MK) - fractions, algebra, geometry

- Technical Knowledge

  Auto and Shop Information (AS) - cars and tools
  Mechanical Comprehension (MC) - physical principles
  Electronics Information (EI) - electricity and electronics

- Science

  General Science (GS) - physical and life sciences

- Speed and Accuracy

  Coding Speed (CS) - matching words and numbers
  Numerical Operations (NO) - simple arithmetic computations

The entire ASVAB, which requires about three hours to administer, is administered to all applicants for enlistment at testing stations located throughout the country and overseas. Typically, the test is administered to groups of examinees in the paper-and-pencil mode; however, there are a few test sites that now use computers.[5] The ASVAB is also given in high schools and post-secondary training institutions, with results used to (a) provide vocational counseling, (b) assist in career exploration and (c) to obtain recruiting leads.[6]

Scores from the each of the tests in the ASVAB are combined into composite scores that are used to help make personnel decisions. For example, the AFQT composite score is defined by the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge scores. Other combinations of ASVAB scores, called aptitude composites, are used to help classify and assign recruits. Table 2 lists the composites currently in use by the Services.

[5]Chapter 2 of this report contains more information about computerized testing.

[6]Chapter 3 describes the Student Testing Program in detail.

Table 2

Current Composite Definitions

Service	Composite	Definition
All	VE	WK + PC
	AFQT	2VE + AR + MK
Army	MM	NO + AS + MC + EI
	CL	AR + MK + VE
	ST	VE + MK + MC + GS
	EL	AR + MK + EI + GS
	CO	CS + AR + MC + AS
	FA	AR + CS + MC + MK
	SC	AR + AS + MC + VE
	OF	NO + AS + MC + VE
	GM	MK + EI + AS + GS
	GT	VE + AR
Navy	ME	VE + MC + AS
	CL	NO + CS + VE
	GT	VE + AR
	EL	AR + MK + EI + GS
	E	AR + GS + 2MK
	EG	MK + AS
	ST	VE + AR + MC
	MR	AR + MC + AS
	HM	VE + MK + GS
	CT	VE + AR + NO + CS
	BC	VE + MK + CS
Air Force	M	MC + GS + 2AS
	A	NO + CS + VE
	G	VE + AR
	E	AR + MK + EI + GS
Marine Corps	MM	AR + EI + MC + AS
	CL	VE + MK + CS
	GT	VE + AR + MC
	EL	AR + MK + EI + GS

Defining Aptitude Composites

Typically, scores from three or four tests are combined to yield an aptitude composite that is used to determine eligibility for assignment to occupational specialties. All Services have four aptitude composites that are related to the same families of specialties:

Mechanical or Mechanical Maintenance
Clerical or Administrative
General or General Technical
Electronics or Electronics Repair

Currently, the Air Force and Marine Corps have only these four composites. The Army and Navy have additional composites to cover occupational areas as diverse as combat arms and boiler technician/engineman/machinist's mate. Even though the Services have composites with the same or similar names, the definitions of the composites are specific to each Service.

There is no standard analytical procedure to determine either (a) how each composite is defined, or (b) how many composites each Service should have; the decisions are based on tradition and interpretation of research results. Validation studies show only the degree of statistical relationship between a score or set of scores and a criterion measure of performance. The analysis informs, but does not determine, decisions about which tests to include in a composite or how many composites should be used. These decisions are interdependent, but will be discussed separately.

The general practice used in the Military Services for defining aptitude composites is to select the tests that in combination have the highest validity for predicting performance in a set of specialties with similar job requirements. The statistical procedure is called test selection, in which the test with the highest validity is selected first, and the test that then adds the most validity is selected second, and so on until the test selection is stopped. Typically the composites contain three or four tests, sometimes only two, and the test selection tends to stop when additional tests add little or nothing to the validity of those already selected.

Normally aptitude composites are defined by the test selection procedure because the results are reasonable. Most composites contain at least one test of mental ability and one or more technical test and/or test of speed and accuracy, as appropriate. Sometimes, however, routine test selection does not produce reasonable results, and then judgment must intervene. For example, when classification tests were validated against training grades in the Marine Corps during the 1980s, the results of test selection indicated that the Auto and Shop Information (AS) test should be included in the composite for infantry specialties. The use of the AS test for infantrymen did not seem reasonable, and Mechanical Comprehension (MC) was proposed instead. The loss in validity by substituting MC for AS was small, and the effects on the quality of personnel decisions were minimal; the MC test was accepted by manpower managers. Later in the 1980s, when the classification tests were validated against hands-on performance tests for Marine Corps infantrymen, the AS test again was picked through test selection for inclusion in the aptitude composite for infantrymen. However, the Marine Corps managers did not change the tests in the composite for infantrymen even though validity could have been improved by replacing MC with AS. These anomalies occur infrequently, and usually both researchers and managers agree on the definitions of composites.

The number of aptitude composites a Service should use is more subject to judgment and administrative convenience than is the definition of the composites. The minimum number of

composites is one, of course, but in that case any classification efficiency is lost. The maximum number would be a separate composite for each specialty to which recruits can be assigned; however, manpower managers and the automated manpower data bases would be severely taxed by so many aptitude composites because the number of specialties open to recruits may be well over 100 in a Service. Having different composites for specialties requiring similar skills and knowledge is not reasonable; many specialties with different training courses do in fact have similar job requirements. For example, sometimes aircraft mechanics for different models of similar aircraft are trained in separate courses, even though the content is similar among the courses. Therefore, with some empirical support, the number of aptitude composites used by the Services ranges from four to eleven. (Including the AFQT composite, there are a total of five to twelve composites used.)

Traditionally, the number of composites has been held to the minimum that Service researchers and managers thought was necessary. Before the days of computers, all computations of composite scores were done by hand, and the computational burden could not exceed the capability of the staffs. The Army started with ten composites in 1949, in the 1950s the number was reduced to eight, and in 1973 it was increased to nine. The Air Force has used four composites since 1958; before that it had used up to eight composites. The Marine Corps followed Army practice until the early 1970s, when it reduced the number of composites to six, and in the 1980s to four. The number of Navy composites was at times not readily apparent; sometimes, the Navy imposed additional standards to a minimum score on an occupational composite (such as a minimum score on a specific test) for assigning recruits to some specialties. In all cases, though, the number of composites has been within the computational capability of each Service.

Besides the number of scores in a composite, another constraint on the computation of composite scores is the weight assigned to each test in a composite. For convenience, most weights have been, and continue to be, one, instead of being proportional to the regression weights. (Over the years, a few tests have been weighted two; additionally, NO was 1/2 weighted in the AFQT prior to 1989.) When the military aptitude composites were defined in the 1980s, all Services either followed traditional practices and kept the number of composites the same, or they reduced the number and used unit weights (see, for example, Booth-Kewly, Foley, & Swanson, 1984). Although the computational capability of the testing centers currently may be sufficient to handle more complex weights, a big constraint on the number of composites is that each Service, plus the Department of Defense, retains an automated personnel record for each person; it contains aptitude composite scores that could be computed each time they are needed, but the computational burden on the users would be great. Also, when changing the number and definition of composites, the test of reasonableness to managers must be met. Explaining the need for large numbers of composites and complex weights, although technically elegant, is not easy.

The number of composites is also constrained by the number of meaningful occupational areas that can be distinguished. Three occupational areas seem to have different job content and may require different aptitudes: maintenance and repair specialties, clerical and administrative specialties, and everything else. The Services do distinguish between mechanical and electronics repair specialties by having separate composites.

The rationale for retaining the current number of composites relies heavily on tradition and convenience. Changing aptitude composites is a complex process, and the Services do not enter into changes without good reasons. Even small changes to a composite, such as substituting one test for another, involve extensive changes throughout the scoring and reporting procedures. Computer programs, scoring manuals, and personnel regulations need to be changed. When new composites are

introduced, or old ones deleted, the changes are even more extensive. In addition to changing the scoring procedures, new reporting forms must be developed, and more importantly, the classification and assignment procedures must be revamped to change the association between composites and specialties. Recruiting and training commands need to be informed about changes, so they can change their procedures accordingly. A side effect of the complexity of changing composites is that sometimes past practices may be perpetuated even though they may be cumbersome and have obvious defects. Unless some clear reasons for changing are found, such as changing the content of classification tests, the composites are likely to continue in their present form.

Even though the Services have used aptitude composites continuously since 1949, arguments about why they are needed have never totally subsided: why not use a single measure of general mental ability? In the 1970s, the theory of validity generalization was advanced as a procedure for inferring the validity of tests (Schmidt & Hunter, 1977). Essentially, the argument for validity generalization is that aptitude tests are valid for all jobs and there is no need for differential aptitude composites. The conclusion from validity generalization is that a single index of ability is sufficient for selecting, classifying, and assigning recruits. If the argument for validity generalization wins out over differential validity, then the use of classification test batteries will be called into question.

Historically, the Services have relied on validity data to justify the use of differential aptitude batteries and composites. However, as the advocates of validity generalization argue, the evidence for differential validity is hardly compelling. The reasons for low differential validity may lie more in the way aptitudes and criterion performance are measured than in the inherent similarities or differences in aptitudes and job requirements. In general, both sets of measurements, aptitudes and performance, have used the paper-and-pencil mode, which relies heavily on words, numbers, and mental concepts rather than hands-on skills in performing job tasks. A quick review of military classification tests and grades in specialty training courses, which typically have served as the criterion of performance, will bring out the difficulty in establishing the differential validity of classification batteries.

Evaluating the skills and abilities of students needed to perform job tasks in an accurate and objective manner is difficult and very expensive. Verbal, quantitative, and conceptual skills and knowledge tend to be emphasized in both the aptitude and criterion measures. The ASVAB, as any paper-and-pencil test used in mass testing programs, is limited in content to items that feature words, numbers, or pictures of objects; examinees are asked to identify meanings of words or concepts, solve problems, or manipulate objects conceptually. Similarly, training grades have been based on paper-and-pencil achievement tests given during the training courses. These similarities contribute greatly to the high degree of overlap among the different tests in an aptitude battery and to the similar validity of most tests for predicting grades in most training courses. Even if job requirements in fact are different for different job families, and people do in fact have differential aptitudes, the measurement techniques in the paper-and-pencil mode tend to prevent the differential nature from emerging in validity studies.

Training grades are sometimes questioned as being appropriate criterion measures for validating aptitude tests because they are not direct measures of job performance. In the 1970s, when the success of the AVF was still unproven and the usefulness of military aptitude testing was suspect, an embarrassment to the military testing community arose from the lack of validity information about how well aptitude tests predicted performance on the job. In fact, a common allegation that grew up in the 1970s was that aptitude tests could predict training grades but not job performance. To help answer questions about the usefulness of aptitude tests as predictors of job performance, the Services

in the 1980s embarked on massive studies to validate the ASVAB directly against measures of job performance. The benchmark of job performance was a hands-on performance test that covered the skills and knowledge required to perform a critical job task. The results, contrary to allegations made by critics of the tests, were that aptitude tests could predict job performance to the same degree of accuracy as they predicted training grades. The confidence of manpower managers for using aptitude tests is strengthened by these results showing the validity of the ASVAB against job performance.[7]

Test Fairness

The Military Services have a long history of concerns about the fairness of tests for minorities: women, Blacks, and Hispanics. Several steps have been taken throughout the past fifty years to ensure test fairness.

When all Services started to allow women to join their ranks during 1942, the Army developed and calibrated a selection test designed especially for women (ARI, 1942, November). Later, special tests for selecting women were used by all Services. The Armed Forces Women's Selection Test (AFWST) was introduced on January 1, 1953 along with forms 3 and 4 of the AFQT. With the addition of tool knowledge items to the AFQT in 1953, the adverse impact for women was unacceptable; similarly, the first versions of the AFWST with the same content as the AGCT (verbal, arithmetic reasoning, and spatial ability) had an undesirable adverse impact for women. So, new forms of the AFWST that did not contain the spatial items were introduced in 1956 (Morton, Houston, Mundy, & Bayroff, 1957, May). These forms contained only verbal and arithmetic reasoning items; the spatial relationships and tool knowledge items were not used to evaluate the general mental ability of females. However, beginning in 1976, minimum enlistment standards for females were set on the AFQT, which contained spatial ability items along with the verbal and arithmetic reasoning items. The adverse impact of the spatial ability items in the AFQT again became an issue in the late 1970s, and one of the reasons for dropping these types of items from the AFQT in 1980 was that the AFQT score would be more fair for females. Since 1980 the AFQT has contained only verbal and math items.

The AGCT was translated into Spanish as early as 1941, titled the Examen Calificacion de Fuerzas Armadas (ECFA), and used instead of the AFQT as the first screen for Puerto Ricans (ARI, 1945, August). Puerto Ricans taken into the Army during the 1950s and 1960s were given special language training as required, and generally went through basic training as a group. Special testing and language training for Puerto Ricans were provided all through the 1980s. Just as for females with the AFWST, so the ECFA for Puerto Ricans helped reduce adverse impact and helped establish the fairness of the military aptitude testing program for Spanish-speaking people.

The AFWST and ECFA were used to help make selection decisions, but after the females and Puerto Ricans qualified for service, classification decisions were made on the basis of the standard classification batteries administered and scored in the same way for all examinees. For classification and assignment purposes, the aptitude composites were the same for all groups; training programs were also the same for all groups. The reasoning was that all groups needed to satisfy the same aptitude standards for assignment to occupational specialties.

---

[7]Chapters 2 and 4 each contain more information about the Job Performance Measurement Project.

Beginning in the Vietnam era and continuing with the AVF era, the Services became more sensitive to the impact of testing on all individuals, noting especially the impact on members of racial/ethnic subgroups and females. The concern remains that personnel decisions, both selection and classification, may not be fair for all groups. Typically, racial/ethnic minorities as a group score lower than whites on aptitude tests, in both the military and civilian areas, and females score below males on technical tests. As a result of this adverse impact, which means lower qualification rates for these minorities than for white males, aptitude tests now are receiving special scrutiny to evaluate their fairness for minorities as predictors of performance in training and on the job.

In the 1970s, the Services started to evaluate the predictive validity of the tests for blacks. Dating back to World War II, even to tests used during World War I, blacks scored lower than whites on virtually all selection and classification tests. The traditional finding is that the mean score for blacks has been about one standard deviation below that of whites. This fact by itself shows adverse impact on, but not necessarily bias against, minorities. Test fairness usually is evaluated in terms of how accurately tests predict performance of minorities compared to white males. The typical finding in the Services during the 1970s and 1980s was that the aptitude tests were equally accurate in predicting training grades for racial/ethnic minorities and whites. In fact, the results showed that the tests tended to overpredict the performance of racial/ethnic minorities. In the 1980s, when the military classification tests were validated against measures of job performance, similar results were found: blacks tended to perform less well on the job compared to whites with the same test scores (Wigdor & Green, 1991). Thus, although blacks as a group have scored below whites, and fewer of them therefore have qualifying scores, the tests are fair for them in the sense that the prediction of performance in training and on the job is about the same for racial/ethnic minorities and whites.

For females, the results are less clearcut than for blacks. The number of females in the Services was relatively small until the late 1970s and 1980s, and the number of females in most specialties is still small, which means that empirical evaluation of the fairness of military tests for females is difficult. The limited data to date indicate that the tests are accurate predictors for females because their performance in training courses and on the job tends to be commensurate with their test scores. Therefore, military aptitude tests appear to be as fair for females as for males.

Since the 1970s, sensitivity to the effects of tests on individuals, especially minorities, remains an important consideration when evaluating tests. Predictive validity, which in former times was a driving force, is still necessary for tests, but predictive validity by itself is not sufficient to justify using a test; the effects on minorities must be weighed carefully in decisions about the use of aptitude tests.

Score Scales and Qualifying Standards

A primary purpose of military aptitude tests is to help set qualifying standards. Qualifying standards imply that the scores show how well people at various levels are expected to perform in training and on the job; that is, standards are set in terms of these expected levels of performance. Standards are changed infrequently, and when they are, the levels of expected performance of people with test scores at the qualifying score should change commensurately. The setting of standards requires that aptitude scores retain their same meaning over time, as long as the standards are being used.

Qualifying standards for the Military Services are set to control two outcomes. One outcome is

to keep failure rates in training courses at acceptable levels; the other is to reflect public policy about who is entitled, or obligated, to serve the country through military service. The score scales for military tests are developed to reflect the proportion of people who meet the mental standards, wherever they are set. Standards that reflect public policy about eligibility to serve are based on proportions of the mobilization population that are accepted or rejected for serving. The score scales are defined to reflect relative standing in the population of potential recruits, and that is how they generally are interpreted.

The ASVAB Score Scales

The ASVAB score scales are expressed as either (a) percentile scores or (b) standard scores.

The percentile score scale ranges from 1 to 99, and each score point contains one percent of the 1944 (World War II) mobilization population. Percentile scores show directly the percentage of the population that scores at or below each point. For example, a percentile score of 21 means that 20 percent of the population scores below that point, and 80 percent is at or above a score of 21. The AFQT scores are now, and always have been, expressed as percentile scores. When minimum AFQT qualifying scores are set at 21, this standard keeps the bottom 20 percent of the population from qualifying.

The other scale used with the ASVAB, the standard score scale, is based on the mean and standard deviation of scores in the mobilization or reference population. During World War II, scores for the AGCT and NGCT were expressed on the standard score scale. The AGCT scale had a mean of 100 and standard deviation of 20. The NGCT scale had a mean of 50 and standard deviation of 10. Use of these scales with military classification tests, including the ASVAB, continues. Each test in the ASVAB is placed on a standard score scale with a mean of 50 and standard deviation of 10. The number of items answered correctly (the raw score) is converted to the standard score scale. Although test raw scores differ in meaning because tests differ in difficulty and number of items, use of the standard score scale places all tests on a common metric and makes the scores comparable. A score of 52 on an ASVAB test is always higher than a score of 48 on the same or any other ASVAB test. The main advantage of standard scores is that they can be added and divided and still retain their original meaning. (Adding and dividing percentile scores is a more questionable procedure, although it is sometimes done.) Standard scores can always be converted to percentile scores for the population, and they therefore also reflect relative standing in a population.

Aptitude composites of the Army and Marine Corps are expressed as standard scores with a mean of 100 and standard deviation of 20 in the population. The Air Force expresses its composites on the percentile score scale. As with the individual tests of the ASVAB, aptitude composite scores for these Services can be compared to each other. The composites are referenced to the same population, and higher scores always indicate higher ability, even in cross-service comparisons. The Navy does not use a common score scale for its aptitude composites; it adds the standard scores for each test in the composites, and this sum is used for setting classification standards. Because the sums of standard scores are not on a common scale, even if composites contain the same number of tests, direct comparisons of relative standing in the population for the different Navy composites cannot be made.

Selection Standards

Selection standards are set by military manpower managers to help control the number of people accessioned into the Services. Currently, selection standards are high because the pool of potential recruits contains large numbers of people with high aptitudes. In the 1970s, recruiting was more difficult and a larger percentage of recruits were marginally qualified. Also, when the draft was in effect during the 1950s and 1960s, a larger proportion of the recruits were in the marginal category than during the 1980s. (The marginal category is defined as AFQT scores of 10 through 30.) Prior to the 1980s, minimum standards played a major role in determining qualification for serving and assignment to specialties.

When recruiting people with high aptitude scores is relatively easy, the actual enlistment standards are high. When recruiting quality people is more difficult, as during the 1970s, the actual enlistment standards are lower. Actual enlistment standards are those prevailing at any one time; people below the actual standards may be told by recruiters that other people with better qualifications are being accepted at this time, and those with scores below the actual standards may be put into a holding status. On the other hand, formal enlistment standards are those established by regulation; actual standards can never go below the formal standards. Formal standards are changed infrequently--usually as a result of other major changes in the personnel system. During the Vietnam War, for example, formal standards were lowered to help meet the need for large numbers of recruits, and the actual standards were equal to the formal standards. Following the war, formal standards were raised, but they were not raised higher during the 1980s when recruiting quality people became easier; on the other hand, actual standards were progressively raised during the 1980s.

Selection standards are set to balance supply and demand. Demand is set by the number of new accessions needed and job requirements; supply of qualified recruits is determined by aptitude test scores, plus other criteria such as educational level. Manpower managers adjust the formal and actual selection standards to obtain the required number of recruits, while maintaining quality or mental ability, as best can be done under the circumstances. In the AVF environment, actual standards are set to meet current market conditions; formal standards help ensure that all accessions will be minimally trainable in specialties with the least cognitive demands. When the actual selection standards are below the minimum classification standards, as they were during the Vietnam War, then the Services have difficulty assigning recruits with low ability to occupational specialties.

During World War II, there were no minimum aptitude standards, although at times there were minimum literacy standards (Uhlaner, 1952, November). Even though all male recruits during World War II took the AGCT or NGCT, these tests were not used for making selection decisions. They were classification tests used to help determine initial assignments, to the extent that the information about test scores and needs of the Services could be matched in a timely manner. In 1943, the literacy standards were removed, and special training units were established. Special literacy and ability tests were developed during the war, but there is little hard data about how the tests were used in practice for making personnel decisions. At that time, the need for physically-able men was the overriding consideration, and probably not much attention was given to the precise placement of those with marginal mental ability.

The establishment of mental standards was started during World War II when classification standards were set to help identify personnel who were qualified for training to become commissioned officers. Even though relatively few people were excluded from enlisting because of low mental ability, the precedent for using aptitude tests by the Military Services for making personnel decisions

became firmly implanted. Selection standards were used after World War II, and both selection and classification standards were in use during the Korean War.

The minimum selection standard for inductees in 1950 was set at an AFQT score of 13. In July 1951, the minimum AFQT score for induction during mobilization was lowered to 10. After the Korean War, standards started edging up for both inductees and enlistees. Then, during the Vietnam War, selection standards were again lowered, and high school graduates who scored 10 or higher on the AFQT were qualified for the draft. In the AVF environment, both formal and actual selection standards have been raised, particularly the actual standards during the 1980s (Eitelberg, Laurence, & Waters, 1984, September).[8]

The quality of recruits in the AVF environment of the 1980s reached an all-time high. The mean aptitude scores and percentage of high school graduates were at their highest levels, and the percentage of recruits in the marginal category was at an all time low. The recruiting environment was able to bear such a high level, and the Services, as prudent employers, endeavored to maintain the high quality. Manpower analysts are constantly studying the supply of people and trying to figure out how to increase the supply at minimal cost. Standards are one of the tools managers use to help control supply. (During the draft era, supply was governed to a large extent by draft quotas. Inductees and draft-induced volunteers were a captive and relatively stable source of qualified manpower.)

Classification Standards

All people who are assigned to an occupational specialty must meet or exceed the classification standards for that specialty. Classification standards for specialties are defined in terms of aptitude composites; sometimes special physical standards, security clearances, or special aptitudes or education are also considered. Only the setting of aptitude or mental standards is discussed here. The setting of standards is not an exact science, and much judgment, plus tradition based on research and experience, enters into the process.

Classification standards, as any entrance standard, are designed to control failure rates in training and on the job. People whose scores are below the minimum standard are judged to have such a low probability of successful performance that they are denied admittance into the training course. The failure rates that help govern classification standards are based on academic performance in training courses for occupational specialties. Standards are set to maintain failure rates for academic reasons at about 10 percent of the student input, or less. (Other reasons for failure, such as acquiring a physical disability, are excluded for purposes of setting aptitude standards.) When the failure rate exceeds 10 percent, the training course is examined to determine the reasons and find a remedy, if possible. Some courses are so difficult, such as some electronics repair courses, that manpower and training managers decide to live with the higher failure rates. Some courses impose relatively low cognitive demands, such as infantry, and the failure rates are below five percent; minimum standards for these courses are maintained anyway because all people in all specialties should have some literacy and problem-solving ability.

Classification standards tend to remain stable, although sometimes external reasons arise to

[8]Appendix A identifies times when selection standards were changed.

change the standards. One such reason occurred in the Army during the early 1970s when a new classification battery was introduced. The main reason for introducing the new battery was that the prediction of training performance was improved over the previous battery; this resulted in reduced failure rates, other things being equal. The Army training managers directed the training schools to reduce the classification standards by one-fourth of a standard deviation, unless they could show good reason why the standards should not be lowered.

Another large scale change to classification standards occurred in 1984 when the ASVAB score scale was changed (Maier & Sims, 1986, July). When the new scale was introduced, the percentage of people in the population of potential recruits who would be qualified was lowered for many specialties compared to the qualification rates on the earlier score scale. To compensate for the lowered percentage of qualified people in the population, classification standards for many specialties, especially for the technical repair specialties, were lowered.

Changing classification standards is only one component in the management of failure rates in training courses. The easiest solution, and in the 1950s and 1960s the one usually preferred by training managers, was to raise the aptitude standard. However, this easy solution is not always the best. Since the 1970s, training managers typically have examined course content to determine if the training tasks are realistic for the job requirements. Before the 1970s, training courses as a rule were not systematically based on entry-level job requirements. The training content may have been unduly difficult with emphasis on understanding theory and abstract concepts rather than on performing concrete job tasks. In modern training management, systematic analysis of job requirements is done routinely, and the course content is established accordingly. Changing of aptitude standards may be done when the training content is realistic and the failure rates are excessive.

Manpower and personnel managers in the Services have the final approval of classification standards. As a rule, they are reluctant to raise standards for courses. Because the pool of highly qualified accessions is limited, when standards are raised for one course, the pool of qualified people for other courses is reduced. Each course should have its fair share of highly-qualified people, and manpower managers are responsible for determining the fair shares. Sometimes standards are raised by manpower managers in response to requests by training schools, and sometimes they are not.

In addition to raising classification standards, another request frequently made by training schools is to change the standards from one aptitude composite to another. To understand the basis for this request, a brief description of the theory and practice of differential classification is in order. The heart of classification decisions is that (a) different people have different aptitudes, and (b) different jobs require different aptitudes. All these differences, of course, are a matter of degree rather than qualitative differences. To the extent that differential aptitudes and job requirements are a fact, the pool of qualified manpower is expanded by measuring differential aptitudes rather than using the pool available made by using aptitude tests that provide only a single measure of ability. For example, to the extent that some people could become better clerks than mechanics, and vice versa, assigning people with relatively high clerical aptitude to clerical jobs results in high performance in these jobs with relatively little loss to the mechanical area; similarly, by assigning people with high mechanical aptitude to mechanical jobs, performance in the clerical area is not much affected. Many people, though, have similar levels of aptitude in most areas, and when they are assigned to one area, the aptitude levels in the other areas are affected commensurately. All aptitude composites from the military tests are correlated, and when people with high aptitude are assigned to one area, the pool of ability in other areas is reduced to some extent. Manpower managers attempt to take the impact on all areas into consideration when making decisions about changing classification standards.

Another consideration in setting classification standards is based on the importance of general mental ability. All specialty training courses require some level of general ability, and all aptitude composites have some component of this ability. Some composites, such as General or General Technical, are designed to measure general ability. Other composites, notably Mechanical, are designed to have a heavier loading of technical content. Over the years, training schools have found that people with high general ability are easier to train, and training managers tend to prefer students with high general ability. The importance of general ability was especially pervasive in the older style training courses that emphasized theory and concepts rather than hands-on skills. The tendency was for trainers to request that general ability be given more weight in classifying students for their courses, either by replacing an aptitude composite with one that had more loading of general ability or by using two composites to determine qualifications. Again, manpower managers evaluate these requests in terms of their impact on the total classification outcomes.

An anecdote about aptitude composites illustrates the importance of general mental ability. During the 1950s and 1960s, Army personnel managers learned to rely heavily on the General Technical composite, which contained tests of verbal and math ability. This composite emerged as the dominant score for evaluating the quality of Army recruits and even of careerists. When the Army classification system was revised in the early 1970s, the General Technical composite was dropped from the classification system. All the new composites contained some component of general ability, and no aptitude composite has emerged since then to enjoy the status previously held by the General Technical composite.

In the current AVF environment, most recruits are qualified for most occupational specialties, and by and large, minimum classification standards exert their effects only for the most difficult specialties that have qualifying scores above the mean or 50th percentile. Currently the Services are able to recruit large numbers of qualified people, and almost all recruits have aptitude scores in the average and above-average range. Very few recruits are only marginally qualified, and these tend to be assigned to the specialties that have low classification standards.

Maintaining the Integrity of Aptitude Test Scores

In former times, the primary technical evaluation of new tests relied on predictive validity and classification efficiency. Since the AVF, however, tests are being evaluated more rigorously. One set of concerns has been that recruiters might increase their numbers by coaching applicants about either test content and/or test-taking strategies . In the AVF environment, all examinees are brought or sent by recruiters to the testing site. Both recruiters and examinees have a vested interest in how high the test scores are. Recruiters have goals to meet, and examinees who fail to qualify cannot help recruiters meet their goals. The examinees are applying for employment, and their test scores help determine their qualifications for enlistment, assignment, and sometimes special bonuses. To the extent that test scores can be raised by recruiters telling examinees how to take the test, or by giving them unauthorized practice, or by using any other means, the validity of the tests is likely to suffer. If the coaching is done selectively, then coached examinees are more likely to qualify than uncoached ones. Both validity and fairness would be degraded.

The military testing community has a longstanding concern about maintaining the integrity of the scores reported for individuals. Threats to the integrity of the scores are of two kinds.

- Some potential enlistees want to raise their scores to qualify for enlistment or obtain special treatment, such as receiving a bonus; sometimes they receive the help of recruiters.

- Some potential inductees wanted to lower their scores and, thereby, evade the draft.

Two different solutions were worked out for these two threats.

Help given by recruiters to examinees became a problem immediately after the first AFQT was introduced in January 1950. It was administered by recruiters to applicants for enlistment at Recruiting Main Stations. (At the time, no one was being inducted because the Services' needs were satisfied through voluntary enlistment. Although peacetime-draft inductions were initiated in November 1948, they were continued for only a three-month period and were not re-instituted until August 1950.) Within a few months after introducing the AFQT, the test scores were found to be inflated compared to expectations based on the score distribution of the World War II population. So in April 1950, the AFQT was given at Joint Examining and Induction Stations for screening inductees, and finally in July 1951, aptitude testing of potential recruits for all Services was consolidated at examining stations. These Military Entrance Processing Stations (MEPS) were then called Armed Forces Examining Stations (AFES). The AFQT scores then returned to expected levels (Uhlaner, 1952, November).[9]

Even with the shift in testing sites, problems continued because recruiters had access to the test items, and they could easily pass on correct answers to applicants for enlistment. During the Vietnam era, the problem of recruiters coaching applicants on the AFQT was exacerbated for two reasons: (a) the same AFQT items had been in use since 1960 and their content was widely known throughout the recruiting community, and (b) the Army introduced enlistment guarantees. AFQT scores were used not only to help make selection decisions, but for those who enlisted with a guaranteed assignment, also to help make classification and assignment decisions. Getting a good AFQT score was especially advantageous for enlistees who participated in this program.

During the late 1970s, the Services were having difficulty meeting their recruiting goals, and coaching applicants on the tests was rampant. Coaching on the ASVAB became such a major problem that Congress held hearings to address the problem. One solution proposed in 1979 and 1980 was that the testing medium be changed from the paper-and-pencil mode to computers. If test items were stored in a computer, then recruiters would have more difficulty gaining access to them. In addition to storing the tests in computers, the proposal was to make the tests computer-adaptive. In computer-adaptive testing, examinees each receive sets of items tailored to their own ability level. Examinees with high ability are given more difficult items, and examinees with lower ability are given easier items. The pool of items is large, and the tendency is for each examinee to receive a different set of items. Testing of each examinee continues until the true level of ability is estimated with a prescribed degree of accuracy. For administrative convenience, a fixed number of items may be presented to each examinee, but the accuracy of estimating true scores is within the prescribed limits. In concept, computer-adaptive testing (CAT) would have largely resolved the coaching problem. Unfortunately, CAT was not ready for operational implementation at the time, and other solutions to the coaching, or as it is also known, test compromise, had to be found.[10]

---

[9]See Appendix A for further information about the establishment of examining stations.

[10]See Chapter 2 for more information about computer-adaptive testing.

Two solutions were developed. The easy one was to increase the number of test items and forms available for use at any one time. Beginning with the ASVAB forms 8, 9, and 10 introduced in October 1980, three forms of the ASVAB were used at a time, each containing two independent sets of AFQT items. (This structure of this ASVAB is described as six heads and three tails, because six sets of AFQT items and three sets of non-AFQT items are in use at a time.) Since 1980, each AFQT contains 50 verbal items, 30 arithmetic reasoning items, and since 1989, 25 mathematics knowledge items. (Until 1989 it contained 50 numerical operations items [half-weighted] instead of the mathematics knowledge items.) With six forms of the AFQT, examinees would need to learn the answers to many items if they wanted to raise their scores through cheating.

The other solution was to develop internal checks to look for discrepant scores. In 1976, the Army developed a procedure to help identify cheaters (Fischl & Ross, 1980, April). The procedure focused on the difference between the Word Knowledge and Arithmetic Reasoning test scores in the AFQT: if the Word Knowledge score for an applicant were unduly high compared to the Arithmetic Reasoning score, cheating was suspected, and the case was investigated.

By 1977, the Marine Corps had developed an improved internal check, called the Pseudo AFQT, to help detect compromise on the AFQT (Sims, 1978, April). The Pseudo AFQT used a set of non-AFQT tests to predict the AFQT score; if the AFQT score were too high, cheating was suspected. The Pseudo AFQT used more information from the ASVAB than did the Army procedure. Instead of relying on the difference between only two tests, both of which were in the AFQT, the content of the Pseudo AFQT was similar to that of the entire AFQT. The Pseudo AFQT became the standard procedure used in the 1980s to check all examinees routinely. Probability tables were provided with the Pseudo AFQT to detect not only suspect examinees, but also suspect recruiters. Examinees with unduly high AFQT scores were retested to verify the accuracy of their scores, and recruiters who had many applicants with unduly high AFQT scores were investigated. The combination of six independent AFQT forms and the Pseudo AFQT, plus the improved recruiting environment of the 1980s, served to hold cheating by applicants and recruiters in check. When cheating on the AFQT was no longer a major problem in the late 1980s, routine use of the Pseudo AFQT was suspended.

The other threat to the integrity of AFQT scores was a form of cheating to *lower* test scores. During the draft era, many people wanted to avoid serving in the military, and one way to avoid the draft was to be classified as unqualified for mental or physical reasons. Physical disabilities are harder to fake than low mental ability. Some registrants for the draft deliberately faked low AFQT scores, and techniques were developed to help identify deliberate failures. All people in category V, those with AFQT percentile scores 1 through 9, were routinely interviewed to help evaluate whether they were true or deliberate failures. Checks were made on the person's educational level and work history to look for inconsistencies. Special tests to evaluate literacy were sometimes given to the failures. If people were judged to have deliberately faked a low score, and were otherwise qualified, they were administratively inducted, and they were called Administrative Acceptees. The number of Administrative Acceptees tended to be about one or two percent of the total number of inductees.[11]

During the draft era, the workhorse procedure to help identify deliberate failures was the deliberate failure key (DFK) obtained from the AFQT. The first Army reports on detecting faking were published in the early 1950s (ARI, 1951, December; ARI, 1952, December). The procedures

[11]See Appendix A for further information about Administrative Acceptees.

for developing and using the DFK for the separate AFQT were closely held in the research community. Only with development of the joint-service ASVAB for testing all applicants did the DFK become commonly known by military testing psychologists. The DFK was used routinely during the draft era to screen failures on the AFQT. In the AVF era, a DFK is developed for the AFQT, but none have been used to date. It is available in case of mobilization; then it would be used once again to help identify people who try to evade serving by faking low mental ability.

The concept underlying the DFK is remarkably simple. For each new ASVAB form, a group of recruits is told to deliberately fake low scores on the AFQT. Their responses are then compared to responses from those who are known to be true failures (applicants for enlistment who score in category V), and item alternatives that show large differences between these two groups are included in the DFK.

Stability of the Military Aptitude Test Score Scales

In the 1960s and 1970s there was some concern in the military testing community about how accurately the World War II score scale reflected relative standing in the then current population of potential recruits. Obtaining a new reference population was expensive, and in the absence of clear problems with the existing score distributions, no efforts were made to define a new score scale. During the Vietnam War, when the draft was fairly representative of the then current youth population, the AFQT score distributions were comparable to the World War II population, and the interpretation made by researchers and manpower managers was that the World War II score scale remained appropriate for setting standards.

When the ASVAB was administered to a nationally representative sample of American youth in 1980, a comparison of the ability of the 1980 population to those tested in earlier times became possible. The nation had gone through large educational and cultural changes between World War II and 1980, and there was no clear indication about how aptitudes had changed over the decades. Comparison of the AFQT scores for the 1980 population and potential recruits during the Vietnam War with the AGCT and NGCT for the World War II population provide a reasonable basis for comparing changes in aptitude between World War II and 1980. The score distributions of males tested during World War II, the Vietnam War, and 1980 are shown in Table 3.

The sources for these percentages are Uhlaner (1952, November) for the World War II distribution and Maier and Sims (1986, July) for the Vietnam and 1980 distributions. The World War II population represents accessions in 1944. The distribution for the Vietnam period is based on 6,322,302 registrants for the draft tested in 1966 through 1970. The 1980 population comprises males of ages 18 through 23 tested in 1980.

The proportion of males in the above-average range, categories I and II, increased from 36 percent in World War II, to 39 percent in the late 1960s, and to 42 percent in 1980. The proportion of males in the below-average range, categories IV and V, decreased from 30 percent during World War II, to 27 percent during the Vietnam War, but increased back to 30 percent in 1980. The median score moved up four percentile score points between World War II and 1980. Generally, the results indicate that the population of young people in America had higher aptitude scores in 1980 than in earlier times.

Table 3

AFQT Score Distributions

AFQT Category	Percentile Score Range	Percentage of Population WWII	Vietnam	1980
I	93-100	8	8	6
II	65-92	28	31	36
III	31-64	34	34	28
IV	10-30	21	18	22
V	1-9	9	9	8

Comparing scores of people tested at different times requires stable score scales. Test items are replaced every few years, which means that the new items must be calibrated, or scaled, to the reference population. For example, the AFQT was replaced in 1953, 1956, and 1960. All these changes to test items could possibly have introduced error into the score scale. However, there was evidence that the new AFQT forms were properly calibrated because the score distributions did not change abruptly when new ones were introduced. This stability of the score scale during the 1950s and 1960s provided confidence for manpower managers that the qualifying standards were serving as intended and that the score of recruits could be taken at their face value.

The stability of the score scale for military tests was evaluated in 1980 through a study conducted by Educational Testing Service (Boldt, 1980, August). The AGCT used during World War II and form 7 of the AFQT, used in the 1960s during the Vietnam era and also used as the reference for calibrating the ASVAB in 1980, were administered to a sample of high school students. The results generally confirmed that the AFQT 7 and the AGCT were on the same score scale. At the low end of the scale, up to a percentile score of 30, the two scales were in almost perfect agreement. Above that point, the two scales diverged in the sample of high school students; the sample of students scored higher on the AGCT than on the AFQT compared to the military samples used in the original scaling of the AFQT 7 to the AGCT in the late 1950s. The two scales diverged by about four percentile points at the median, and up to seven percentile score points in the above-average range. In a re-analysis of the data for the student sample, some AFQT 7 scores were found to be aberrant from the other score distributions. When the aberrant AFQT scores were removed, the AGCT and the AFQT scales were in close agreement up to a percentile score of 50, and the differences in the top half were smaller than in the original analyses (Maier & Sims, 1986, July). The re-analysis lent further support to the stability of the score scale for military aptitude tests from 1950 until the mid 1970s.

Structure of Military Selection and Classification Testing

Prior to the AVF, each Service used its own classification battery to help assign recruits to occupational specialties, together with the joint-service AFQT. The first test given in the old procedure was the AFQT. People who met the selection standards were accepted into the Service; later they were given the classification tests to help evaluate which classification standards they met and to help determine their assignment to occupational specialties. (This testing procedure of having separate selection and classification tests given at different times is called two-stage testing.)

Military aptitude testing assumed a significant new shape in 1973 when the draft was suspended and all recruits were obtained through voluntary enlistment. The ASVAB was conceptualized as a way of simplifying the testing procedures. One-stage testing, in which all tests are given at the same time and both selection and classification decisions can be made simultaneously, was introduced to facilitate decisions by applicants and the Services. By having all information needed about qualification for all Services, an applicant could more easily shop around among the Services without going through a lot of extra effort. In addition, testing was decentralized in the early days of the AVF by the opening of nearly a thousand testing stations throughout the country; testing was also conducted at times convenient for applicants, such as during evenings and on Saturdays.

Since 1976, all applicants take the entire ASVAB in one testing session, and selection and assignment decisions are frequently made simultaneously; that is, the applicant may be enlisted for a specific occupational specialty and even be given a training class start date. This type of decision is part of the enlistment-guarantee or contract program.

Even with one-stage testing for decision making, most applicants are screened by recruiters with a paper-and-pencil Enlistment Screening Test (EST) or a Computer Adaptive Screening Test (CAST) prior to taking the ASVAB. Scores on these brief tests are used to predict the AFQT score that the person is likely to receive. The screening test scores have no official status; they are not used to make personnel decisions. They are used primarily to guide recruiters in the amount of effort they may want to expend in pursuing a person. People obviously unqualified for enlistment are not likely to be sent forward for testing with the ASVAB. The use of screening tests dates back to the late 1940s when they were used to help cut down on the expense of shipping people to examining stations who were likely to fail the aptitude tests.

In addition to issues around one-stage and two-stage testing, the Services have lived with the tension between (a) a single test score which measures general mental ability, and (b) classification tests which measure different aptitudes. The primary tests used in World War II, the AGCT and NGCT, both provided a single score that reflected general ability. The single index of ability, however, was supplemented with additional tests that covered the technical and clerical domains. Tests of mechanical aptitude and information, spatial ability, and clerical aptitude (speed and accuracy) were developed and used during the war to help make classification decisions. When the peacetime draft was instituted in 1948, the decision was made to use a single measure of general ability as the basis for determining mental qualifications for serving; this test, the AFQT, was modeled on the AGCT (Uhlaner, 1952, November).

For this overview purpose, it is important to note that military selection testing and classification testing were reorganized in the 1970s to accommodate the AVF. From the Services' point of view, a massive change in attitudes and ways of doing business was required to effect the transition from the draft to the AVF environment. The military testing community also had to go through a massive

change away from the traditional Service classification batteries and procedures to a common joint-service battery that met the needs of all the Services.

The AVF also forced other decisions and tradeoffs in the personnel system, in addition to those in aptitude testing. One big tradeoff was between quantity and quality of recruits. In the AVF, all accessions must be recruited, and recruiting is expensive. Quality recruits, those who are high school graduates with AFQT scores of 50 or better, are more expensive to recruit than people with lower test scores or nongraduates. The Services prefer quality recruits because they are more likely to finish their term of service, they are easier to train, and they are more flexible in their job assignments.

How many quality recruits can the Services afford? How many do they need to operate and maintain the new technology and weaponry introduced in the 1980s? How much training can compensate for aptitude? These are some of the questions that face manpower managers, training managers, and public policy makers in the AVF era. The issues were of course also prevalent during the draft era, but then the issue of quantity versus quality in recruits could to some extent be resolved through inducting greater numbers of people. The issue was most pronounced in the Army, as the other Services tended to have enough quality recruits, many of them induced to enlist because of the prospect of being drafted. When the Army needed more quality recruits, one solution was to increase draft quotas.

Aptitude testing is designed to remain a constant source of information about the ability of recruits throughout all changes to the personnel system. When working properly, it provides accurate information about the expected performance of recruits and helps inform managers and policy makers about the effects of their decisions and policies.

As alluded to several times in this overview, the beginning of the AVF and the development of the joint-service ASVAB were watershed times of the military testing program. Virtually everything--from writing items, through scaling new test forms, to management of the testing program--were restructured, and the professional integrity of military selection and classification testing during the 1980s conformed to high professional standards.

CHAPTER 2

MANAGING THE MILITARY SELECTION AND CLASSIFICATION TESTING PROGRAM

The military personnel system is constantly evolving in response to factors such as changing doctrines about warfare, degrees of mobilization, and public attitudes toward the military. As conditions change, manpower managers respond by changing policies, developing new initiatives, and restructuring existing procedures. In this chapter, the organizations and agencies that develop, operate, and manage the joint-service Military Testing Program (MTP) are discussed.

The MTP was born in the crisis of World War II with the origin of the large scale testing programs. During World War I, Army tests were administered to recruits, but military testing was not used much between the two wars, and the origins of the current MTP lie in World War II. The Armed Forces Qualification Test (AFQT) was developed to meet another major change in American history--the institution of the peacetime draft in 1948. The introduction of the Service classification batteries, beginning in 1949, was initiated by the research organizations as an outgrowth of their research programs, and not in response to major external events. Following World War II, military testing was a stable enterprise during the 1950s and 1960s, but in the 1970s major changes seemed to become the rule of the day. One major change in the personnel system occurred in the early 1970s as the military was switching from the draft to the All Volunteer Force (AVF). The MTP was dramatically revamped to help prepare for the mammoth change in the personnel system.

Management of the MTP has become increasingly complex since the joint-service Armed Services Vocational Aptitude Battery (ASVAB) was introduced in 1976 to facilitate the accessioning process in the AVF environment. In addition to accommodating the needs and interests of all Services, the current MTP has come under increased scrutiny and pressures from social critics of testing, the professional testing community, and political groups. The effects of military tests on subgroups are of special concern in current testing practices and management. In the face of this increased managerial and public attention, many complex technical issues, especially test fairness and validation of the tests in the joint-service arena, are now driving forces in the MTP. The attempts of the military testing community to cope with the new demands underlie the content of this chapter.

MANAGING THE DAY-TO-DAY OPERATIONS OF THE MTP

Current Management Structure

In an organization as large and complex as the Department of Defense, which includes four Services, the management structure is necessarily complex. In this part, the functions of major organizations are briefly described as they pertain to the MTP. For most of these organizations, testing is only a small part of their responsibilities, as they have numerous other functions. An exception is the Personnel Testing Division of the Defense Manpower Data Center (DMDC), also

known as the Testing Center, which is devoted exclusively to research and development of military tests.

The major organizations, and a brief description of their functions in the MTP, follows. They are all joint-service in operation or composition; the outcomes of these bodies affect all Services.

- **The Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P).** The Directorate for Accession Policy in OASD-FM&P (OASD-FM&P-AP) provides the daily management of the MTP, having approval authority for most decisions. The Assistant Secretary of Defense (ASD-FM&P) approves major actions but normally is not concerned with day-to-day decisions and operations.

- **Manpower Accession Policy Steering Committee.** The Steering Committee is composed of flag officers from each Service with responsibility for military manpower policy, plus the Commander of MEPCOM (see below), and is chaired by the Director for Accession Policy. It reviews and approves actions and decisions that affect the ASVAB and other components of the MTP. The Steering Committee is responsible for decisions concerning test administration procedures, test content, conversion tables, and requirements for troop support needed for the development of new tests. New undertakings in the MTP are reviewed and approved by the Steering Committee; controversial issues are either forwarded to OASD-FM&P for resolution or sent to subcommittees for further study.

- **The Military Entrance Processing Command (MEPCOM).** MEPCOM is responsible for processing potential recruits (to determine mental, medical, and moral qualifications), plus processing potential commissioned officers and people joining the reserve components. It has a group that performs analyses of testing data and makes technical recommendations about day-to-day administration of the ASVAB, plus approximately 65 examining stations, currently called Military Entrance Processing Stations (MEPS), and approximately 1,000 local stations called Mobile Examining Testing Stations (METS). MEPCOM is also responsible for the printing of all testing materials, including the support materials for the Student Testing Program (STP), and for investigating cases of potential compromise of the ASVAB.

- **Manpower Accession Policy Working Group (MAPWG).** MAPWG is the current designation of the old ASVAB Working Group that was organized in 1974. The MAPWG, meeting quarterly since 1979, provides a forum for joint-service review and discussion of test analyses, plans, results, and proposals that affect the MTP. The full MAPWG is made up of (a) a technical committee, and (b) a policy committee. The technical committee has responsibility for the technical adequacy of the ASVAB and other instruments projected for use in the joint-service arena. It is chaired by the Testing Center, and each Service and MEPCOM provides a representative. This committee is augmented on occasion by additional technical representatives from the Services to address specific efforts, such as recommending new tests to try out for potential inclusion in the ASVAB. Separately, the policy committee reviews how the tests are used in making personnel decisions; the Service representatives are from the military manpower policy offices, plus MEPCOM. The policy committee is chaired by a Service representative. The MAPWG makes recommendations to the Steering Committee about the development of the ASVAB, and normally the Steering Committee approves unless the issues are controversial. (The Coast Guard also has a representative on the MAPWG because it uses the ASVAB for making personnel decisions for its enlisted people.)

- **The Testing Center**. The Testing Center, or Personnel Testing Division of DMDC, was formed on October 1, 1989, to assume the responsibilities of executive agent for research and development of the ASVAB (and potentially for other testing instruments that may be used in the joint-service arena for making personnel decisions about recruits). The Testing Center has responsibility for test production, analyses of new and existing tests, preparation of supporting materials, and research on psychometric procedures.

- **Defense Advisory Committee on Military Personnel Testing (DAC)**. Although members of the DAC are not employees of the Department of Defense, and the DAC has no operational authority, it exercises considerable control over the MTP through its review of all aspects of the MTP. The DAC focuses on technical issues, but it is also sensitive to policy questions that surround the technical issues. The Steering Committee is reluctant to approve recommendations from the MAPWG about technical matters unless the DAC has previously reviewed and approved the matter. The DAC has met about three times a year since its formation in 1981.

In addition to these joint-service organizations, each Service has its own organizations with responsibility for Service-specific testing research and development, operational use, and policy decisions. Historically, until 1976, each Service was responsible for its own testing program. Currently, each Service retains control over how to use the ASVAB results for classification. The Services also have specific tests that may be given at the MEPS to applicants for that Service prior to enlistment or, at recruit centers following enlistment. The following Service agencies are directly involved with aptitude testing.

- **Research Facilities**. The Services have the following laboratories or capabilities to perform research and development of testing:
  Army Research Institute for the Behavioral and Social Sciences (ARI)
  Navy Personnel Research and Development Center (NPRDC)
  Armstrong Laboratory (AL), formerly called the Air Force Human Resources Laboratory (AFHRL)
  Center for Naval Analyses (CNA), performing research on the ASVAB for the Marine Corps since the mid 1970s

  The focus of attention at these agencies includes (a) validation of the existing ASVAB to determine the aptitude composites for classifying recruits, and (b) analyses to set qualification standards. The laboratories also conduct research to develop and analyze new tests that may be used in Service-specific testing or in the joint-service arena. Representatives from these research facilities serve on the MAPWG technical committee. Other responsibilities include recruiting, training, and other components of the personnel system, such as the effectiveness of enlistment incentives, welfare of dependents, training methods, and human factors. Zeidner and Drucker (1988) have prepared a comprehensive history of ARI that describes the scope of research activities in the Army and briefly in the other Services.

- **Military Manpower Policy Offices**. Each Service has an agency concerned with setting policy on manpower and personnel issues such as helping set mental standards for selection and classification of recruits, for re-enlistment, and for promotion. A representative from the policy office of each Service is a member of the MAPWG policy committee.

The above groups, joint-service and Service-specific, constitute the management of the MTP. Since the first joint-service ASVAB was introduced in 1976 for selecting and classifying military

recruits, these groups have learned how to work together toward the common goal of maintaining and improving the MTP. As a rule, before any action is taken, decisions and proposed actions about testing operations and policy are reviewed in the joint-service arena to make sure that the interests of all Services are met. Exceptions to joint-service review include decisions about the definitions of aptitude composites, supplementary testing of applicants or recruits, classification standards, and research on new measures; each Service makes its own decisions about these matters.

Previous Management Structures

A joint-service committee was formed in 1948 to design and oversee the first AFQT. After the AFQT was in operation and test administration became centralized at examining stations, the joint-service involvement was ad hoc. The Army, as executive agent for the AFQT and the operation of the examining stations, had primary responsibility for providing research and development, maintaining the score scale, and administering the AFQT to registrants for the draft and applicants for enlistment for all Services. The Army activities necessary to carry out these responsibilities were reviewed by the Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P).[12]

After the development and implementation of the AFQT, the next major joint-service venture that resulted in new testing operations was the design and development of the first ASVAB for use in the Department of Defense Student Testing Program (STP).[13] Again, a joint-service committee was established for this purpose (with the Army serving once again as executive agent), and it was also disbanded after the first form (ASVAB 1) was introduced in 1968. Even in the early days of the ASVAB Working Group, from 1974 until 1979, meetings were held as needed rather than on a regular basis. Similarly, the Steering Committee did not meet regularly until 1980.

When the Air Force became executive agent for the ASVAB in 1972, its joint-service testing activities were reviewed by OASD-FM&P. Only with the development of forms 5, 6, and 7 of the ASVAB (ASVAB 5/6/7), starting in 1974, did joint-service involvement become effective at a detailed technical level on a continuing basis, addressing issues such as deciding on the distribution of item difficulties, performing detailed review of the items, and reviewing preliminary results. The working relationships in the early years of joint-service participation on the ASVAB were sometimes marked by dissension instead of harmonious work toward the common goals of a common testing program. The participants needed time to learn how to work effectively together.

The formation of the Testing Center on October 1, 1989 was several years in the making, and the reasons for its formation extend back to the 1950s and 1960s. The focal issue around forming the Testing Center was the level of support provided for aptitude testing over the years. As with most things in the MTP, the starting point was during World War II, and so it is for the location or home of military aptitude testing.

During World War II, the Army, Navy, and Army Air Force each had a separate research

---

[12]The title of the OASD-FM&P has changed several times since its inception; OASD-FM&P is used in the document rather than tracing the name used at the time when a reference is made to it.

[13]Chapter 3 of this report provides information about use of the ASVAB in the Student Testing Program.

organization to develop tests for selecting and classifying people. The research organizations were part of the personnel system, funded and sponsored by the Service personnel offices. The Army facility was part of the Adjutant General's Office, and the Navy's part of the Bureau of Personnel (BUPERS). The research facilities were large during the war, and in the 1950s some still had over 100 people on the staff, with the bulk of their efforts devoted to test development and validation.

During the 1950s, the research organizations started efforts in other areas. As classification and assignments became formalized, concerns grew about the effectiveness of these procedures, which in turn led to interest in what people do on the job and how more efficient work procedures could be developed. Research on human factors and training were other areas that grew up in the research organizations.

Changes introduced by manpower managers that affected test administration procedures and qualification standards were, of course, also made during the past 50 years. Starting in the 1950s and continuing until MEPCOM was formed in 1976, the US Army Recruiting Command (USAREC) was in charge of operating the examining stations, which tested and processed applicants for the Services and all for induction. USAREC prepared an annual report, called the "Qualitative Distribution of Military Manpower Program" (US Army Recruiting Command, 1972), that contained statistics on all the people processed during a fiscal year. One of the interesting sections of the report from the perspective of this report is a listing of historical notes.[14]

Army as Executive Agent for the AFQT. When the AFQT was formulated, the initial intent was to replace the existing forms with new ones every three years. That goal was met only the first time, when forms 3 and 4 were introduced on January 1, 1953, exactly three years after forms 1 and 2. Three and one-half years later, forms 5 and 6 appeared on August 1, 1956; forms 7 and 8 were ready on July 1, 1960, almost four years later (Bayroff, 1963, May). Forms 9 and 10 never did appear, although by 1973, some 13 years later, new items had been written for them. Item writing in the early 1970s was done under contract, but before the items could be analyzed, use of the separate AFQT was made optional and all development activities ceased.

Forms 1 through 8 were prepared by in-house staff of the Army laboratory. In the early days of the AFQT, people were available to write new items, try them out, and prepare the test booklets and supporting materials. By the mid 1960s, the number of people working directly on the AFQT was 4 or 5, including managerial and support staff. In 1966, the AFQT research staff was diverted to work on the ASVAB.

From 1966 through 1972, the existence of only two operational forms of the AFQT became a problem and embarrassment. The problem arose from coaching of applicants on the AFQT during the Vietnam buildup; because of the limited number of items in use, recruiters could readily learn their content and coach examinees on the test. The embarrassment arose because there was no adequate response by the research community to pleas from personnel managers for new items that would help reduce coaching.

In the early 1960s, the funding for some testing research was shifted from the personnel communities to Research and Development (R&D) funds. The Army, for example, started funding the Army Research Institute for the Behavioral and Social Sciences (ARI) with R&D funds in

[14]The set of historical notes in the final report, for fiscal year 1972, is included in this report as Appendix A.

December 1961 (Zeidner and Drucker, 1988). The organizations in the Army, Navy, and Air Force became laboratories or centers, just as they were called in the military hardware arena.

From an R&D funding perspective, much of aptitude testing could hardly be justified as research. The development of new test forms that were parallel to existing ones seemed more like a maintenance activity than an R&D activity. A justification offered for R&D funding was that only trained researchers could adequately develop new forms of tests, and therefore the activity belonged in a research facility and could be funded as R&D. The AFQT, which remained unchanged from 1953 until 1973, was a good case in point about the use of R&D funds for an essentially maintenance activity. The Service laboratories during this time started expanding the scope of their research areas to efforts that more clearly qualified for R&D funding, and over the years, as staff migrated out of testing into other areas, departing members frequently were not replaced, and funding for testing tended to shrink. Military aptitude testing fell on hard times.

Air Force as Executive Agent for the ASVAB. In 1972, when the Air Force replaced the Army as executive agent for the ASVAB, the OASD-FM&P expectation was that the program would be adequately staffed to produce new tests in a timely manner. New forms did appear on a reasonably regular schedule: forms 8, 9, and 10 appeared on October 1, 1980; forms 11, 12, and 13 on October 1, 1984; and forms 15, 16, and 17 on January 1, 1989. The number of researchers working directly on the ASVAB, however, was not adequate to produce new forms on schedule and also carry out the other duties required to maintain the joint-service and student testing programs. The number of people whose primary responsibility was research and development of the ASVAB varied from 3 to 6, about the same as for the AFQT in its declining years.

In 1979, when the miscalibration of the ASVAB became apparent, the inadequacy of funding and staffing for the ASVAB became clear to OASD-FM&P.[15] An immediate action was to provide $1 million of Operations and Maintenance (O&M) funds to the Air Force to supplement the existing R&D funds. The managers recognized that much of the ASVAB effort was maintenance of the existing battery and the R&D funds should be used to support R&D activities. The O&M funds were provided annually to the ASVAB program until 1989.

The extra money provided to the Air Force for the ASVAB was not used to increase in-house staffing; about half was sent to the other Services each year for ASVAB-related R&D, and most of the remainder was used to support contracts managed by the Air Force. The contracts were used to help prepare new forms, which enabled reasonable timely introduction of new items.

Personnel Testing Center as Executive Agent for the ASVAB

A major event that led to the formation of the Testing Center was a recommendation by the DAC in 1986 that the research and development activities for the MTP be centralized to improve efficiency and effectiveness. The Air Force, as executive agent for the ASVAB, had produced new test forms, but had limited resources for other activities, such as the DoD STP or research on equating methodology. A separate executive agent, the Navy, had responsibility for research on computer-adaptive testing (CAT). In the 1980s, research on a self-description instrument to replace educational credentials in selection standards was intensified, and again the Navy was designated executive agent

[15]The ASVAB miscalibration episode is described in Chapter 4 of this report.

by OASD-FM&P. A third large effort that impinged on the ASVAB, but was carried on outside of the MAPWG, was the Job Performance Measurement (JPM) project. The Air Force, as executive agent for the ASVAB, and the MAPWG were not directly involved in this project, although it had clear implications for selection and classification.[16]

In its 1986 report, the DAC strongly urged the Department of Defense to study the feasibility of centralizing the activities necessary for the ASVAB and CAT programs, and expanding the numbers of centralized psychometric researchers and research managers. They recommended that the central facility have personnel dedicated to the day-to-day responsibility for tasks such as test development and analysis and compliance with professional standards. The expected outcome would be more efficient and responsive testing programs. CAT may have been singled out by the DAC because at that time it was thought to be about ready for implementation and, therefore, could be integrated into the operational testing program. Although the JPM project and self-description instrument were not mentioned by the DAC in their recommendation for inclusion in the centralized facility, probably because they were not ready for operational use at that time, it was clear that they would become candidates for inclusion when they were ready for operational use.

In response to DAC's strong urging, a task group of the MAPWG was convened in April 1987 by OASD-FM&P-AP to consider how the management and conduct of the joint-service testing program could be improved. The task group, with representatives from all Services and MEPCOM, submitted a report to the Chair of the MAPWG in May 1987, who in turn reported to the Director for Accession Policy. The thrust of the report was that the functions of a testing program -- quality control and analyses, test production, and development of psychometric procedures -- should be fulfilled by a dedicated staff with appropriate training and experience. The report of the task group said that the projected staff should consist of 35 people, of whom 20 would be devoted to test production, 8 to psychometric development, and 7 to quality control and analyses. In addition, $2 million should be available for contract support, especially in support of the STP, for a total budget of about $6 million.

The report of the task group was not received well in either the MAPWG or the Services because the logical outcome of the recommendations was the establishment of a testing center housed in the Department of Defense. The concentration of these functions in one facility under the direct control of OASD-FM&P was seen as a threat to the traditional functions and prerogatives of the Services and MEPCOM. Agreements reached in the deliberations of the task group about centralizing the functions of the MTP dissipated quickly when the members returned to their parent organizations, and the report as submitted in May 1987 contained the following disclaimer: "The chair has sole responsibility for the content of this report. The other members of the task group made vital contributions."

No public actions were taken, nor was there any public discussion of the DAC or task group recommendations until fall 1988. Then in October 1988, a series of memoranda was initiated in OASD-FM&P that culminated on November 21, 1988, with a memorandum from the ASD-FM&P to the Assistant Secretaries of the Services for Manpower. This memorandum directed that the Testing Center be established effective October 1, 1989, and that it be housed in DMDC. In fact, the Testing Center was up and running as of October 1, 1989.

---

[16]The JPM project is discussed more fully in the next part of this chapter.

The Testing Center currently is of more modest proportions than envisioned by the task group. The number of staff is closer to 20 than 35, and the current budget is closer to $4 million than $6 million.

The staff of the Testing Center has more diverse training and experience than was typically found in the Service research laboratories. The Service laboratories traditionally hired testing psychologists who tended to have backgrounds in industrial and employment testing. They were supported at first by a large staff of statistical clerks, who later were replaced by computer specialists. In the past, item writing and editing were performed by psychologists rather than specialists in these areas, and in the years since the 1970s, item writing was contracted out. Preparation of testing materials, such as test booklets and manuals, was performed by testing psychologists, usually as an additional duty. In the Testing Center, the staff includes editors experienced in test production and quality control of testing materials. The staff also includes people trained in education and vocational counseling. The mainstay of the staff is expertise in psychometric methods.

The primary function of the Testing Center is to develop and maintain the ASVAB and related materials, and the staff is dedicated to fulfilling this function. The parent organization of the Testing Center, DMDC, maintains personnel data on all members of the Services. In addition to being a data repository, it performs special analyses on request of OASD-FM&P and the Services, and it has a research capability to conduct surveys of Service members and civilians. The management of DMDC has experience in directing the mixture of research and operational activities of the Testing Center.

The formation of DMDC in the early 1970s is instructive to help understand Service reactions to proposals about forming a centralized testing center under OASD-FM&P. The original title of DMDC was Manpower Research and Data Analysis Center (MARDAC). The key word in the title was analysis. The Services were reluctant at that time, as they still are, to have outside agencies perform analyses of their people. The Services by and large want to be in control of how the analyses are conducted and reported and of the conclusions drawn from the results. The prospect of an organization named MARDAC was troublesome; the title was changed to DMDC, with the analytic capability controlled, and DMDC has been thriving since.

The Testing Center as initially organized was set up to perform the functions traditionally served by the executive agent for research and development of testing. These include preparation of new forms parallel to existing versions, scaling the new forms to the existing score scale, and preparing all testing materials. Procedures to collect data for the new forms on joint-service samples of recruits and applicants have been well established, begun in 1974 and refined during the 1980s, by the executive agent and the Services. Technical details of scaling and equating are still being worked out, but in substance the procedures are adequate to maintain the accuracy of the score scale. In these traditional and basic functions, the Testing Center executes established procedures with little controversy or concern among the Services.

MANAGING MAJOR CHANGES TO THE MTP

In this part of the report, major events in the MTP are recounted from the perspective of how they were managed and how the technical community responded.

Inflation of AFQT Scores

In 1948, the AFQT was developed by a joint-service committee under the general direction of OASD-FM&P. Following the testing practices after World War II, the AFQT initially was administered by Service recruiters. The AFQT score distributions for the first few months of use were examined by researchers and managers, and the scores were found to be excessively high. The solution was to remove the AFQT from the hands of the recruiters and centralize testing at examining stations; that practice was established during World War II and has continued ever since. This decision was made at the direction of OASD-FM&P (Uhlaner, 1952, November). The inflation of AFQT scores found in 1950 was the first recorded crisis in the MTP that resulted in intervention at the highest levels, and use of the AFQT had just gotten started. In hindsight, with decades of experience in administering tests to applicants for enlistment, the outcome of allowing recruiters to evaluate mental qualifications is obvious, but in 1948 and 1949 recruiters routinely administered aptitude tests to applicants, and the procedures for the AFQT were just following existing practice. Recruiters could provide the ultimate in individualized testing; they could administer tests at any time and place that was mutually convenient. Unfortunately, with those testing arrangements, they could also arrange for the examinees to get any desired score. The response of the manpower managers to this first crisis in the MTP was swift, decisive, and effective.

Addition and Deletion of Tool Knowledge Items

Army documentation is silent about how the decision was reached and approved to add Tool Knowledge items, presented in a pictorial format, to the AFQT in 1953. The reason was to reduce the correlation of AFQT with years of education, but the process was not discussed (Bayroff, 1963, May). The next change to AFQT content was discussed 16 years later, in 1969, when the decision was made to delete the Tool Knowledge items from the AFQT because many functional illiterates could qualify for service on the nonverbal portions of the AFQT (Spatial Perception and Tool Knowledge items); these people subsequently had high failure rates in their training courses. In addition, the predictive validity of the Tool Knowledge items was low. Eventually both the Tool Knowledge and Spatial Perception items were deleted from the ASVAB.

The decision at the technical level to delete the Tool Knowledge items from the AFQT was reached through an ad hoc meeting held at the Army research laboratory in August 1969. The technical representatives of the other Services were in the Washington DC area for a convention, and they met to discuss the content of the AFQT. They quickly agreed to delete the Tool Knowledge items, and the committee disbanded. No action, of course, was taken until 1973 when forms 7 and 8 of the AFQT were no longer used by all Services. How the recommendation of the joint-service committee was staffed through OASD-FM&P was not part of the technical documentation, nor was it an issue that caused much discussion in the research community.

Supplementary Testing

The next major change in the MTP occurred in 1958 when Congress authorized supplementary testing of people with marginally qualifying AFQT scores, those with AFQT scores in category IV (percentile scores 10 through 30) (Bayroff, 1963, May). Also in 1958, the Air Force introduced the Airman Qualifying Examination (AQE) for testing Air Force applicants at recruiting stations to help make selection and classification decisions (Weeks, Mullins, & Vitola, 1975). The introduction of supplementary testing was initiated by the Department of Defense to improve personnel decisions, and was not in response to a crisis or other major change. Although the Service reports make no mention of OASD-FM&P's involvement in initiating supplementary testing, a change of this magnitude could not have been made without approval by the OASD-FM&P. With supplementary testing, Army recruits in the marginal category had to qualify on two aptitude composites with scores of 90 or better on the Army standard score scale (one-half of a standard deviation below the mean or higher). The first supplementary testing of registrants for induction and applicants for enlistment in the Army was with the Army Classification Battery (ACB), which at that time required over four hours of testing time. Thus people with marginally qualifying AFQT scores had to take well over five hours of testing: the AFQT plus the ACB. The burden on examinees and examining stations must have been overwhelming.

In September 1961, the Army introduced the Army Qualification Battery (AQB), an abbreviated version of the ACB, which incorporated the four parts of the AFQT and other tests from the ACB (Bayroff, 1963). The AQB tests contained 25 items each, as did each of the four parts of the AFQT. The total testing time for the AQB was about three hours. In the 1960s, all Services administered a shortened version of the classification batteries at examining stations. Use of the AQB came into its own during the Vietnam buildup, when the Army used it to make personnel decisions for enlistees. For enlistees with guaranteed assignment to specific specialties, subsequent testing at recruit centers with the ACB was redundant, although it still was done until 1973. Supplementary testing at examining stations became the precursor to the conduct of all initial testing at examining stations.

The ASVAB for High School Students

The first ASVAB, initiated in 1966 for use in testing high school students, was in response to a minor crisis generated by educators who complained about the burden of being approached by recruiters from all Services offering to administer their Service test batteries. OASD-FM&P restored order to testing in high schools by directing that only one military test be given at the schools. As with the AFQT, a joint-service committee was established to develop the first ASVAB. The committee designed the research effort, reviewed the results, and agreed on the tests to be put into the ASVAB. Also, as with the AFQT, when the initial work was done, the committee disbanded, and the executive agent generally worked alone in developing new forms.

The Watershed Time for Aptitude Testing in the 1970s

In 1973, use of the common AFQT was made optional by the ASD-FM&P, and the Services could then, if they chose, derive an AFQT score from their classification batteries. The Army had introduced a new classification battery in May 1973, and it then started obtaining an AFQT score from its new classification battery. The Air Force switched to a version of the ASVAB, form 3, in September 1973, and the Marine Corps switched to that same test in July 1974; both Services began

to derive an AFQT score from this battery. The Navy continued to use its Basic Test Battery, and it too derived an AFQT score from its battery. After 1974, the separate AFQT, which had been in use since January 1950, was no longer administered.

The intent of the manpower managers when allowing the Services to obtain an AFQT score from these different test batteries was that they all should be on the same score scale. In separate development efforts, each was linked to the AFQT, so ideally the scores should have conveyed the same meaning in terms of expected performance and relative standing in the population. However, none of them was directly linked to another, and except for the Navy, there was no evidence that the score scales did have the same meaning. (The Navy during this period conducted a study to evaluate the similarity of the scales for the AFQT scores obtained from the Navy Basic Test Battery and the separate AFQT used in the 1970s; the results showed that the distributions of the AFQT scores were similar, and therefore the score scales were comparable.)

In addition to lingering questions about a common meaning to the score scales, the multiple test batteries were logistically cumbersome. Beginning in 1973, all classification testing was moved from recruit reception or training centers to examining stations. Instead of one hour of testing for the AFQT, three or more hours were required for each classification battery. The test administrators had to give the different batteries to applicants for enlistment in the different Services. From a joint-service and OASD-FM&P perspective, a major problem with the multiple batteries in the early AVF era was that applicants could not easily shop around among the Services. To determine eligibility for enlistment in different Services, an applicant would have had to take the tests specific to each Service. The concern on the part of manpower managers was that some applicants would become discouraged by all this testing and walk away. In a word, military selection and classification testing and personnel decisions in the mid-1970s were chaotic.

It was during this period that the ASVAB was conceptualized as the joint-service selection and classification battery that would restore order to the accessioning process. It was to be given at examining stations to all applicants for all Services, and the same set of scores were to be used to determine eligibility for any Service.

The All-Volunteer Force

Beginning with the AVF, management of military aptitude testing by OASD-FM&P became visible and direct. Previously, the research staffs working on the AFQT and all Service batteries went about their business without much concern or awareness of what was transpiring at the OASD-FM&P level. The researchers were working on technical problems, and the policy issues by and large were not their immediate concern. The Service classification batteries used from 1973 until 1976 also were not under the daily scrutiny or management of OASD-FM&P. Only with development of the joint-service ASVAB for testing military applicants for enlistment (forms 5/6/7) did the technical and policy communities, at all levels, become intermingled; the close working relationships have remained, and even intensified, over the years.

In 1974 and 1975, OASD-FM&P prodded the ASVAB Working Group to get ASVAB 5/6/7 into place as soon as possible. The technical community was not accustomed to this kind of pressure and scrutiny. In earlier times schedules were taken more lightly by researchers; the prevailing attitude was that new tests would be introduced when they were ready. However, the Working Group did respond to get ASVAB 5/6/7 introduced by the modified deadline of January 1, 1976, after the initial

date set by OASD-FM&P of September 1, 1975 could not be met.

The ASVAB Miscalibration

With the ASVAB miscalibration, the attitudes and behavior of the technical community began to change, although not immediately. As explained in Chapter 4 of this report, the ASVAB miscalibration episode unfolded slowly, except for the fix made at the upper end of the score scale in the summer of 1976. Yet as early as June 1976, information was available that the low end was also inflated. On June 11, 1976, the Air Force as executive agent sent a memo to the Army presenting alternative AFQT conversion tables for ASVAB 5/6/7. This information did not lead to changes to the conversion tables or to further data collection by the Working Group to determine their accuracy. Neither did the policy committee nor manpower managers at all levels become concerned about possible errors in personnel decisions at the low end of the score scale. Finally in 1979, the Steering Committee did become engaged, as did OASD-FM&P, and corrective action was quickly taken to fix the score scale.

A major outcome of the ASVAB miscalibration was that management supervision became intensified. The Working Group, the Steering Committee, and the DAC all started meeting regularly, and they scrutinized the MTP in detail. (Since then, all major actions have required prior approval by these bodies before being started.) During the 1980s, the executive agent routinely started providing data on the tryout and scaling of new forms to all Services. The Services used these data to conduct their own analyses to check on the results presented by the executive agent.

The 1980 ASVAB Score Scale

A crash program was initiated in 1980 to get form 8a of the ASVAB ready for administering to a national sample of American youth in the summer of 1980 for the purpose of constructing a new ASVAB score scale. The fix to the ASVAB score scale in 1980 was timely, and the outcome has been increased rigor in scaling the ASVAB. The urgency in defining a new reference population in 1979-80 was real, and the response had to be swift or else the opportunity to test an existing sample of American youth would have passed away. The study was managed at the OASD-FM&P level, and the MAPWG was kept informed.[17]

The ASVAB Content Changes

An area of great silence in the public documentation of the ASVAB is how test content has been determined since 1974. The preliminary plan for the ASVAB 5/6/7 prepared by the first ASVAB Working Group called for tests of 25 items each, or more for some tests, plus 527 interest items. The testing time would have been excessive, far exceeding four hours, and it was pared back to about three hours (Wiskoff, Zimmerman, DuBois, Borman, & Park, 1991, December). The ASVAB 8/9/10 did have 25 items in most tests, no interest measure, several new tests, and not all of the old ones. The documentation of how the changes were made is sparse.

---

[17]Chapter 4 provides information about the development of the 1980 ASVAB score scale.

Decisions about paring back the preliminary plan for the ASVAB 5/6/7 are easy to understand. Because of the three-hour limit on total testing time and the decision that all Services should be able to compute their then existing aptitude composites, the ACB 73 became the model to follow. The ACB 73 had 20 items in most tests and 87 interest items. The final forms of the ASVAB 6 and 7 also had 20 items in most tests (except for Word Knowledge and Electronics Information, each of which had 30 items) and the same 87 interest items. (The ASVAB 5, used in the STP, did not contain any interest items because some people thought some of the interest items were too military in content and were inappropriate for use with high school students.) However, the resulting AFQT for ASVAB 5/6/7 consisted of only 70 items (30 for Word Knowledge and 20 each for Arithmetic Reasoning and Space Perception). Such a short AFQT was troublesome because of marginal reliability and ease of compromising.

Test content was adjusted in the late 1970s when the ASVAB 8/9/10 was formulated. One change was to lengthen the test by increasing the number of items in most tests to 25 (conforming to the preliminary plan for the ASVAB 5/6/7). Another consideration for lengthening the test was reliability of aptitude composites. The Navy at that time, and still currently, had some composites with only two tests. Composites with only 40 items seemed too short, whereas 50 items appeared a more reasonable number for making personnel decisions. The total time was further increased through the addition of the Paragraph Comprehension (PC) test, which added 12 minutes. If the tests were to be lengthened and testing time remain about the same, or preferably even be shortened, then some tests had to be deleted.

One easy decision by the Working Group was to delete the interest inventory because of low predictive validity. When the interest items were used for making personnel decisions, whatever validity they demonstrated in the research studies tended to dissipate. Applicants and recruiters could easily discern the correct answers and thereby increase their scores. In follow-up validation studies, the interest measures were found to have lost much of their validity. The predictive validity of the four scores obtained from the interest inventory, corrected for range restriction, were as shown in Table 4.

Table 4

Predictive Validity of Interest Measures in the ASVAB 6/7

Interest Score	Clerical	Electronics	Mechanical	Combat
Original	.41	.44	.34	.28
Follow-up	.26	.24	.20	.20

The original validity coefficients were based on samples of Army recruits tested in the late 1960s as part of a research effort to evaluate an experimental classification battery (Maier & Fuchs, 1972, September). Based on these results, the interest inventory was incorporated into the ACB 73, and subsequently into the ASVAB 6/7, introduced in January 1976. The follow-up results were obtained on samples of Marine recruits tested with the ASVAB 6/7 who started their training courses in 1977 and 1978 (Sims & Hiatt, 1981, February). In both studies, the sample sizes were large and the number of training courses included for each validity coefficient was about 5 or 6.

Most of the other changes to the ASVAB were easy to understand. Deleting the General Information test was easy, because it was used only by the Army and the test content overlapped verbal and technical content. The Attention to Detail test, which involved discriminating between the letters *c* and *o*, also was easy to delete; accurate printing of the items was virtually impossible. Replacing Attention to Detail with Coding Speed was a natural because Coding Speed had been in the Service classification batteries since World War II. For some reason the Electronics Information test had 30 items in the ASVAB 5/6/7, but the number was cut back to 20 in ASVAB 8/9/10. Electronics Information did have marginal validity for predicting performance in high-level electronics training courses, but it had good validity for mechanical repair specialties.

The change in content from the ASVAB 5/6/7 to the ASVAB 8/9/10 that raised the most questions was the deletion of the Space Perception (SP) test. SP was part of the separate AFQT from 1950 on, as it was of the AGCT, and it was part of the AFQT for the ASVAB 5/6/7. One reason for deleting SP from the AFQT was its adverse impact on females. Another reason it was dropped was that it had low mean incremental validity across all specialties and Services. As a general predictor, similar to verbal and mathematics abilities, it added little to predictive validity, and therefore could reasonably have been deleted from the AFQT score. The differential validity of SP across the range of occupational specialties, however, was not systematically evaluated at the time of the decision to drop it, and had it been, perhaps it might have been retained in the battery. But at the time, it had no active supporters, and it was deleted from the ASVAB.

With the deletion of SP from the ASVAB 8/9/10, a new definition of the AFQT had to be formulated. Two tests were retained because of tradition and high general validity: Word Knowledge (WK) and Arithmetic Reasoning (AR). However, both were lengthened to help control coaching on the AFQT, which was rampant in the late 1970s: WK was lengthened from 30 items in the ASVAB 5/6/7 to 35 items in the ASVAB 8/9/10, and AR was lengthened from 20 to 30 items. The new PC test was added to improve the measurement of literacy and to help control coaching. A fourth test, Numerical Operations (NO), was also added to help alleviate coaching. What no one foresaw at the time, and what no one could anticipate because of the limited experience with using speeded tests for selection decisions, were the other problems attendant to speeded tests, described in Chapter 4 of this report.

The AFQT for the ASVAB 8/9/10 was computed as the number of items answered correctly, as it had been since the beginning of the AFQT, although at times a correction for guessing was applied. However, NO had a larger standard deviation than the other tests in the AFQT, so the NO score was divided by 2 to make the standard deviation comparable, or slightly smaller, than the other tests in the AFQT. This definition of AFQT (WK + PC + AR + NO/2) remained in place until January 1989, when NO was replaced by the Mathematics Knowledge (MK) test.

The reason for adding MK to the AFQT was that, of the remaining tests of the ASVAB, it had the smallest adverse impact on females. Coding Speed could not be used in the AFQT for the same reasons that NO was dropped. Traditionally, emphasis had been directed toward keeping the AFQT free of educational achievement tests, (one reason for the addition of the Tool Knowledge items in 1953), but since MK was one of the best predictors in the ASVAB for the full range of specialties, it was included in the AFQT, and the accuracy of selection decisions increased.

The content of the ASVAB has remained unchanged since October 1980, when forms 8, 9, and 10 were introduced. The change to the definition of the AFQT in January 1989 did not involve any substantive changes to the ASVAB. The Services did complete validation studies of the ASVAB in

the 1980s, and some of them redefined some of their aptitude composites. After the big changes to the content of the ASVAB in the 1970s, the decade of the 1980s was marked by stability.

MANAGING TEST-VALIDATION EFFORTS

The Job Performance Measurement Project

The JPM Project was not fully integrated into the MTP; it was viewed by the MAPWG as peripheral to the mainstream efforts to develop and calibrate the ASVAB. The MAPWG involvement with the JPM Project was largely limited to some overlap among working group members, especially among the Service policy representatives, and occasional information exchanges. In the same vein, the JPM Working Group as a body did not view its charter as anything beyond validating the existing ASVAB. Although individual Services did try out new tests as predictors of job performance, the efforts were uncoordinated and in most cases only some of the specialties were used for validating the new predictors. Currently, after about 10 years of research to validate the ASVAB and new predictors, some validity data for new predictors are available, but aside from the existing ASVAB, no new predictor was evaluated by all Services on all occupational specialties included in the JPM Project.

The significance of not systematically trying out new predictors is becoming apparent to the MAPWG and manpower managers in the 1990s. The efforts underway to change the content of the ASVAB in the early 1990s could have benefitted from comprehensive validity data on new predictors against the measures of hands-on job performance tests. Some of the tests developed by some of the Services in the JPM Project may have promise for inclusion in the ASVAB. In the absence of validity data from its own recruits, Services are reluctant to agree to changes that would affect them.

The JPM Project was not designed to provide joint-service validity data on new predictors. In hindsight, the inclusion of such a purpose in the design would have greatly enhanced the value of the project. As it is, much is now known about the predictive validity of existing ASVAB tests in the joint-service arena, but very little about new predictors.

Joint-Service Evaluation of Test Fairness

About the time that the Testing Center was formed, the General Accounting Office (GAO) issued a draft report on the effectiveness of military training programs; the report also covered the validity of the ASVAB for predicting the performance of females and racial/ethnic minorities in training courses for technical specialties. Based on preliminary and somewhat limited analyses, they concluded that (a) the ASVAB was not as valid for females as for males, and (b) there was adverse impact for females and racial/ethnic minorities. As usual, the draft report was sent to the Department of Defense for comment prior to final publication. The ASVAB testing staff criticized aspects of the GAO study, noting the small sample sizes, but the issue of test fairness was accepted as an issue for further research and evaluation. OASD-FM&P promised that additional analyses based on adequate samples would be performed and that efforts would be continued to find predictors with less adverse impact. The Testing Center had the lead in conducting the additional analyses and in revising the

ASVAB.

The analyses to determine the validity of the ASVAB for females and racial/ethnic minorities involved the Services providing performance data to the Testing Center for analyses. The Services reviewed the analyses plan and the results. All Services cooperated in carrying out the validation effort, except the Marine Corps, which carried out its own analyses. The results showed that, contrary to the GAO preliminary findings, the ASVAB is not biased against females and racial/ethnic minorities in technical specialties (Wise et al., 1992, September). The issue of test fairness, in the sense of similar regression of performance on ASVAB scores, has been fairly well resolved in favor of the ASVAB. This joint-service validation of the ASVAB was a first for the Testing Center as the ASVAB executive agent.

Joint-Service Validation of Enhanced Computer-Administered Testing (ECAT)

A joint-service validation effort was initiated in 1990 to evaluate the ASVAB and new predictors that could be used in the ASVAB. This study was designed to compare the validity of a set of new tests, called Enhanced Computer-Administered Testing (ECAT), to the ASVAB in all Services. The intent of this study was to obtain data that could inform decisions about changing the content of the ASVAB in the mid-1990s. The Navy served as the executive agent for this effort.

The ECAT was an outgrowth of attempts to develop CAT, the computer-adaptive test. In the late 1980s, the military testing community realized that the feasibility of the ECAT for operational use in testing applicants was dependent on the inclusion of new predictors that cover test content that were not possible to cover with the existing paper-and-pencil or CAT ASVAB. Some ECAT tests were developed under the JPM Project, and some under other Service research projects to evaluate new predictors. In 1989, a special committee was formed by OASD-FM&P-AP to nominate tests for inclusion in the ECAT project. Data collection began in 1990 to validate the tests in the ECAT project, and all Services provided validity data for evaluating the ECAT tests. Analysis of the data was projected for completion in 1993.

In the ECAT validation, a call went from the executive agent to the other Services to provide validation data. The response was varied. The Army provided data for three combat arms specialties, and the Marine Corps provided data for mechanics tested with job-performance tests. Both the Army and Marine Corps were collecting these data for other purposes, and they agreed to provide them to the ECAT project. The Air Force agreed to provide training grades for four specialties, but because of the reduced flow of students, both the number of specialties and the sample sizes were reduced from the original intent. The Navy from the beginning agreed to provide the bulk of the validation data, with the criterion measure being performance in training courses. The research design and analysis plan for the ECAT validation were worked out by the executive agent in response to the data made available by the Services.

The value of the ECAT project remains to be determined for informing decisions about changing the ASVAB. One problem was that the sample sizes did not turn out to be as large as projected and desired. With the decreasing number of recruits in the early 1990s, the flow of students through training courses has been cut back, and some courses in the ECAT project have been eliminated. Another, and more fundamental problem, was the variety of criterion measures used for evaluating performance, and the lack of standardization across the Services. The Navy provided the bulk of the samples, and the criterion measure for the Navy specialties was performance in training courses.

The Air Force also provided data for a few specialties using training grades as the criterion measure. The Army provided data for three specialties, with performance on gunnery simulators as the criterion measure. The Marine Corps used hands-on performance tests administered to maintenance mechanics for ground equipment and aircraft as part of its JPM effort.

The extent to which decisions about changing the ASVAB will be informed by the ECAT validation results at this stage is problematic. A small case in point concerns the usefulness of psychomotor tests for predicting performance in the combat arms specialties. The Army had been using psychomotor tests to help place people into gunnery training courses; proficiency of the trainees was evaluated through performance on gunnery simulators. Army technical representatives reported that the effort was successful. On the other hand, the Marine Corps used a psychomotor test as a predictor of job performance of infantrymen in its JPM project. The predictor consisted of simulated firing of a pistol at moving targets on a video screen. The criterion measure included firing the M-16 rifle at targets that popped up while the examinee was navigating through terrain. The predictive validity of the simulated firing against live firing of the M-16 rifle was marginal, and negligible against the total performance of infantrymen.

The question of adding psychomotor tests, and other types of tests, to the ASVAB rests on resolving issues such as this. The Marine Corps representatives are not likely to give wholehearted support to using psychomotor tests for setting qualification standards for the combat arms specialties, whereas the Army representatives are likely to remain, as they have been all along, highly supportive.

Computer-Adaptive Testing

A joint-service validation effort was undertaken in the mid-1980s to compare the validity of a computer-adaptive version of the ASVAB with that of the paper-and-pencil version. The Navy, as executive agent for computerized testing, performed the analyses. The Services provided training grades for people in selected specialties, and these were correlated with their ASVAB scores, in either the computerized or paper-and-pencil mode of administration. The results showed that both versions were equally effective as predictors of training grades.

Computer-adaptive testing permits individualized testing with flexible start and finish times. Paper-and-pencil testing, by contrast, has a fixed start time for the group of examinees and ordinarily a fixed finishing time. The time limits for some tests can be shortened by the test administrator if all examinees are clearly finished before the time limit is reached, but the testing schedule must allow for the full block of time. With the lock step arrangements for paper-and-pencil testing, an examinee who is late for the testing session is turned away and required to return for another session. With computerized testing, examinees may be able to start taking the test whenever they show up and the computers are operating and available. However, the individualized scheduling possible with computerized testing requires major changes in testing operations that affect MEPCOM, recruiters, and examinees.

An advisory panel composed of Service recruiting managers and MEPCOM operators was formed in 1990 to advise the Testing Center and OASD-FM&P on new concepts of operations that might be employed with computer-adaptive testing. A concept that has been popular with the Services is to extend the notion of individualized testing to include choice of what tests are given when. One extreme outcome is that each Service would have a unique set of supplementary tests, to be administered to applicants or recruits, in addition to a common core such as an expanded version of

the AFQT. In effect, this extreme outcome would be a return to pre-1976 when each Service had its own classification battery and only the AFQT was common. Less extreme concepts are to have greater uniformity in test content and place of administration across Services. Changes of the magnitude to testing operations envisioned with computerized testing require the review and approval of the Steering Committee and the ASD-FM&P. If new computers need to be purchased, then the change involves more extensive review and approval in the Department of Defense.

The intent of the ECAT project was to evaluate tests for possible inclusion in the ASVAB and thereby help justify using computers in the MTP. Preliminary data for the ECAT tests about increasing the predictive validity of the ASVAB through adding the ECAT tests are not encouraging. Prior research on these instruments by the Services has shown marginal incremental validity for predicting either training grades or job performance measures, or both. The most promising candidates for inclusion in the ASVAB are spatial ability tests, as these have consistently the highest incremental validity of the ECAT tests. Spatial ability can be measured in the paper-and-pencil mode, as it has been for decades. Military classification tests had spatial tests until 1980 when the Space Perception test was deleted from the ASVAB. The justification for switching from paper-and-pencil testing to computerized testing is more likely to be on the basis of individualized testing, which would benefit both recruiters and applicants, than from improved validity and improved quality of personnel decisions.

Development of the Adaptability Screening Profile (ASP)

A self-description instrument, called the Adaptability Screening Profile (ASP), has been under development since the mid 1980s for joint-service use as a predictor of attrition through the first term of enlistment. The development of the ASP has been somewhat in response to a crisis. Congressional pressure has been placed on OASD-FM&P either to change the educational tiers used in making personnel decisions or to provide an alternative to the tiers.

Educational level is related to attrition because of the general characteristics of people in the various groups. The Army capitalized on the behaviors and attitudes that are related to attrition through the development of a self-description inventory, called the Military Applicant Profile (MAP), in the 1970s. The MAP contained items that tapped biographical data about educational, social, and work experiences, and attitudes towards social institutions and experiences. The instrument was used by the Army in the early 1980s to help screen selected applicants, especially young nongraduates. The other Services were conducting research on similar instruments during this period.

Currently, the Department of Defense uses three tiers of educational credentials for categorizing applicants because in all Services educational level has a strong relationship to attrition. High school graduates have the highest completion rates, and nongraduates the lowest, with alternative credential holders in between. The top tier initially included high school diploma graduates and people with college education. This group is the most favored for enlistment, and quality recruits are defined as high school graduates with AFQT scores of 50 or higher. The middle group initially included those with alternative credentials, such as the General Educational Development (GED) program, home study, or adult education. The bottom tier included nongraduates. Over the years, experience has been that about 20 percent of the high school diploma graduates, 30 percent of those with alternative credentials, and over 40 percent of the nongraduates fail to complete their first term of enlistment. The Services keep the number of nongraduates to a minimum and attempt to control the number in the middle tier.

Congressional pressure built up in the mid 1980s for the Department of Defense to change the definitions of the educational tiers. Some constituents complained to their Congressman that their educational programs were being discriminated against by being placed in a lower tier than they should have been. Specifically, some representatives of adult education programs said that people who complete adult education programs should be placed in Tier One along with diploma graduates. Proponents of the GED program also have questioned the placement of their participants in the middle tier. The tiers were initially defined in the early 1980s on the basis of empirical relationships of various credentials to attrition from the Services. Before the tiers were defined, each Service made their own rules about educational credentials, and they were not consistent. The consistent definitions drew attention from interested groups, and controversies arose about who is placed in which tier.

When congressional pressure built up in the mid 1980s, OASD-FM&P-AP designated the Navy as executive agent to develop a joint-service instrument to predict attrition, and the result was the ASP. Development of the ASP was not under the purview of the MAPWG. Instead, special joint-service groups met early in the development of the ASP to coordinate the effort. In later years the MAPWG was kept informed of developments, and it has provided input on occasion to the executive agent and Director for Accession Policy. The DAC and the Steering Committee were also kept informed about the ASP, and they too have provided input about its development and projected use for making personnel decisions.

Use of self-description instruments for making selection decisions remains controversial. As found with the interest items in ACB 73, people can easily inflate their scores when the desired answer is apparent. They can fake answers to make themselves look good. The MAPWG in the mid 1980s raised questions about use of self-description inventories for making selection decisions. The DAC in the late 1980s made stronger statements about the need to verify the amount of distortion to the score scale from faked responses and to evaluate the impact of faking on validity of the scores. The Steering Committee, when briefed on plans to use the ASP on a trial basis for making selection decisions, would not approve its use.

The Navy as executive agent for ASP had developed a comprehensive system to track ASP scores if and when they were used for making selection decisions, but skeptics did not endorse its use. Another issue that the Steering Committee raised was the possibility that ASP would screen out some high school graduates who otherwise would be qualified for enlistment. The controversies around the use of ASP precluded its introduction, even on a trial basis, for making personnel decisions, and the issue was forwarded to the ASD-FM&P in 1991 for resolution.

The controversy surrounding introduction of the ASP contrasts with the introduction of the MAP by the Army in the late 1970s. The Army started using the MAP for screening nongraduates, and no follow-up studies were done to evaluate distortion of scores or of predictive validity when the items were used operationally instead of in a research environment. Use of the MAP was suspended by the Army in the 1980s, but not on the basis of empirical analyses as were proposed for the ASP. The ASP as a joint-service effort was held to a higher standard than the MAP, which reflects the increased rigor of military testing procedures and the closer managerial control.

The contrast between the experience of the MAP and ASP, which are separated in time by the ASVAB miscalibration episode, is instructive for understanding the emerging environment of the MTP. Management of the MTP is becoming increasingly centralized in the 1990s. In the 1980s, research efforts that affected the MTP tended to be more fragmented. The JPM project perhaps warranted a separate working group to coordinate the effort, but as noted earlier the lack of

coordinated research on evaluating new predictors is now having its effect through the lack of joint-service data for informing decisions about changing the ASVAB. The fragmentation of efforts may have been necessary given the amount of resources available to support the MTP. The Testing Center now provides an opportunity to centralize more of the research and development functions that affect the MTP.

Joint-Service Validation of the ASVAB

Validation of the ASVAB assumed new direction in the 1980s. Previous validation efforts were conducted by each Service on its own people. Even the AFQT was validated within the Services. In the late 1980s, joint-service validation efforts were initiated for the purpose of revising the ASVAB. In addition, joint-service efforts have been underway to develop a self-description instrument that could be used to help make personnel decisions.

Currently, responsibility for the ASVAB is still fragmented in at least one major component. The current procedures for joint-service validation of the ASVAB as a predictor of performance are inadequate. Evaluation of Service-specific tests has an advantage over evaluation of the ASVAB in that the research staff, policy staff, and users are all working for the same Service. The sense of ownership is common among the participants for Service-specific tests. The ASVAB and the ASP, from their beginnings, have been joint-service endeavors, and both met resistance from individual Services in the beginning. The ASVAB did not enjoy the same harmonious working relationships among researchers, policy staff, and users that mark the Service-specific testing programs. In the early days of the ASVAB, the fragmented responsibility for the ASVAB among the major participants -- the Service acting as executive agent for research and development, OASD-FM&P providing managerial oversight, and MEPCOM serving as the test operator -- tended to preclude a sense of common ownership. The lack of attention and action to fix the low end of the score scale in 1976 may have stemmed in part from the diffuse responsibilities for the ASVAB. No single Service or office "owned" the ASVAB, as the Army did for developing, administering, and using the AFQT.

Decisions about changing the ASVAB have joint-service consequences, and prudence suggests that joint-service validation be used to protect the interests of all Services. Decisions about changing the ASVAB could possibly be made at a policy level in the absence of solid empirical validity data. Such an outcome may be forced by circumstances, but the lack of empirical evidence for changing test content is counter to the long tradition of military testing.

The validation of aptitude tests in the joint-service arena is still very much in the formative stages. Just as in 1974, when the Services had limited experience in developing aptitude tests in the joint-service arena, and the first efforts left much to be desired, so these early joint-service efforts to validate tests for joint-service use so far have had serious shortcomings. In the validation area, each Service is proceeding with what it wants to do, and any problems with combining the validity data will be faced later. No project plans were developed prior to the data collection that specified how the data would be collected, analyzed, and used for making decisions about changing the ASVAB. In the absence of research design and analyses plans, the decisions will need to be made post hoc by whomever is involved at the time.

The design of a joint-service validation effort is no different from that of a Service-specific effort. No competent researcher would compare the absolute validity of two predictors when the criterion measure for one is training grades and the other is supervisors' ratings for the other. The latter

measure is virtually bound to result in lower validity coefficients. And yet the current ECAT validation effort has similar disparate criterion measures. One rule, then, for joint-service validation efforts should be that the criterion measures are comparable for all tests in the studies, both within and across Services; this would permit comparisons of the absolute value of all validity coefficients. Of course, the coefficients must also be on a common metric, which means that the sample values must be converted to population estimates.

A second rule for designing joint-service efforts is that the Services and OASD-FM&P agree on the questions to be answered by the effort. A basic question for evaluating tests is whether they should have high general validity for most specialties, or should the validity be for specific families of specialties such as electronics. In the ECAT validation, one possible outcome is that the psychomotor tests have specialized validity, as the Army found for predicting simulated gunnery tasks for combat arms specialties. (If psychomotor tests are good predictors for combat arms specialties, but not for other specialties, and if they have low intercorrelation with the cognitive tests in the ASVAB, then they are ideal differential measures.) The Marine Corps and Navy also have gunner specialties; if psychomotor tests are added to the ASVAB, would the Marine Corps and Navy also have to demonstrate differential validity for these tests? Currently there are no efforts underway to evaluate the psychomotor tests, or any other new tests, for similar specialties across the Services, to say nothing of using comparable criterion measures.

The validity of existing ASVAB tests can be compared across Services because of their long use and frequent validation against training grades. The validity of the ASVAB against hands-on job performance tests, however, shows that joint-service comparisons cannot always be taken at face value, even when the criterion measure is nominally the same. The validity of the AFQT against hands-on tests for similar combat arms specialties ranged from 0.66 for Marine Corps machine gunners, to 0.15 for Army cannon crewmen. The Marine Corps tested four combat arms specialties, and the range of validity coefficients was 0.38 to 0.66; the Army tested three combat arms specialties, and the range was 0.15 to 0.34 (Wigdor & Green, 1991). When job-knowledge tests were used as the criterion measure of performance, the Army and Marine Corps validity coefficients were more comparable. Establishing the comparability of criterion measures involves more than just having the same general procedures, such as supervisors' ratings or hands-on tests. The relevance of the criterion measure to actual job performance is crucial in interpreting validity coefficients, and the care taken to ensure relevance of the criterion measure is another component of designing joint-service validation efforts.

The validity of the ASVAB is of more fundamental importance than the accuracy of the score scale, even though it is typically less visible. Rigorous joint-service procedures have been developed, and accepted as legitimate to scale the ASVAB and check on the accuracy of the score scale; the rigor of the score scale grew during the 1980s in response to errors in the scores reported through the ASVAB miscalibration episode. However, no one currently is devoting the same rigor and control over the research procedures for validating the ASVAB scores as predictors of performance. The necessary resources to design and carry out a proper validation effort have not been made available to the military testing community; managerial control over test validation efforts is not as rigorous as it is over the accuracy of the score scale. The joint-service validation efforts currently have been accorded only the same degree of respect and support as the initial scaling of the ASVAB in 1974 and 1975. Although the validation of the ASVAB is not in the same crisis mode that grew out of the ASVAB miscalibration, with care exercised now to validate the ASVAB properly, no crisis is likely to arise, and the integrity of the MTP can be maintained.

CHAPTER 3

THE DEPARTMENT OF DEFENSE STUDENT TESTING PROGRAM

The Department of Defense Student Testing Program (STP) had a modest beginning in 1968 with the introduction of the first Armed Services Vocational Aptitude Battery (ASVAB). The STP grew out of efforts to obtain recruiting leads by using the Service classification batteries to test students in schools. The Air Force was the first Service to test students, using the Airman Qualifying Exam in 1958. Soon after, schools were approached by recruiters from other Services offering to give their tests. In 1966, the Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P) directed the Services to develop a common classification battery for testing in the schools. Form 1 of the ASVAB was ready two years later, in time for the 1968-69 school year. It was accompanied by a counselor manual that provided some information on using the test scores in vocational counseling. The prominence of materials to help counselors and students interpret the test scores expanded over the years, and the ASVAB 18/19 Career Exploration Program (CEP), introduced in 1992, provides a wealth of materials that conform to current professional recommendations for vocational counseling and career exploration.

The ASVAB Career Exploration Program (CEP)

The CEP is the latest generation of the ASVAB materials for use in high schools and postsecondary schools. It is the culmination of about ten years of efforts to improve the program provided to counselors and students. In 1987, a panel of counseling experts was convened by the Directorate for Accession Policy in OASD-FM&P (OASD-FM&P-AP); the panel made sweeping recommendations, most of which are incorporated into the CEP. Since the late 1980s, the Defense Advisory Committee on Military Personnel Testing (DAC) has had at least one member that is expert in vocational counseling and career exploration, and the DAC, especially through this member, has provided advice on the STP. Full-time expertise for the STP has been provided since the early 1980s, when OASD-FM&P-AP obtained contractual support to develop counseling materials; in the mid-1980s, an expert in vocational counseling and career exploration became a member of the OASD-FM&P staff.

The key components of the CEP are (a) an interest inventory to supplement the aptitude scores provided by the ASVAB, (b) an updated counselor manual, and (c) a student workbook that includes a chart for matching search variables that describe students and occupations. The ASVAB score reports that are delivered to schools and students have been revamped to improve their usefulness.

The Self-Directed Search (SDS)™, developed by John L. Holland (1985), is the interest inventory that has been incorporated into the CEP. Designed to be self-administered and self-scored, the SDS is one of the most widely used interest inventories in American schools. It is included in the student workbook, entitled *Exploring Careers: The ASVAB Workbook* (U.S. Department of Defense, 1992a), together with instructions for answering the items and scoring the responses. Inclusion of the

SDS marks a departure from prior practices in the STP, when only the ASVAB scores were reported to schools and students. The inclusion of a vocational interest inventory increases the usefulness of the CEP for students and counselors because interest scores are useful in helping direct the search activities into occupational areas. The efficiency and effectiveness of career exploration is likely to be increased through incorporation of interest scores into the search process.

The Department of Defense has arranged with the publisher to use the SDS for three years in the CEP, from 1992 until 1995. In the meantime, it is developing a new interest inventory for use in the CEP. The fact that a new interest inventory is being developed specifically for use in the STP implies a long-term commitment to provide interest scores, as well as aptitude test scores, to students.

The student workbook activities allow linkage of a student's interests and aptitude scores with work values and educational plans to suggest occupations that the student may want to explore further. The workbook provides instructions to users and uses a cartoon format to present concepts of career exploration. The culmination of activities in the workbook is facilitated by an OCCU-FIND chart in which the above search variables are matched to occupations.

A student's ASVAB scores are translated to an ASVAB Code that is used with the OCCU-FIND chart. The ASVAB Code is also reported to schools and students as the Academic Ability score; it is based on a student's verbal and math test results. There are five codes, each representing a section of the Percentile Score Range:

ASVAB Code	Percentile Score Range
1	90 - 99
2	70 - 89
3	50 - 69
4	30 - 49
5	01 - 29

The ASVAB Code is used to help identify occupations whose incumbents tend to have aptitude levels similar to the student. The association between the ASVAB codes and occupations was developed empirically. Occupations in the *Dictionary of Occupational Titles*, published by the U.S. Department of Labor (1991), have been rated on numerous characteristics by panels of occupational analysts. The following ratings of occupations are combined to form an occupational complexity score: (a) general educational development requirements; (b) specific vocational preparation; (c) general, verbal, and numerical aptitude ratings; and (d) repetitive work temperament rating. The occupations in the dictionary were ranked on their complexity score and divided into five levels. The levels of occupational complexity were then associated with the five ASVAB Codes. The intent of using aptitude scores in the CEP is to encourage students to focus their efforts toward occupations in which they are likely to be satisfactory performers while exploring a broad range of occupations.

After students complete the OCCU-FIND and identify the occupations that best match their interests, aptitude level, and personal preferences, they can learn more about specific occupations through the use of other student activities that are suggested in the workbook. The activities include reading about occupations, talking to people who know about the occupations, and actually gaining relevant experience in the occupations.

Supporting Materials for Earlier Versions of the STP

The one document that has been available since the introduction of the STP, aside from the ASVAB itself, has been a counselor manual. The first manual was published in 1968, in time for introduction of the ASVAB 1; it contained 43 pages, of which only one was devoted to using the ASVAB test scores in counseling while 27 were devoted to listing civilian and military occupations associated with the ASVAB scores. In contrast, the counselor manual for the CEP has 147 pages, with numerous case examples scattered throughout the manual. The association of military and civilian occupations currently is covered in a separate document. The manuals have expanded over the years to present increasingly detailed information to help counselors explain and use the ASVAB scores. The current counselor manual, of course, also describes the proper use of the interest inventory scores.

Since the late 1980s, the STP has also included a book, called *Military Careers: A Guide to Military Occupations and Selected Military Career Paths* (U.S. Department of Defense, 1992b), to describe the variety of military occupational specialties, both enlisted and officer, available to young people, and to list related civilian occupations.

Content of the ASVAB and Scores Reported in the STP

The ASVAB 1, introduced for school year 1968-69, reported three sets of aptitude composite scores: one for the Navy, one for the Air Force, and one for the Army and Marine Corps together. The composite scores reported in school year 1968-69 were as follows:

Composite	Service
General Technical	All
Clerical/Administrative	All
Electronics	All
General Mechanical	Army/Marine Corps
Motor Mechanical	Army/Marine Corps
Mechanical	Navy/Air Force

The tests in some composites varied for different Services, even though the label may have been the same.

Apparently the three sets of aptitude composite scores proved unworkable. In the summer of 1969, a change to the counselor manual presented a new set of scores to be reported to the schools. These were percentile scores for the nine tests in the ASVAB 1 and percentile scores for five composites, shown in Table 5. These scores were also reported for form 2 of the ASVAB that was introduced in January 1973.

Table 5

Tests in Forms 1 and 2 of the ASVAB and Composite Scores

Test Names

Word Knowledge (WK)
Arithmetic Reasoning (AR)
Tool Knowledge (TK)
Space Perception (SP)
Mechanical Comprehension(MC)
Shop Information (SI)
Automotive Information (AI)
Electronics Information(EI)
Coding Speed (CS)

Aptitude Composite	Definition
Electronics	(MC + 2EI)/3
General Mechanical	(SP + 2SI)/3
Motor Mechanical	(MC + 2AI)/3
Clerical/Administrative	(WK + CS)/2
General Technical	(WK + AR)/2

In July 1976, the ASVAB 2 was replaced by form 5; because this form had different content, different scores were reported. A score sheet was provided for each student, but they were delivered by the counselors because the test scores were reported to the schools. A new aptitude composite, Communications, was added in 1976, and the tests in the remaining composites, except for General Technical, were changed from the previous version. The tests in the ASVAB 5 and the composite scores reported are shown in Table 6.

In the school year 1977-78, the aptitude composites reported to schools were radically revised. The rationale underlying the aptitude composites used in 1976 and earlier was that they could be used by students and counselors to indicate chances of performing successfully in related occupational fields, and the intercorrelation among the composites was not a driving concern for the developers of the STP. However, during the school year 1976-77, the intercorrelation among the aptitude composites became a serious concern because of a review of the ASVAB by Professor Lee Cronbach that first appeared in February 1977 (Cronbach, 1979, January). (His review is treated at some length in the next section of this chapter.) For purposes here, the salient point is that the composites were criticized by Cronbach as being too highly intercorrelated for counseling purposes; students and counselors would likely overinterpret differences among the composite scores.

During that same time period, six new composite scores to replace the previous ones were developed (Fischl, Ross, & McBride, 1979, April). The new composites and their definitions are also shown in Table 6. These scores were reported to students and schools until July 1984, when

Table 6

Tests in Form 5 of the ASVAB and Composite Scores

Test Names

Word Knowledge (WK)
Arithmetic Reasoning (AR)
Space Perception (SP)
Mechanical Comprehension (MC)
Shop Information (SI)
Automotive Information (AI)
Electronics Information (EI)
Mathematics Knowledge (MK)
Numerical Operations (NO)
Attention to Detail (AD)
General Science (GS)
General Information (GI)

Composite scores reported in school year 1976-77

Aptitude Composite	Definition
Electronics/Electrical (EL)	AR + EI
Communications (CO)	AR + SP + MC
General Technical (GT)	AR + WK
Motor Mechanical (MM)	MK + MC + AI
General Mechanical (GM)	AR + SP + SI
Clerical/Administrative (CL)	WK + NO + AD

Composite scores reported in school years 1977-84

Aptitude Composite	Definition
Verbal (VE)	WK + GS + GI
Analytic, Quantitative (AQ)	AR + MK
Academic Ability (AA)	WK + AR
Clerical (CL)	NO + 3AD
Mechanical (ME)	SP + MC
Trade Technical (TT)	SI + AI

the ASVAB 5 was replaced by the ASVAB 14 in the STP. (The 12 test scores for the ASVAB 5 continued to be reported to schools, even though Cronbach had criticized this practice because of their low reliability.)

In July 1984, the ASVAB 14 was introduced into the STP. (The ASVAB 14 was previously used in the joint-service program for testing applicants; it was then form 9.) The tests and composites reported for the ASVAB 14 are shown in Table 7.

With the ASVAB 14, the reporting of scores on individuals tests was stopped, and the clerical, mechanical, and trade technical factor scores also were deleted. Four new occupational composites were added to the verbal and math factor scores retained from ASVAB 5, along with the Academic Ability score. The last four composites in Table 7, called occupational composites, were used in the *Military Careers Guide* to show the chances of qualifying for at least one specialty in each cluster of military occupational specialties.

With the introduction of the ASVAB 14, the applicant testing program and the STP were for first time truly parallel programs. Because the same tests were used in both programs, identical composites could be computed for both students and applicants.

Changes in the scores reported to students and schools came about for various reasons. When test content was changed, the tests in some composites also had to change. Some changes in the composite scores came from the military testing community, such as adding the occupational composites in an attempt to make the scores more useful to students, schools, and the military services. Other changes came about because of external review of the ASVAB and the STP by testing and counseling experts. The next section recounts the salient reviews that changed the course of the ASVAB and the STP.

Professional Reviews of the ASVAB and STP

The most controversial area for the STP over the past 25 years has centered on (a) the scores reported to schools and students, and (b) the stated or implied interpretation of them for vocational counseling and career exploration.

The first and most devastating review of the ASVAB and the STP was by Professor Cronbach; it was distributed to the Congress and Department of Defense in February 1977, before it was released to be published in the professional literature in 1979 (Cronbach, 1979, January). Cronbach was especially critical of the following:

- the six aptitude composites reported for the ASVAB 5 - because of their high intercorrelation;
- the association of the composite scores with occupations, some of which lacked credibility (e.g., stating that the Communications composite, which contained mechanical and spatial abilities, could be used for translators and fashion designers);
- the interpretation of score levels in terms of red (stop), yellow (caution), and green (go) on the student score results sheet. The red, yellow, and green score bands had no empirical justification for civilian occupations, and there was no attempt to consider occupational complexity in the association of test scores and occupations.

After his review appeared, factor composite scores were developed to replace the earlier composites. The factor composite scores were related to specific occupations, as were the earlier

Table 7

Tests in Form 14 of the ASVAB and Composite Scores

Test Names

Word Knowledge (WK)
Paragraph Comprehension (PC)
Arithmetic Reasoning (AR)
Mathematics Knowledge (MK)
Auto & Shop Information (AS)
Mechanical Comprehension (MC)
Electronics Information (EI)
Numerical Operations (NO)
Coding Speed (CS)
General Science (GS)

Composite	Definition
Academic Ability (AA)	VE (WK +PC) + AR
Verbal (VBL)	VE + GS
Math (MTH)	AR + MK
Mechanical and Crafts (MC)	AR + MC + AS + EI
Business and Clerical (BC)	VE + MK + CS
Electronics and Electrical (EE)	AR + MK + GS + MC
Health, Social, Technology (HST)	VE + AR + MC

composites, but Cronbach thought that the new association was more reasonable. The red, yellow, and green bands disappeared from the student results sheet. The number of schools and students participating in the STP declined immediately after Cronbach's review, but started creeping up after the fixes were made, and eventually reached the former levels of about one million students tested each year.

In 1985, the ASVAB 14 was reviewed by Professor Arthur Jensen (Jensen, 1985, April). Although he was favorably impressed with the psychometric quality of the ASVAB, Jenson, in the same vein as Cronbach, was critical of the occupational composites because of their high intercorrelation. He recognized that the occupational composites were useful for purposes of military selection and assignment decisions (hence their continued use by the military services), but they were practically useless as a basis for making differential occupational choices by individuals. He proposed that a single score of general mental ability be used in lieu of the four occupational composites. He also requested that information be provided about the aptitude scores for youth who later become successfully engaged in various occupations.

The same criticisms of the occupational composites were again voiced by a panel of counseling

experts in November 1987, and the occupational composites were dropped with the introduction of the CEP in 1992.

Marketing the STP

Recruiting interests have been the driving force behind the STP. The program was conceived by the recruiting commands, and local recruiters have continued to maintain contact with schools with the intent of obtaining recruiting leads. The cost of conducting the STP (except for research and development, including printing and distribution of the materials) is carried largely by the Military Entrance Processing Command (MEPCOM) and recruiting commands. The STP continues in existence because recruiting commands want the leads obtained from the students who take the ASVAB.

The support of the STP by the military services is prominently displayed by recruiters and in the written materials. However, the role of the military in the STP was not always so apparent; the issue came to a head in the mid-1970s through the efforts of Congressman Mosher from Ohio. His concern in this area was the evident obfuscation of the connection between the STP and military recruiting.

The connection between the STP and military recruiting is now explicit. During the 1980s, schools were given eight recruiting contact options, shown in Table 8. Schools participating in the STP chose one of the eight options, each of which clearly implies that recruiters ordinarily are involved in using the ASVAB results from the STP. Some schools did choose option 8--no recruiter contact based on the STP--and they continued to test year after year; individuals from those schools could have their scores released to recruiters by signing a release form.

Congressman Mosher also voiced concern about the storage and use of test scores and personal data obtained in the STP. Currently, the policy about storing and using test scores and personal data is carefully spelled out and adhered to by the Department of Defense in a Privacy Act statement that students must sign to receive their test scores. The policy is as follows:

> Personal data supplied by students will be used only by the military services for recruiting, school counselors for career guidance, and Department of Defense researchers; the data will not be divulged for any other purpose; test scores are provided only to the school counselors and to recruiting commands; test results and identifying personal data are maintained by the Services for up to two years; after two years individual test scores, identified by student name and social security number, are retained by the Department of Defense for research purposes only; test scores are not released outside the Department of Defense, except to the Coast Guard.

A resolution was effected in 1976 by OSD-FM&P that satisfied another of Congressman Mosher's main concerns: claims that the ASVAB test results were directly applicable to counseling students for civilian occupations. The issue of validity of the ASVAB for civilian occupations essentially was postponed through a promise by OSD-FM&P that studies would be conducted. The validity of the ASVAB in the STP is reviewed in a following section.

Table 8

Recruiter Contact Options for the STP

1. No special instructions. Release results to recruiting military services 7 days after test scores are mailed to the schools.

2. Release results to recruiters 60 days after test scores are mailed. No recruiter contact prior to that time.

3. Release results to recruiters 90 days after test scores are mailed. No recruiter contact prior to that time.

4. Release results to recruiters 120 days after test scores are mailed. No recruiter contract prior to that time.

5. Release results to recruiters at the end of the school year. No recruiter contract prior to that time.

6. Release results to recruiting military services 7 days after test scores are mailed. No telephone solicitations by recruiter based on the student names provided with the listing of student results.

7. Not valid for enlistment purposes. Results not released to recruiting military services.

8. No recruiter contact from this listing of student results. Results not released to recruiting military services.

The marketing of the STP assumed a new dimension in December 1972. Prior to that time, the scores could only be used to provide recruiting leads; after that time, the ASVAB scores obtained in the STP were valid for enlistment in the military services. The same composite scores continued to be reported to schools for use in vocational counseling and career exploration; the new dimension was that the Services could compute their own aptitude composite scores from the STP results, and these scores could be used to help make military personnel decisions about selection and classification.

Administration of the STP

When Air Force recruiters started offering the Airman Qualifying Examination to high schools in 1958, the tests were administered and scored by Air Force personnel. When the other Services followed suit, they too administered and scored their own tests. So, when the ASVAB was introduced in 1968 joint-service arrangements had to be made for administering and scoring the ASVAB. The initial arrangements required that the Services shared the responsibility for contacting, administering, and scoring the ASVAB in designated schools. The schools in a region were divided among the Services, and each Service then was responsible for the ASVAB testing in assigned schools.

Then in 1973, responsibility for the ASVAB was centralized at a new organization set up specifically to handle the ASVAB: the Armed Forces Vocational Testing Group (AFVTG). AFVTG was run by the Air Force and was staffed to print the ASVAB materials, score the answer sheets, report scores to schools, and prepare interpretative materials such as the counselor manual. Test administrators were provided by military examining stations near the schools.

New scoring procedures were introduced by AFVTG. Whereas formerly conventional answer sheets were used with the ASVAB, AFVTG started using mark-sense cards, which had oval spaces for recording answers; they were fed through a machine that read the marks and converted them to punched holes in IBM cards. The IBM cards could then be read by IBM machines and computers.

In 1976, the AFVTG was disbanded and the responsibility for STP was moved to MEPCOM. Then in the early 1980s, responsibility for preparing STP materials was gradually shifted from MEPCOM to OASD-FM&P-AP, with the Air Force as executive agent for research and development of the ASVAB. MEPCOM in the 1980s continued to operate the STP, but it was no longer the primary preparer of interpretative materials for the STP. Since 1989, preparation of all STP materials is the responsibility of the Personnel Testing Division of the Defense Manpower Data Center, also known as the Testing Center, which functions as the executive agent for research and development of the ASVAB.

In the early 1980s, new optical readers were purchased by MEPCOM to score the ASVAB answer sheets, so scoring the ASVAB for the STP became feasible at the military examining stations. The turnaround time to report scores back to schools and recruiters was shortened to a matter of days rather than weeks.

The foregoing description of how the STP was administered since 1968 reveals that several organizations have been involved in the operation of the STP. Responsibility for research and development of the ASVAB similarly has been moved around in the past 25 years. The Army had been the executive agent for the AFQT since its inception in 1948 when the first joint-service committee was assembled to develop the AFQT; the Army performed most of the research tasks. In 1966, when OASD-FM&P directed the Services to develop a joint-service test for use in high schools, the Army again was designated executive agent.

The research design for ASVAB 1 was worked up by a joint-service committee representing the four Services. Three Service classification batteries were then in use, and all three batteries were administered to samples of recruits from each Service. The tests in the three batteries were intercorrelated to identify the common tests. The nine tests listed earlier in Table 5 for ASVAB 1 were those common to the Service batteries and the AFQT. Tool Knowledge was in the AFQT, but not in the Service classification batteries; because ASVAB 1 had to provide an accurate estimate of an AFQT score, Tool Knowledge was included in ASVAB 1. The common content of the Service classification batteries essentially covered the content of each of them. Development of ASVAB 1 has been reported by the Army (Bayroff & Fuchs, 1970, February).

The Army was also responsible for developing forms 2 and 3 of the ASVAB, with form 2 introduced in January 1973 as a replacement for ASVAB 1, and form 3 used by the Air Force and Marine Corps to test applicants (Seeley, Fischl, & Hicks, 1978, February).

In 1972 and 1973, the military testing program went through major changes: use of the AFQT was made optional, classification testing was moved from recruit centers to examining stations,

AFVTG was formed, and the Air Force was designated by OSD-FM&P to be executive agent for the ASVAB.

Validity of the ASVAB for the STP

The argument advanced by the military testing community to support the use of the ASVAB in the STP has been that validity for military occupations generalizes to similar civilian occupations. This argument was implied in the first ASVAB counselor manual of 1968 when comparable military and civilian occupations were listed, and over the years it has become more formally stated in the STP documents. The validity of the ASVAB covers the full range of military specialties, from those exclusive to the military, such as combat arms and ordnance handlers, to those also found in the civilian economy, such as repair, clerical, medical service. The correspondence of military and civilian occupations has been well documented through comparisons of job requirements (U.S. Department of Defense, 1992c); generalizing ASVAB validity from military to similar civilian occupations therefore is well justified.

A continuing effort to link the ASVAB and the General Aptitude Test Battery (GATB), published by the Department of Labor and widely used in employment testing for civilian occupations, has been underway since the mid-1970s. In 1976, the ASVAB 5 and GATB were administered to high school students (Kettner, 1976, October); the correlations between tests that have similar content are shown in Table 9. (Note that the GATB does not have an MK test, only an AR math test.)

Table 9

Correlation of the ASVAB and GATB Tests

Samples of high school students

	Grade 11		Grade 12	
Test	Females	Males	Females	Males
Verbal tests	.69	.73	.81	.69
Math tests				
AR*AR	.68	.77	.77	.72
AR*MK	.71	.72	.75	.75
Spatial tests	.57	.71	.72	.69

Samples of military recruits

Test	R
Verbal	.59
Math Tests	
AR*AR	.61
AR*MK	.54
Spatial (SPxMC)	.53

In 1991, the ASVAB and the GATB were administered to samples of military recruits. The correlations among similar tests are also shown in Table 9. The correlations among similar tests were high, similar to those found earlier on samples of high school students. Note that the generally lower correlations for recruits reflect the greater homogeneity of recruits compared to high school students. In general, these high correlations lend support to the argument that the ASVAB is a valid predictor of performance in civilian occupations.

In the early 1980s, a study using the procedures of validity generalization was conducted to evaluate the degree of overlap between the GATB and the ASVAB and then to generalize predictive validity from the GATB for civilian occupations to the ASVAB (Hunter, Crosson, & Friedman, 1985). The findings showed that the predictive validity found for the GATB against civilian occupations could be generalized to the ASVAB because the two batteries had similar content; both were good measures of general mental ability. The validity of the GATB also applied to the ASVAB, because the GATB had been extensively validated as a predictor of civilian occupations through validity generalization.

The ASVAB has also been validated against course grades in high schools. The predictive validity coefficients for a broad range of courses, from academic to vocational, range from about 0.3 to over 0.5 (Jensen & Valentine, 1976, March; Fairbank, Welsh, & Sawin, 1990, September).

When Congressman Mosher complained in 1974 that the ASVAB had not been validated for civilian occupations, and therefore its use in vocational counseling for civilian occupations was questionable, a validity study was promised. In 1985, Professor Jensen made a similar request in his review of ASVAB 14 (Jensen, 1985, April).

In 1992, a study to validate the current ASVAB against supervisors' ratings of job performance was completed for 11 civilian occupations (Holmgren & Dalldorf, 1993). The sample sizes ranged from about 200 to over 350 workers in each occupation. The workers in each occupation covered a wide range of age and length of experience in the occupation. The validity coefficients in the samples are shown in Table 10; correlations of AFQT, Technical, and Speed composites with overall job performance ratings by degree of selection are provided. These values were consistent with those found against supervisors' ratings in military specialties and were comparable to those found for the GATB in similar civilian occupations.

However, from a counseling and career-exploration perspective, the study that validated the ASVAB against supervisors' ratings was deficient in that a concurrent validity study was used instead of a predictive validity study. Workers of various ages and amounts of work experience took the ASVAB and received performance ratings concurrently with testing. On the other hand, high school students tend to be homogeneous in age and work experience and want to know how their current ASVAB results will predict performance one to five years in the future. Although concurrent and predictive validity generally are comparable in military samples, critics can still fault the validation study as not directly relevant for the use made of the ASVAB by students for career exploration.

Table 10

Validity of the ASVAB for Predicting Performance in Civilian Occupations

Measure	Low Selectivity* # Jobs	Mean Val**.	High Selectivity # Jobs	Mean Val.
AFQT	8	.24	3	.41
Technical	6	.29	5	.45
Speed	11	.23	0	n/a

* The sample for a particular job was considered selective if the sample mean on the measure in question was more than one-third of a standard deviation higher than the mean for the entire youth population.

** Validities are corrected for range restriction and criterion unreliability, but not adjusted for age or experience.

Issues About Reporting ASVAB Scores in the STP

The question about predictive versus concurrent validity of the ASVAB is only one of the concerns about how aptitude tests are used in vocational counseling and career exploration. In a preceding section, the various scores reported to schools and students from 1968 until 1992 were traced. (See Table 5 and Table 6.) The record is characterized by flip flops in going from aptitude composites to factor composites, then to both, and scores being reported for individual tests versus not being reported. In the CEP, scores on individual tests are reported to students on the result sheet, which was never done before. In the ASVAB 5, they were reported to school counselors who could readily pass them on to students; counselors were assumed to be available to interpret the test scores. The changing practices of reporting ASVAB scores reflects the ambiguous nature of aptitude tests in vocational counseling and career exploration.

What are the proper aptitude scores to report to schools and students, and what kinds of interpretative statements should accompany the scores? The reviewers cited earlier (Cronbach, Jensen, the panel of counseling experts) were unanimous in denouncing the aptitude composite scores because they were so highly intercorrelated, yet the military testing community kept putting them back in. A safe conclusion seems to be that the professional counseling community does not like correlated scores, whereas military testing psychologists concerned with personnel decisions are willing to use them.

The military testing community put aptitude composites in the STP from the beginning. From the point of view of predicting success in occupations, aptitude composite scores are the ones to report. Aptitude composites that contain general mental ability (verbal and math) plus specialized tests (technical and perhaps speed) are somewhat more valid predictors of training and job performance than are factor scores. To the extent that aptitude scores are used by counselors and students to predict performance in future occupations, the aptitude composites are the scores to use.

The number and content of composite scores have also been troublesome areas. By variously

reporting four, five, or six aptitude composite scores in the STP, the implication is that the scores carry differential meaning about predicting performance. The high intercorrelation among the composites, however, reduces whatever differential validity there may be among job requirements; therefore, reporting that many composites is questionable. The high intercorrelations mean that for most students, the composite scores have about the same value and are virtually interchangeable. For other students, however, large differences may occur (through error of measurement), and the tendency might be to overinterpret these differences as indicating greater potential in one area than another. For some students, of course, there are true differences in aptitudes for different types of occupations, and the reported differences in aptitude composite scores are meaningful. Unfortunately, for any individual there is no way of distinguishing scores based on true differences from scores based on error of measurement. A common practice used in interpreting test scores for individuals, and followed in the STP, is to say that differences larger than two standard errors of measurement are meaningful as probably indicating true differences in potential. These statements are probabilities, and there is no way of knowing how the probabilities apply to any given set of test scores.

The solution used in the CEP is to provide only one score as a predictor of performance in occupations: the ASVAB Code. This code in the CEP is the composite of verbal and math scores, and the validity for most occupations is about as high as would be obtained from adding technical and speeded tests.

Technical and speeded tests have troublesome properties for the STP. Technical tests (AS, MC, and EI) typically produce scores that have large gender differences. Females might easily feel discouraged from entering occupations for which the technical tests are valid predictors. With more training and experience in technical areas, many females might improve their scores and become proficient workers in technical occupations. No one yet has figured out how to construct a technical test with little or no adverse gender impact.

Speeded tests have lost much of their predictive validity in the military services. In the 1950s and 1960s, speeded tests were valid predictors of performance in clerical and administrative occupations. However, in the 1980s, military validation studies tended to show little unique validity for speeded tests. Also, speeded tests are susceptible to testing conditions, motivation of examinees, and practice. Their use in military aptitude composites has been cut back, and they are not used in the composites for the CEP.

The ASVAB is billed as a multiple aptitude battery with ten tests.[18] To report only one score of general mental ability for a multiple aptitude battery does not seem reasonable. Following Cronbach's criticism of the aptitude composites, five factor composite scores were developed in 1977. These were used until 1984, when the four occupational composites were introduced for the ASVAB 14. Verbal and math factor scores were retained in 1984, and again in 1992 for the CEP. The virtue of factor scores is that they are less intercorrelated than the aptitude composite score, but at the cost of some predictive validity.

Another solution might be the reporting of scores on individual tests without combining them into composites. Scores on individual tests were reported in the STP to schools beginning in 1969. However, Cronbach criticized the reporting of test scores to schools in 1977 because of their low

[18]The label of multiple aptitude battery is a replacement for differential aptitude batteries because in modern usage the word "differential" has become associated with test bias and, therefore, conveys a negative image.

reliability. They were dropped in 1984 because of their low reliability. Now they are once again being reported with the 1992 CEP, even though their reliability was not changed. When the CEP is reviewed by the professional testing and counseling community, some reviewer surely will be critical of reporting test scores to students because the differences are likely to be overinterpreted as differential predictors of performance in different occupational areas.

This lengthy discussion of test validity for the STP is intended to convey that no good solutions have been found for reporting ASVAB scores. The problem is not unique to the ASVAB, and other test publishers face the same dilemmas. The problem is more pronounced for the ASVAB because of the broader coverage of test content than is found for most batteries given in schools.

The reporting of both aptitude and interest scores in the CEP appears to strike a nice balance that serves the needs of the students and does not invite overinterpretation. The old occupational composites, as the critics argued, did invite overinterpretation for orienting students toward some occupational areas and away from others. As noted, the high intercorrelations meant that errors of measurement produced much of the observed differences in scores among the occupational composites; hence, the scores appeared to carry more information about relative strengths and weaknesses than they actually did.

A second criticism of using occupational composites for helping students focus on one or another occupational area is that the scores are a confounding of aptitude, interest, and opportunity to learn. People who acquire the skills and knowledge in an area, such as mathematics or technical, have demonstrated both a willingness and ability to acquire proficiency. A low score may mean a lack of interest, ability, or opportunity. A high score may include all three components. If a person has a higher composite score in one occupational area, what is the reasonable suggestion for exploring that area; or conversely, what does a low score mean for career exploration? Scores on the occupational areas do not provide clear clues about directions for students to take in their exploration.

In the CEP, interests and aptitudes are treated as two separate entities. The interest inventory provides indications about occupational areas in which the student may find work satisfaction. From a measurement point of view, the interests are relatively independent of aptitudes. The interest measure does contain items that ask for self-evaluations of ability in the various occupational areas, but there are no objective measures of aptitude confounded with the interest measures. Because interest and aptitude are separated in the CEP, students can decide how much emphasis they want to place on interests and how much they want to place on aptitude in their exploration of occupations.

The individual test scores reported on the Student Results Sheet can also be used by students to help them focus on certain occupations, or they can be ignored. If students want to consider relative strengths and weaknesses, as reflected in the test scores, the information is available. But no interpretation of the test scores is built into the exploration process; students and counselors are required to take the initiative if they want to use the test scores in exploring occupations.

A desired outcome from the all the information provided by scores in the CEP is that students will weigh the available information about themselves and use it in directing their search activities for those occupations that meet their interests, aptitudes, and other personal preferences related to the world of work. The CEP is designed to facilitate attaining this outcome.

Relationship Between the STP and the Joint-service Program

When the ASVAB began to be used for making military personnel decisions in 1973, the Services were operating a dual testing system. From 1973 until 1984, the military aptitude composite scores obtained from the STP were not identical to those obtained from enlistment testing. For the ASVAB 5, the differences affected only the Army, which gave an interest inventory to applicants. (The interest inventory was not included in the STP because some items had an excessive military flavor.) When ASVAB 2 was used in the STP, from 1973 until 1976, both the Army and Navy operated with a dual set of aptitude composites. The Air Force and Marine Corps were not affected because their classification batteries overlapped the ASVAB in the STP. The Army and Navy worked out procedures for dealing with the dual set of aptitude composites, and no one in the military testing community raised questions about the effects of the STP composites on predictive validity and the level of performance expected from recruits classified using the STP composites.

The dual classification system raises some questions about validity generalization from military to civilian occupations. The rationale for validity generalization is sound, and it has been confirmed through extensive analyses. The issue concerns what specifically is being generalized from the military composites to civilian occupations. Generalizing in terms of general mental ability is sound, and this practice has been the one supported through analyses. Generalizing from specific definitions of aptitude composites, which was implied when aptitude composites were reported in the STP, is not as clearcut.

The four occupational composites (mechanical, clerical, general, electronics) are common to all Services, but the particular tests in these composites sometimes vary from Service to Service and from time to time. For example, see the various definitions of the Mechanical composite used since 1973 shown in Table 11.

Relative to the Table 11 definitions of the Mechanical composite, how exact is validity generalization? If this set of composites can be generalized from military to civilian occupations, why does not one definition generalize from one Service to another? An argument advanced by the Services is that the job requirements in each Service are unique, and common composites simply would not work. If the Services do in fact have unique job requirements, and the same definition is not appropriate for all Services, then how can the validity of the Mechanical, or any other, composite be generalized from military to civilian occupations?

Generalizing aptitude composites across Services raises the specter of common composites, which infringes on the Service prerogatives to define their own aptitude composites as they see fit. Service prerogatives have carried the day in the classification arena, and they are likely to continue. To have common composites would require major policy decisions about how aptitude tests are used in selecting and classifying recruits. The discrepancy between the validity generalization argument and Service practices has not been a compelling reason for changing in the past; it probably will not be in the future either. Just as the place of aptitude testing in career exploration is ambiguous, so aptitude testing in the military services is subject to policy constraints and traditional practices.

In summary, the STP has had its ups and downs. The current CEP meets high professional standards for vocational counseling. The commitment by the Department of Defense to develop and maintain a quality STP is now institutionalized with the Testing Center staffed properly for the STP. From a technical perspective, the future for the STP looks promising. Issues of (a) how to report and use aptitude scores in career exploration, and of (b) the relationship of the STP to the military

selection and classification testing program remain to be resolved.

Table 11

Definitions of the Mechanical Composite by the Services and in the STP

Years 1973-1975

Army*	MK, EI, SI, AI, CM**
Marine Corps	AR, EI, SI, AI
Navy	WK, MC, SI
Air Force	TK, MC, SI, AI
STP	MC, 2AI

Years 1976-1984 (ASVAB 5)

Army	MK, EI, SI, AI
Marine Corps	MK, EI, SI, AI
Navy	WK, MC, SI
Air Force	MC, SP, AI
STP	MK, MC, AI

Years 1976-1980 (ASVAB 6/7)

Army	MK, EI, SI, AI, CM**
Marine Corps	MK, EI, SI, AI, CM**
Navy	WK, MC, SI
Air Force	MC, SP, AI
STP	none

Years 1980-1984 (ASVAB 8-14)

Army	NO, AS, MC, EI
Marine Corps	AR, AS, MC, EI
Navy	VE, AS, MC
Air Force	GS, 2AS, MC
STP	AR, AS, MC, EI

* The Army composite is based on ACB 73; the others on ASVAB 2 or 3.

** CM is an interest score in the mechanical area; the other symbols are defined in earlier tables in this Chapter 3. Although the definitions vary, they tend to have common elements of verbal or math and technical tests.

CHAPTER 4

NORMING AND SCALING MILITARY SELECTION AND CLASSIFICATION TESTS

Development of the World War II Score Scale

When the peacetime draft was initiated in 1948, a joint-service test had to be developed for screening potential recruits. The Armed Forces Qualification Test (AFQT) was formulated, calibrated, and introduced for operational use on January 1, 1950. The development of the World War II score scale is described by J. E. Uhlaner (Uhlaner, 1952, November), who was widely recognized as the father of the AFQT. The content of the first AFQT was based on the Army General Classification Test (AGCT), a test that measured verbal skills, arithmetic reasoning, and spatial ability. That decision was relatively easy because these types of items were known through numerous validation studies to have high predictive validity for the full range of occupational specialties. The difficult part was defining the population for constructing test norms.

The two tests widely administered during World War II were (a) the AGCT, given to over nine million men, and (b) the Navy General Classification Test (NGCT), given to over three million men. The two tests had been scaled separately. The decision was made in the late 1940s by the military testing community to combine the score distributions of the AGCT and NGCT to define the World War II mobilization population. But first the two tests had to be placed on a common score scale.

The entire Army and Navy classification batteries, which contained the AGCT and the NGCT, were administered in 1947 to a sample of 1,052 Navy recruits (ARI, 1949, April). The calibration of the AGCT and NGCT was embedded in a larger study to determine the similarity of the Army and Navy classification batteries with the intent of improving the coverage of both. The AGCT and NGCT scores were extracted from the classification batteries and calibrated to each other. The two score scales agreed closely in the mid and upper range (100 to 130 on the AGCT score scale), but below the mean, the two scales differed by up to one-fifth of a standard deviation, with the NGCT scores lower. The NGCT scores were adjusted to match the AGCT, and the two score distributions were combined. The total World War II score distribution included scores from both enlisted recruits and commissioned officers for all Services. This group defined the World War II mobilization population and served as the World War II reference population.

The AGCT and NGCT had different content, but the military testing community at the time decided to pool the score distributions anyway. The NGCT had 100 items composed of verbal analogies and vocabulary, while the AGCT had 140 items composed of vocabulary, arithmetic reasoning, and spatial ability (block counting). Reasoning ability in the NGCT was measured through words, whereas in the AGCT it was measured through numbers.

Content differences also occurred between the AGCT given to recruits during World War II and the AGCT given to the calibration sample in 1947. In 1947 the AGCT had four types of items: reading and vocabulary (56 items, 25 minute time limit), arithmetic reasoning (56 items, 35 minutes), arithmetic computation (56 items, 15 minutes), and pattern analysis (60 items, 20 minutes). The pattern analysis items replaced the block counting items in the original AGCT.

The version of the AGCT that was calibrated to the NGCT was forms 3a and 3b, introduced for operational use in 1945 and given to about 350,000 men (Staff, 1947). It was an improved version of the AGCT forms 1c and 1d that was widely used during the war. The procedures for defining the World War II reference population and calibrating the AFQT were reviewed and approved by a panel of testing and statistical experts from government and private industry (Uhlaner, 1952, November).

The reason that the AGCT and NGCT score distributions were combined, even though the tests had different content, was that the focus of the military testing community at that time was on producing an acceptable test in time to meet the schedule imposed by initiation of the peacetime draft. The military testing community in the late 1940s was faced with the real problem of scaling the first AFQT and having it ready for use on January 1, 1950. At that time there was no other readily available population of examinees for developing the AFQT score scale, and omitting either Navy or Army recruits from the population would have been more serious than combining the score distributions for two tests that had different content. The testing community in the 1940s looked for a reasonable and feasible solution, and the recruits from World War II were the obvious choice.[19]

Once the World War II reference population and score scale were defined, they became the standard for calibrating the AFQT and Service classification batteries. The conversion from test raw scores to standard scores and percentile scores was fixed, and it served the military testing community for three decades. For purposes of defining the score scale, the total distribution of the AGCT and NGCT scores served well, and over the years it provided a stable basis for setting qualification standards.

The AFQT percentile score scale had 101 points instead of the usual 100. A score of zero was assigned to people who could only sign their names and answer a few questions. The AFQT score category V was said to contain the bottom 10 percent of the mobilization population (scores 0 through 9), and the common statement was that the bottom 10 percent of the population was not qualified to serve. The top score on the AFQT was 100, and 8 percent of the population scored in category I, percentile scores 93 through 100.

In the 1980s, the AFQT score scale was revised to contain only 100 points, with the score of zero deleted. For administrative convenience, the top score was set at 99, which meant that only two columns had to be reserved for the AFQT score in the automated data systems. The AFQT scores traditionally have been defined as the percentage of people that scores at or above each score point; by this definition, the top score is 100, but instead it is called 99. Thus, in the current AFQT score scale, category I includes 8 percent of the population, and category V contains 9 percent.

A questionable decision was made in the 1950s when the content of the AFQT was changed from the original in 1953; concerns about the effects on the score scale did not prevent the changes to test

---

[19]The recruits actually used to compute the score distributions during World War II were samples of input during 1944; hence, the World War II scale is sometimes referred to as the 1944 Score Scale.

content. When form 3 of the AFQT was introduced in 1953, a new type of item was added to measure knowledge of tool functions. The reason for adding the Tool Knowledge items was to reduce literacy requirements and decrease the correlation between the AFQT and years of education. This version of the AFQT, and the next two sets (forms 5 and 6 introduced in 1956 and forms 7 and 8 introduced in 1960) continued to be calibrated to the AGCT, and the percentile scores were interpreted as representing the World War II mobilization population. No one conducted a study to evaluate the effects of changing test content on the score scale.

Calibrating the AFQT and Classification Tests in the 1950s and 1960s

The procedures for calibrating the AFQT evolved gradually until forms 7 and 8 were developed in the late 1950s; these were the last separate AFQT forms to be developed. The calibration of the AFQT 7/8 involved two different samples that differed in ability.[20]

- Recruits from all Services took the AFQT 7/8 and the reference test, form 1c of the AGCT; their scores were used to determine the score scale above a percentile score of 30.

- Registrants for the draft first took the operational form of the AFQT 5/6; those whose operational scores were 30 or below then took AFQT 7/8 and the reference test, form 1c of the AGCT.

The two samples were tested under similar conditions, with comparable motivation to take both the new AFQT and reference test. The scores from the recruits and registrants were combined to produce the score distributions used for calibrating the new AFQT to the World War II score scale (Bayroff & Anderson, 1963, May).

Whenever a new form of the AFQT was introduced, the score distributions were carefully examined to make sure that no abrupt changes in the proportions of examinees in the various score categories occurred. No problems with the calibrations were found, and the tests continued in use until they were replaced.

In contrast to the procedures for calibrating the AFQT, which were reasonably rigorous, the procedures for calibrating the classification tests by each Service did not necessarily result in accurate score scales. (The intent in calibrating classification tests was to maintain approximately the same meaning of the test scores when new forms were introduced and to have a reasonable tie to the World War II reference population.) In a typical calibration of the Service classification tests, the new tests would be given to samples of recruits, which means that the low end of the score distribution was either absent or poorly represented. As a rule, operational AFQT scores or operational classification test scores would be used as the reference for anchoring the new classification tests to the existing score scale. The sample of recruits would be stratified on their operational scores to ensure 10 percent of the sample in each decile; that is, the number of cases in each decile was weighted to produce the desired percentage. Typically, cases at the top and bottom of the distribution would have

[20]Applicants for enlistment were not tested along with the registrants; the applicants had been prescreened with a short aptitude test. In the days of the draft and draft-induced volunteers, applicants for enlistment generally were not given extra testing burdens.

large weights, and those in the mid-range would have weights less than one. The weighted distributions of scores on the various classification tests would then be converted to the score scale used by a Service. The Army tests had a mean of 100 and standard deviation of 20; the Navy tests had a mean of 50 and standard deviation of 10; the Air Force did not convert test scores to a standard scale prior to computing aptitude composites. (The term often used to describe the calibration of classification tests was "standardization," which implied that the raw scores were converted to a standard scale. No great precision was claimed for the score scales of classification tests, and none was required as it was for the AFQT.)

Because the uses of the AFQT and classification tests were different, they required different degrees of accuracy in the meaning of the score scales. The AFQT was, and continues to be, used for making selection decisions where standards are set in terms of percentages of the mobilization population that would be qualified for the outcome. The outcome might be eligibility for enlistment, or for a contract or a bonus. The AFQT selection standards have been from the beginning stated in terms of the population of potential recruits that would qualify, or conversely, not qualify. If the AFQT score scale was in error, then the intent of the qualification standard was subverted. The care taken to ensure that the AFQT was properly calibrated to the reference population was necessary, and military manpower managers, and the Congress, had much confidence in the accuracy of the scores.

Calibrating the ASVAB 5/6/7 in the 1970s

The formal date for initiating the ASVAB as the first joint-service selection and classification battery was May 1974, through a memorandum from the Assistant Secretary of Defense for Force Management and Personnel (ASD-FM&P). The joint-service battery that was in use for the Student Testing Program (STP) was to serve as the model for the joint-service enlistment test. A working group, called the ASVAB Working Group, with technical and policy representatives from each Service, was established to develop the ASVAB for enlistment purposes. The development process included making decisions about (a) test content, (b) calibration of the battery, and (c) uses of the test for making personnel decisions.[21] The work was scheduled to be accomplished in a little over one year; it was actually completed in about 20 months, resulting in an implementation date of January 1, 1976. (For perspective, the time required in the 1980s to produce replacement forms for the ASVAB was over four years, more than twice as long.) There was little precedent for developing a classification battery in the joint-service arena, and the times were hectic for all participants.

The ASVAB 5/6/7 was to be placed on the World War II score scale; this would retain the traditional meaning of the test scores and would enable use of existing qualification standards. The scaling procedures had severe constraints placed on them by operational demands. First, the tests had to be ready for administering to all applicants for enlistment by January 1, 1976. Second, the scaling could not interfere unduly with the accessioning process, lest applicants become disgruntled and decide not to enlist. Given these constraints, the Working Group searched for procedures that might accomplish the scaling in a timely and acceptable manner.

The design for calibrating the ASVAB in 1975 was consistent with the traditional procedures for calibrating the AFQT. The calibration of the ASVAB was intended to be more rigorous than traditionally employed for classification tests.

---

[21]The development of the ASVAB 5/6/7 is described in Appendix C of this report.

- Army applicants took the operational form of the AFQT from the Army Classification Battery (ACB 73) as the reference test and the ASVAB 5/6/7; those whose operational AFQT scores were 50 or below were used to calibrate the bottom half of the score scale.

- Air Force and Navy recruits took the ASVAB 5/6/7 and the ASVAB form 3 as the reference test; their scores were used to calibrate the top half of the score scale.

The scores from the applicants and recruits were then combined to produce a full-range sample for placing the ASVAB 5/6/7 on the World War II score scale.

In principle, the procedures should have produced a satisfactory scaling, providing all went well. Similar procedures had been used previously with the AFQT, and even less precise procedures with Service classification batteries, and they had worked before. But the score scale for the ASVAB 5/6/7 was seriously inflated, which resulted in hundreds of thousands of erroneous personnel decisions during the late 1970s.

Impact of the ASVAB Miscalibration

The ASVAB miscalibration was caused by a serious error in converting raw scores (number of items correct) to percentile and standard scores. The converted, or scaled scores, were too high compared to the traditional meaning of the score scales used for military selection and classification tests. Consequently, the ability of the people applying for enlistment was overestimated.

Within two months after the ASVAB 5/6/7 was introduced in 1976, the Navy found that the percentage of recruits in the above-average range on the AFQT showed a large increase compared to earlier months when the Navy was using its Basic Test Battery. The Services quickly conducted analyses and found that the score scale was in error, and adjusted conversion tables were introduced in July 1976. The adjustments to the conversion from the AFQT raw scores to percentile scores were limited to the top half of the score scale, and virtually no changes were made to the bottom half; neither were the conversion tables used for computing aptitude composite scores changed.

However, the AFQT percentile scores obtained from the ASVAB 5/6/7 were particularly inflated in percentile scores 21 through 49. Recruits with an AFQT score reported as 50 actually should have received a score of 41, and recruits with a reported score of 46 should have gotten a score of 31, the bottom of the AFQT score category III (scores from 31 through 64). All recruits with the original AFQT scores of 31 through 45 were at that time counted as being in category III, when in fact their correct scores would put them in category IV (scores from 10 though 30), often referred to the group with marginal ability. The original AFQT percentile score of 31 should have been a 17. In the late 1970s, the minimum qualifying AFQT score for enlistment was 16; recruits who came in with these scores had corrected AFQT scores of 12. The fact that no people in category V (AFQT percentile scores 1 through 9) were accessioned was the only good news as the story of the ASVAB miscalibration began to unfold; the basic rule was that no military recruits could be in category V.

Inflated percentile scores at critical points on the Armed Forces Qualification Test (AFQT) score scale for the ASVAB 5/6/7 are shown in Table 12. The adjusted scores shown were those remaining in the score scale after July 1976.

Table 12

ASVAB 5/6/7 AFQT Percentile Scores

Adjusted Scores*	Corrected Scores**
10	9
16	12
21	14
28	16
31	17
37	21
46	31
50	41
55	50
60	58
65	65
77	80
80	82
93	91
94	93

* scores actually used during 1976-1980 for personnel decisions
**corresponding scores accurately calibrated to the World War II score scale

During the time (from January 1, 1976 until September 30, 1980) that the adjusted, but still inflated, score scale for the ASVAB 5/6/7 was in use, over 400,000 recruits, about 30 percent of the total, were in the AFQT score category IV. In 1980, the Army thought that only 10 percent of the recruits were in category IV, and so reported to Congress, when in fact according to the correct AFQT score scale, over 50 percent were in category IV (Laurence & Ramsberger, 1991).

Another figure that shows the impact of the scaling error is that over 350,000 recruits during this five-year span would not have been eligible to enlist under the prevailing selection standards, which included both the AFQT and the aptitude composite scores, if the ASVAB 5/6/7 had been properly scaled (Eitelberg, Laurence, & Waters, 1984, September). To help compensate for the inadvertently lowered qualification standards arising from the miscalibration, the Army raised classification standards from 1977 until 1980 for over 50 occupational specialties because of excessive failure rates (Maier & Truss, 1983, March).

The effects of the ASVAB miscalibration on the quality of recruits illustrates a fundamental fact about testing programs used to help make personnel decisions: score scales define policy. Manpower managers establish qualifying standards based on expected performance and relative standing in the population; standards are stated in terms of test scores, either percentile or standard scores, and these scores are defined by the conversion of raw scores to scaled scores. If the conversion is accurate,

then there is congruence between stated policy about qualification standards and effects in the population of potential recruits. But if the conversion is not accurate, then personnel decisions based on the reported scores result in consequences not anticipated by the managers. This relationship between policy decisions about standards and the accuracy of converting raw scores to scaled scores became painfully obvious to manpower managers in 1979 and 1980 as the ASVAB miscalibration episode unfolded.

Reasons Why the ASVAB Miscalibration Occurred

If similar, or even less rigorous procedures for calibrating the AFQT and classification batteries worked satisfactorily before, why did they not work for the ASVAB in 1975? In brief, the answer seems to lie in the changed conditions that worked against the design adopted by the ASVAB Working Group in 1974 and 1975.

Noteworthy were the differences between (a) registrants for induction, as were used for the AFQT in the 1960s, (b) recruits, as were used for scaling the top half of the score scale in the 1970s, and (c) applicants for enlistment tested at examining stations, as were used for scaling the bottom half of the scale in the 1970s.

When registrants appeared at examining stations, they were a captive audience and essentially had to do as they were told. When they were told that they would have an additional 2 1/2 hours of testing, they had no recourse. Because both the reference test (AGCT) and the new AFQT were experimental tests, the level of motivation to take either tended to be the same; there was little effect on the accuracy of the calibration for the new AFQT.

A similar argument applied to the test scores obtained on samples of recruits. Recruits, even more than registrants, were captive audiences who did as they were told. When they were given extra testing, they simply took it as part of their assignments. Again, any poor motivation affected both the new and reference tests equally, and the calibration was expected to be accurate.

However, in 1975 when the ASVAB was calibrated on samples of applicants at examining stations, the testing environment was radically different from that for registrants. In 1975, applicants were a precious commodity that had to be treated with respect and consideration, lest any of them become disenchanted and walk away. Extra testing was feared to be an undue and onerous burden, and its negative effect had to be minimized. Even if individual applicants would not object to the extra testing, they had protectors in their recruiters. A reasonable decision by the Working Group was to minimize the burden on applicants. In hindsight, the way that the decision was carried out by the examining stations worked to inflate the score scale for the ASVAB. Some reasons why the miscalibration occurred will be discussed, based on work by Maier and Truss (1983, March).

A key decision was to give the new tests only to Army applicants with operational AFQT scores below 50. This decision follows directly from the earlier practice of testing only those registrants with AFQT scores below 31. Crucial differences, however, arise that can be discerned in hindsight with over 15 years of experience in scaling new forms of the ASVAB on applicants. One of the first questions is how test administrators would know whom to include in the sample. Because the defining characteristic was applicants with AFQT scores below 50, and applicants with AFQT scores of 50 and above would not count in meeting quotas of examinees, a reasonable procedure for test administrators was to give the ASVAB only to those applicants who were known to have AFQT

scores below 50. The research design called for counterbalanced order of administering the ACB 73 and the ASVAB 5/6/7, but there was little incentive for the test administrators to give the new ASVAB before the operational AFQT scores were known. In this way, only those Army applicants with AFQT scores below 50, who helped the examining stations meet their quotas of examinees, would be given the extra testing.

A reasonable procedure from the examining stations' perspective was to give the tests to Army examinees who were being processed for shipment to training centers and who therefore had taken the ACB at an earlier time. Their AFQT scores (from the ACB) were known, and the people had more time for taking extra tests. Army applicants who had failed to qualify for enlistment would tend to have been excluded from the extra testing with the ASVAB 5/6/7. The effect of excluding people with low scores on the reference test was to lower the raw scores on the new test at the bottom of the scale.

One other effect that would contribute to the inflated percentile scores on the new test was that many applicants were coached on the operational AFQT. The effect of coaching was of course to raise the AFQT raw scores above their true value. To the extent that the AFQT raw scores were raised by coaching, the effect on the calibration was to inflate the score scale.

Perhaps the decision with the greatest consequence was the one to use the operational ACB 73 test scores as the reference for calibrating the ASVAB 5/6/7. To the extent that people were tested with the new ASVAB after taking the operational tests, examinees would tend to answer fewer items correctly on the new tests than normally would be the case. One big factor is motivation. Applicants ordinarily do their best on the operational tests because they are applying for a job. When they took the new ASVAB, many knew that the scores would not affect their employment opportunities, so they tended not to try as hard. Also, they were fatigued by the time they took the extra tests. The ACB 73 and the ASVAB 5/6/7 each required at least three hours of testing time, which meant that some examinees took six hours of testing at a time. As is well known, test scores tend to go down on tests taken later in a session. Fatigue coupled with poor motivation would have tended to depress raw scores on the ASVAB 5/6/7.

The effects of (a) poor motivation, (b) fatigue, and (c) excluding people with low scores on the reference test, all worked in the direction of depressing the raw scores on the new ASVAB. When these depressed raw scores were calibrated to the operational AFQT scores, which did not suffer from the same depressing effects, the outcome was inflated percentile scores referenced to the World War II score scale.

The score inflation at the high end of the scale, detected by the Navy within two months after the ASVAB 5/6/7 was introduced, arose independently of the inflation at the low end. Recall that the upper half of the score scale was calibrated on samples of Air Force and Navy recruits who took the ASVAB 3 as the reference test. Differences in scoring formulas can account for the score inflation. The scoring for the ASVAB 3 involved a correction for guessing: one-third of the number of wrong answers was subtracted from the number right. The ASVAB 5/6/7 was scored as the number of right answers, with no correction for guessing. Apparently when the ASVAB 3 was scored for the sample of recruits used in the calibration study, the correction for guessing was not included; the scores in effect were inflated. Then, when the ASVAB 5/6/7 was scaled to the unduly high ASVAB 3 scores, the resulting score scale was inflated.

Inflation of the score scale arises when raw scores on the new test are too low (or raw scores on

the reference test are too high) in the calibration sample compared to the scores that would be obtained under proper testing conditions. To illustrate these effects, a hypothetical example will be given. Suppose that the raw scores on the new ASVAB which was being calibrated were systematically lowered by five raw score points in the calibration sample, but the scores on the reference test were not lowered systematically. Say that under proper testing conditions a raw score of 25 should be converted to percentile score of 40, and a raw score of 20 to a percentile score of 30. Then in the calibration sample, a raw score of 20 would be converted to a percentile score of 40 instead of 30. When the new test was introduced for operational use, the raw scores would not be systematically lowered, and people would tend to have higher raw scores. The operational score scale would be inflated by 10 percentile score points in this example because of the error in converting raw scores to percentile scores arising from bad testing conditions in the calibration sample.

Using the old research design to collect calibration data was doomed to creating a large inflation of the ASVAB score scale in the low end because testing conditions for applicants were sufficiently different from those for registrants.

The Unfolding of the ASVAB Miscalibration

The inflation of the ASVAB score scale was severe, and the inflation at the low end was not fixed for almost five years. When the ASVAB miscalibration was reported in 1979 and 1980, questions naturally arose about the motivation and competence of the military testing community. At that time no one had a clear understanding of all the things that went wrong and why decisions and actions, and inactions, were taken as they were. Some of the pieces will be put together here to help provide insight about the course of events during the turbulent times of the late 1970s for the military testing community.

The ASVAB Working Group in 1974 came out of the same tradition that produced the World War II score scale and changes to the AFQT. The Working Group attempted to find reasonable and feasible solutions to the calibration problems at the time. Unfortunately, their solutions did not work as well as those adopted for the AFQT, and errors arose.

The Marine Corps in 1978 made the first serious charge that the score scale in the bottom half was inflated (Sims, 1978, April), and this led to a request for further data. The charge was repeated in 1979 based on better data (Sims & Truss, 1980, April). A follow-on study was done in the summer of 1979 to evaluate the earlier results, and the new analysis essentially confirmed the 1979 results presented by the Marine Corps (Maier & Grafton, 1980, August).

A review of the studies is in order to help put the concerns in perspective. The Marine Corps data presented in 1978 was based on a sample of Marine recruits who were retested with an earlier version of the Army Classification Battery (ACB 61). The results were suspect for two reasons. The most obvious was that the sample was restricted to Marine recruits, which meant that relatively few people had AFQT scores below 30, and none had scores below 20. No one knew how well the results would generalize to the full range of applicants. The other reason for questioning the results was that the accuracy of the score scale for the ACB 61 was not known. As discussed earlier, the scaling of Service classification batteries, especially at the low end, was not as precise as the testing community demands today. A legitimate response by manpower managers was to request better data.

However, when the Marine Corps responded in 1979 with better data and the inflation was

confirmed, the amount of inflation was different in the two studies. The new study was still based on samples of Marine recruits, but the reference test for evaluating the scaling of the ASVAB 5/6/7 was form 7 of the AFQT, which was known to be scaled accurately.

Later, in the summer of 1979, a study was designed by the ASVAB Working Group to evaluate the scaling of the ASVAB 5/6/7 with a sample of applicants that included the full range of ability. The AFQT 7 was given as the reference test to the sample, along with either form 6 or 7 of the ASVAB. The data were collected in July 1979, and by September 1979, a preliminary analysis essentially confirmed the amount of inflation reported by the Marine Corps earlier in 1979 (Maier & Grafton, 1980, August). However, the design of the study of applicants was open to some of the same criticisms made against the calibration of the ASVAB 5/6/7. In this study, operational ASVAB raw scores were calibrated to the AFQT 7; the reference test was given under experimental testing conditions, and it was calibrated to a test given under operational testing conditions. Motivation to take the AFQT 7 may have been low because it did not count for enlistment. Also, some applicants were coached on the operational ASVAB. The effects of these improper testing conditions, to the extent they were operating, would have been to show more inflation in the original scaling of ASVAB 5/6/7 than would have been true under proper testing conditions. Analysis of the data at the time showed that poor motivation and coaching had only small effects on the calibration, but critics could fault the design. In late 1979, as the story of the ASVAB miscalibration was unfolding, no one knew with certainty just how much inflation there really was in the ASVAB scores.

The answer given to manpower managers by the testing community in early 1980 about the lack of shift in the score distributions at the low end was in terms of coaching on the ACB 73. The effects of the coaching were to reduce the percentage in category IV and raise the percentage in category III. When forms 6 and 7 were introduced on January 1, 1976 for testing applicants, coaching was not immediately possible because recruiters did not have access to the test items. Scores on the ASVAB 6/7 therefore would be lower than on the ACB 73. No one knew at that time how much coaching was taking place on ACB 73 and how much score distributions were affected. The testing community hypothesized that the effects of coaching on the ACB 73 and score inflation on the ASVAB 6/7 balanced each other and would result in no systematic shift of scores at the low end. But, no one could speak with much certainty.

Fortunately, a more definitive study to evaluate the accuracy of the ASVAB 5/6/7 score scale was carried out in 1980. A contract was let to Educational Testing Service for administering the AFQT portion of the ASVAB 7 along with the AFQT 7 to a sample of high school students. The results from this sample confirmed the score inflation; the score inflation found for the sample of applicants was also found for high school students. The amount of inflation was in close agreement for the AFQT percentile scores 1 through 30, and above that point, the scaled scores in the high school sample were lower than in the applicant sample (Boldt, 1980, August). Because the results for high school students, who were not coached on either test or were differentially motivated, agreed with those for the applicants at the low end, the inflation of the ASVAB 5/6/7 score scale in the bottom half was confirmed. (The disagreement between the two samples at the upper end of the score scale was of little concern to the testing community. Various samples of military examinees showed agreement about the correct conversion of raw scores to percentile scores in the upper end, and the divergence of the high school students could be attributed to systematic differences in the samples.)

All the studies on the scaling of the ASVAB 5/6/7 were reviewed by a panel of eminent testing psychologists, Drs. Melvin Novick, Robert Linn, and Richard Jaeger, in 1980. The panel agreed that (a) the ASVAB score scale was inflated, and (b) the results for the sample of applicants most nearly

represented the correct scaling. The question of whether or not the ASVAB 5/6/7 score scale was inflated, and by how much, at long last was resolved.

After close to five years of lowered qualification standards, the score scale was restored to its intended meaning in July 1980. This allowed applicants with scores from the ASVAB 5/6/7 to enter service with their correct test information.

Aftermath of the ASVAB Miscalibration

Once the studies were deemed conclusive, the response of manpower managers to the ASVAB miscalibration was swift and decisive. The first response was to direct that the score scale be fixed as soon as possible. New forms of the ASVAB--forms 8, 9, and 10--had been under development since 1977, and by 1979 the items had been tried out and were ready for assembly into test booklets and calibration. A decision had been made in late 1979 by manpower managers to continue using the inflated scores until the new forms were ready for use. The testing community promised the managers that the new forms would be ready on October 1, 1980, about one year later. During the year, the following actions were completed for the new forms of the ASVAB.

- Test booklets were printed for the calibration.
- Data were collected in January and February 1980 on samples of applicants and recruits.
- Conversion tables were prepared by May 1980.
- Computer programs were written to handle the new tests and aptitude composites.
- Testing materials were distributed to the 1,000 or so testing stations.

The new tests on the correct score scale were introduced on October 1, 1980, and a state of near normalcy was achieved.

The second response by manpower managers was to address the question of qualifying standards. An immediate problem was that the qualifying standards actually in effect on September 30, 1980 would have been raised substantially on October 1, 1980 when the correct score scale was introduced. The minimum qualifying AFQT score of that time was 16 for the Army and 21 for the other Services. With no other changes other than introducing the correct score scale, the standards as of October 1, 1980 would have been raised to the AFQT scores of 28 for the Army and 37 for the other Services. At that time the Services were having trouble meeting their recruiting goals even with the lowered standards then in effect. OASD-FM&P requested the Services to lower their qualifying standards to the levels actually in effect for the ASVAB 5/6/7. All Services except the Marine Corps complied by lowering their standards.[22]

The long term response to the quality of enlisted accessions in the 1980s by the Department of Defense and the Congress was to improve quality by increasing recruiting resources and military pay.

---

[22]The Marine Corps maintained their standards and said that they would recruit to those standards. The Marine Corps did meet recruiting goals in 1981 even with the effectively higher standards.

The effects were highly successful, as the quality of enlisted accessions during the latter 1980s was at an all time high. Actual standards in the 1980s kept creeping up in response to the availability of quality recruits. The tradeoff between quantity and quality in the setting of qualifying standards at first was to maintain quantity by lowering quality, but then as the recruiting market improved, quality was raised with no loss of quantity.

The third major response by manpower managers was to exercise closer supervision of the military testing community. A steering committee, composed of flag officers responsible for military manpower policy and the commander of the Military Enlistment Processing Command (MEPCOM), and chaired by the Director for Accession Policy (OASD-FM&P-AP), started meeting regularly to review the testing program. The panel of testing psychologists (Novick, Linn, and Jaeger) became the core of the first Defense Advisory Committee for Military Personnel Testing, or the DAC as it has been known. The DAC has remained in continuous existence to provide technical review and advice, passing judgment on virtually every aspect of the testing program. The public scrutiny has had an immeasurable effect on improving the technical rigor for calibrating military aptitude tests.

The fourth major response of the manpower managers was to direct that a new reference population be tested with the ASVAB and that a new score scale be constructed. This action resulted in the 1980 score scale.

Development of the 1980 Youth Population Scale

Until the ASVAB miscalibration was reported to manpower managers, they did not ask many questions about the technical side of aptitude testing. They assumed that the scores were reasonably accurate for predicting performance and showing relative standing in the current population of potential recruits. With the unfolding of the ASVAB miscalibration, however, managers became uneasy about the meaning of aptitude test scores and how that meaning was determined.

The first round of questions concerned the meaning of test scores in the population of potential recruits. Managers had difficulty comprehending how scores on the current classification battery in 1980 could be tied to the World War II reference population. When they realized that the military testing community did not know how the ASVAB scores were distributed in the then current youth population, they were incredulous. The response was to initiate immediate action to obtain a new score scale for the ASVAB that reflected relative standing in the youth population of the 1980s.

Developing a score scale based on a nationally representative sample of American youth was a major undertaking, and the military testing community at that time had no experience in designing such a study. Fortunately, the Department of Defense was at that time cooperating with the Department of Labor in studying the behavior of American youth in the labor market. Arrangements were soon made with the National Opinion Research Center (NORC) to administer the ASVAB to a sample of youths aged 16 through 23.[23]

Scaling the ASVAB with a nationally representative sample of youth was expensive, even with a

[23]More information about the design of the study and characteristics of the sample is contained in the report *Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery* (OASD-FM&P, 1982).

sample in place. The cost in 1980 funds was estimated to be about $3 million; OASD-FM&P lacked the money to fund a manpower project of such scope. However, the money problem was resolved quickly and effectively. The then Assistant Secretary of Defense (FM&P) called a meeting of the Service military manpower chiefs to inform them about the need to conduct a study for obtaining a new ASVAB score scale; given the dire consequences of the inflated ASVAB scores for the quality of recruits, there was no objection to the need for the study. At the end of the meeting, the Assistant Secretary told the Service chiefs that they would be required to provide the funding. (The formula for dividing costs among the Services in joint-service manpower projects typically has been in proportion to the percentage of recruits accessioned each year; usually the Army provides about 40 percent, the Navy and Air Force about 20 to 25 percent each, and the Marine Corps the remainder.) Together the Services sent the $3 million to OASD-FM&P, and the study was initiated in early 1980.

Administration of the ASVAB Form 8a was conducted in the summer and fall of 1980. Local test administrators were hired by NORC and trained to give the ASVAB, and the testing was done in locations and at times convenient for the examinees and administrators. The data from this administration was examined by Drs. Darrell Bock and Robert Mislevy (Bock & Mislevy, 1981); they concluded that the ASVAB tests and the data collection procedures provided a suitable basis for estimating the aptitudes of American youth and computing a new ASVAB score scale. The examinees, both males and females, were of ages 16 through 23 at the time of testing; the group aged 18 through 23 was designated as the 1980 Youth Population, and their scores were used for constructing a new score scale. Development of the 1980 score scale with the 1980 Youth Population is described more fully by Maier and Sims (1986, July).

The 1980 score scale solved the problems of interpreting the ASVAB scores for the population of potential recruits. The ASVAB raw scores were first placed on a standard score scale with a mean of 50 and a standard deviation of 10. The test standard scores were then added and rescaled to produce scaled scores for the AFQT, expressed as percentile scores, and aptitude composite scores for the Air Force, also expressed as percentile scores, and the Army and Marine Corps, expressed as standard scores. No additional scaling was done for the Navy aptitude composites.

The two big advantages of the 1980 score scale over the World War II score scale were (a) the reference population was nationally representative, and (b) all tests and composites of new forms could be equated to the reference test, the ASVAB 8a, because the entire battery was administered to the reference population. Therefore, all ASVAB scores in the 1980s accurately reflected relative standing in the current population of potential recruits.

Equating the ASVAB in the 1980s

The distinction between scaling and equating tests may be instructive at this point. Strictly speaking, military aptitude tests prior to the 1980 score scale could only be calibrated, and not equated, to the World War II score scale. Equating tests requires that they measure the same content equally well; that is, they need to have the same types of items, be of the same length, have the same time limits, and be administered under the same testing conditions. The new forms of the ASVAB developed in the 1980s generally met these requirements, and they can be said to be equated to the ASVAB 8a, the rereference test. Tests on the World War II score scale did not meet these requirements. The AGCT administered during World War II was used as the reference test for calibrating the AFQT and Service classification batteries. Because the content of the AGCT was different from that of the AFQT and classification batteries, equating was not possible.

Calibration or scaling means to place a test on the same score scale as another test, such that they both have essentially the same score distributions. Tests calibrated to the same score scale can be used for setting qualification standards in which a specified proportion of the population is selected out and the remainder selected in. Equated tests are interchangeable and people would get the same scores on either test, except for errors of measurement. On tests that are only calibrated to each other, people are likely to get different scores because they do not measure the same things. The equating of the ASVAB scores since the 1980 score scale was introduced permits more precise comparisons of scores across time and test forms than was possible previously.

The 1980 score scale was thought to be ready for introduction on October 1, 1983, along with forms 11, 12, and 13 of the ASVAB, which were to replace forms 8, 9, and 10. However, in the summer of 1983, the Marine Corps reported that the scores for speeded tests (Numerical Operations and Coding Speed) appeared anomalous; whereas females typically have higher scores on speeded tests than males, the scores of the 1980 Youth Population, which consisted of females and males, were consistently lower than samples of male military recruits (Sims & Maier, 1983, June). Introduction of the 1980 score scale and the ASVAB 11/12/13 were delayed until the issue of the anomalous scores could be resolved.

The Air Force examined the answer sheets used by military recruits and the 1980 Youth Population and hypothesized that the differences on the speeded tests could be attributed to the size and shape of the response spaces on the answer sheets. The answer sheets used with the 1980 Youth Population had circles for recording responses, whereas the answer sheets used with recruits had rectangles; perhaps circles took longer to fill in than rectangles. A study was conducted by the ASVAB Working Group in which recruits filled in both kinds of answer sheets, and the differences on speeded test scores for this sample matched the differences found between the 1980 Youth Population and military examinees. An adjustment to the speeded test scores was developed, and the scores of military examinees who filled in rectangles were made comparable to the scores of the 1980 Youth Population who filled in circles (Wegner & Ree, 1985, July).

After the 1980 score scale was fixed, it was introduced on October 1, 1984, along with the ASVAB 11/12/13. However, another problem with the test scores was found shortly after introduction. Examination of the AFQT score distributions showed a sudden drop of about 2 percentile score points at the median. Again the Working Group examined the testing materials for possible causes. The culprit this time was the print format for the speeded tests. The spacing of test items was different in the ASVAB 11/12/13 compared to the ASVAB 8a, and the differences made the new tests more difficult to read. Some recruits who were truly qualified for bonuses and other desirable outcomes, were denied these outcomes because of the low AFQT scores they received. It took over 18 months to discover the error in the AFQT scoring, determine the reason for it, and prepare new conversion tables.

Manpower managers were chagrined that the error occurred in the first place and especially that it took so long to fix. Manpower policy representatives from the Services were directed by the Director for Accession Policy in OASD-FM&P to develop procedures that would help forestall similar problems. The committee prepared a report in May 1986 (ASVAB Working Group, 1986, May), and significant improvements in the scaling of the ASVAB were formulated.

The most significant change in the procedures recommended by the panel was the addition of the procedure to verify the accuracy of the score scale through an Initial Operational Test and Evaluation (IOT&E) *before* the introduction of new forms for continued operational use. (Prior to the

recommendation, an IOT&E for new forms of the ASVAB was conducted after the tests had been introduced, so any errors in the initial score scale would affect personnel decisions for extended periods, as with the ASVAB 11/12/13.) Following the recommendation, an IOT&E for each new set of forms has been conducted for two or three months, after which the tests are withdrawn and the accuracy of the initial score scale analyzed. After the score scale has been found to be correct, the new forms are introduced for continued operational use. Under the new procedures, the IOT&E serves as a check on the initial scaling of new forms; if the scale were in error, the effects on personnel decisions would last for only the two or three months that the IOT&E is carried out. The risk of erroneous personnel decisions because of a faulty score scale is minimized.

Problems with the speeded tests persisted, so in 1989 the Numerical Operations test was removed from the AFQT, and the Mathematics Knowledge test was substituted for it. (However, Numerical Operations continues in the ASVAB and is used in some aptitude composites.)

The focus of attention by the Working Group in the 1980s was on accurate scaling of the ASVAB. New forms of the ASVAB, when introduced in 1984 and 1989, had exactly the same appearance as the ASVAB reference form, 8a. The cloning was necessary to maintain the accuracy of the score scale. No deviations in print format, directions to examinees, or content of the test booklets were permitted by the Working Group. Even errors in example items were not corrected. The specter of the ASVAB miscalibration still hung heavily over the military testing community, and the focus on the accuracy of the score scale is well placed.

A minor issue arising from the use of 8a as the reference form was in the initial scaling of forms 18 through 22 in 1990 and 1991. The intercorrelation among the tests of form 8a was high; some tests, notably Auto and Shop Information and General Science, tended to have item content that could be learned through experiences of everyday living rather than specifically through formal training in these content areas. The new forms developed in the late 1980s tended to have fewer items that could be learned through everyday experience and to have content that was more specific to these content areas. The correlation of the Auto/Shop Information and General Science tests tended to be lower with other tests in the ASVAB in the new forms compared to form 8a. The result was a slight contraction of the score scale that could affect a small number of people with scores in the range of minimum qualifying scores for aptitude composites that include these two tests. In the early 1990s, when the rigor of the score is emphasized, the effects of the lower intercorrelation on the score scale proves troublesome. In earlier times, the lower intercorrelation would have been viewed as a boon because of increased differential validity, and the effects on the accuracy of the score scale would have gone largely unnoticed.

The rigor in maintaining the accuracy of the score scale is new in military aptitude testing. Prior to the ASVAB miscalibration episode, test content was changed with relatively little concern about effects on the score scale. The military testing community in earlier times was more concerned about test validity, or the accuracy of personnel decisions, than about the strict accuracy of the score scale.

Validating the ASVAB: the Job Performance Measurement Project

In addition to having concerns about the accuracy of the ASVAB score scale for setting qualification standards, manpower managers in 1979 and 1980 once again started probing more deeply into the validity of test scores for making personnel decisions. One position taken by some people was that the Services had been living with the erroneous test scores for close to five years, and

the jobs were still getting done. The conclusion drawn by critics of testing was that the prominence of aptitude tests in making personnel decisions should be downgraded, if not eliminated all together. Most managers at the time wanted to keep using the aptitude tests, which of course happened, but they wanted better information about the predictive validity of the ASVAB.

Early in 1980, efforts to validate the ASVAB against measures of job performance were initiated by OASD-FM&P. These efforts culminated in the Job Performance Measurement (JPM) Project. The JPM Project began in the early 1980s, and a new joint-service committee, the JPM Working Group, chaired by OASD-FM&P-AP, was formed. In addition, a contract was let by the Department of Defense to the National Academy of Sciences (NAS) to form a committee of testing and personnel experts to provide technical oversight of the JPM Project. The JPM Working Group and NAS Committee are still in existence in the early 1990s. Annual reports to the House Committee on Appropriations have been submitted by the Department of Defense.

The early focus of the JPM Project was to develop measures of job performance. The NAS Committee decided that the benchmark of job performance was a hands-on test of job performance. All Services developed such measures, and supplemented them with written tests of job knowledge and perhaps supervisors' ratings of job performance. The annual reports to the House Committee on Appropriations featured the predictive validity of the AFQT against the hands-on performance tests. The conclusion drawn from the earlier work to validate the ASVAB against job performance is that the ASVAB does predict job performance in the variety of occupational specialties, and therefore the ASVAB provides a sound basis for making personnel decisions.

APPENDIX A

TWENTY-FIRST ANNUAL REPORT
QUALITATIVE DISTRIBUTION OF MILITARY MANPOWER PROGRAM

FISCAL YEAR 1972

RCS - DD - M (A) 664
HEADQUARTERS
UNITED STATES ARMY RECRUITING COMMAND
HAMPTON, VIRGINIA 23669

HISTORICAL NOTES

1. Several months after the beginning of the Korean conflict in 1950, it became evident that there was a disparity in the quality of the accessions among the four services, and, as a result of this, on 2 April 1951 the Secretary of Defense forwarded a memorandum to the Secretaries of the Army, Navy, and Air Force stating that the following policies would become effective 1 May 1951 in order to assure an equitable distribution of military manpower.

   a. Voluntary enlistments be continued.

   b. Identical mental and physical standards for acceptability be provided for both enlistments and inductions.

   c. Qualitative distribution be maintained by quota control.

   d. Armed Forces Examining Stations be established to carry out the qualitative distribution program.

2. Beginning in May 1951 and continuing until the establishment of the Armed Forces Examining Stations, the Navy, the Marine Corps, and the Army-Air Force examined and reported their own chargeable accessions separately. Chargeable accessions were defined as: Men, 17 years of age or older, who had not previously served in any of the Armed Forces or who had served on active duty in any of the Armed Forces for a period of less than six months since 16 September 1940 and who became members of the Armed Forces through either enlistment or induction. Aviation Cadets, Officer Candidates, and Reservists were excepted.

3. On 30 June 1951, in order to conform to the provisions of the Universal Military Training and Service Act of 1951, the Department of Defense issued a directive reducing the minimum acceptance standards for the thirteenth to the tenth percentile, based on the Armed Forces Qualification Test (AFQT). On 2 July 1951, Department of Defense issued another directive which:

a. Established the Armed Forces Examining Stations Policy Board in the Office of the Assistant Secretary of Defense (Manpower and Personnel).

b. Designated the Army as executive agent for Armed Forces Examining Stations.

c. Fixed responsibility for fiscal matters with Comptroller, Office of the Assistant Secretary of Defense.

d. Designated commanding officers of individual Armed Forces Examining Stations.

e. Included the Marine Corps as a full partner in the program.

f. Established 1 September 1951 as the beginning date of the program.

4. In November 1951, tables of distribution based on planned chargeable accessions were established for the Armed Forces Examining Stations and initial staffing was authorized for each service on the following basis: 50% for the Army, 30% for the Air Force, 15% for the Navy, and 5% for the Marine Corps. Effective 16 July 1954, the percentages were changed to provide for staffing by the four services on an actual workload basis.

5. The high mental rejection rate of registrants throughout the nation made it necessary for the Department of the Army to take the following actions in order to cope with the problem.

a. As of 1 January 1952, a personnel psychologist was assigned to each Armed Forces Examining Station for the purpose of maintaining standardized and uniform mental testing procedures. Procedures were developed whereby certain categories of registrants could be administratively accepted notwithstanding the fact that they failed to achieve a passing score on the AFQT.

b. Effective 14 June 1957, procedures for categorization of Administrative Acceptances were further revised to restrict administrative acceptance to registrants determined to have deliberately failed the AFQT examination.

6. The Examen Calificacion de Fuerzas Armadas, ECFA-1 (Puerto Rican Screening test), became operational in the Puerto Rican Examining Station on 1 October 1953. This was replaced by ECFA-2 and ECFA-3 on 6 January 1959. These screening tests in the Spanish language include terms and phraseology peculiar to the Puerto Rican population.

7. In July 1958, the Universal Military Training and Service Act was amended by Public Law 85-564 (85th Congress) permitting the President to modify the minimum requirements of acceptance except in time of war or national emergency. The minimum mental standards were changed by Executive Order No. 101776 (28 July 1958) and DA messages by requiring all Mental Group IV (AFQT) registrants to be further examined by the Army Qualification Battery (AQB) at AFES. The minimum score was set at 90 or higher in any two aptitude areas. The AQB, not given to registrants examined in Puerto Rico, is intended to screen out those individuals who do not have the necessary basic aptitudes to undergo military training, thus reducing the number who might later be separated as inept or unsuitable.

8. As of 1 July 1960, the US Army Recruiting District Headquarters were authorized a personnel

psychologist, an enlisted assistant personnel psychologist, and a statistical clerk for the purpose of aiding in the administration of the Armed Forces examining and induction activities at US Army Recruiting Main Stations.

9. Beginning 31 July 1961, the processing of Cuban volunteers into the US Armed Forces was implemented.

10. On 1 December 1961, the AFES Policy Board was discontinued.

11. Responsibility for operation and control of Armed forces Examining Stations was delegated by Headquarters, Department of the Army, to Commanding General, United States Continental Army Command effective 1 July 1962.

12. Effective 1 May 1963, the minimum mental standard for induction into the Armed Forces was raised to require a minimum GT score of 80, plus two scores of 90 or higher in any of the other AQB aptitude areas.

13. Effective 1 October 1964, the United States Army Recruiting Command (USAREC) was established with responsibility for operation and control of the Armed Forces Examining and Induction Stations and all recruiting activities within the Continental United States.

14. On 1 July 1965, USAREC was reorganized by consolidating First and Second Recruiting Districts and changing the Armed Forces Examining Stations and Armed Forces Induction stations to Armed Forces Examining and Entrance Stations.

15. Further revision of mental standards for enlistment and induction was accomplished in November 1965 and April 1966. Minimum induction standards required an AFQT score of 16-30 with an additional requirement for non-high school graduates of GT score of 80 plus two AQB aptitude area scores of 90.

16. On 1 July 1966, USAREC was designated a Class II activity under DCSPER-DA.

17. Effective 1 October 1966, the Secretary of Defense initiated Project 100,000. The objectives of this program were:

    a. To give men who were previously ineligible, an opportunity to serve in Armed Forces, providing they met "New Standards" or qualified under the Medically Remedial Enlistment Program (MREP).

    b. To train men as effective soldiers without degrading military missions or effectiveness.

    c. To give these men an opportunity to learn skills, aptitudes and habits which could be taken back to civilian life to assist them in becoming productive citizens.

18. In February 1966, with the establishment of a second AFEES in the metropolitan New York City area, the total number of AFEES rose to 74.

19. In October 1966, mental standards for induction were revised, establishing the general requirement for an AFQT score of 31 or higher, but permitting an AFQT of 16-30 for high

school graduates or an AFQT of 10-30 and two AQB scores of at least 90 for non-high school graduates.

20. In December 1966, mental standards for induction were further revised requiring an AFQT score of 31 or above, AFQT 10-30 for high school graduates, AFQT 16-30 plus one 90 on the AQB, and AFQT 10-15 plus two 90's on the AQB for non-high-school graduates. (Note that these standards were obtained from the Fiscal Year 1968 report.)

21. On 1 July 1967, USAREC placed the Recruiting Main Stations (RMS) and AFEES in Alaska and Hawaii under the control of the Sixth Recruiting District. The Third Recruiting District assumed control of the AFEES in Puerto Rico.

22. Effective in September 1967, the minimum mental standard for enlistment was lowered to require a minimum AFQT score of 10-15 for high school graduates and an AFQT score of 10-15 with 2 AQB scores of at least 90 for non-high school graduates.

23. Effective 18 March 1969, with the revision of AR 601-270, the definition for chargeable accessions was changed. Local interpretation was authorized to facilitate reporting and data compilation; therefore, the standard DoD definition included in chapter 2, section III, paragraph 2-12d, AR 601-270 is interpreted as follows: Chargeable accessions are (1) those inductions into the active forces of any of the armed services and (2) those male enlistments into the regular components of any of the Armed Forces of the United States from civilian life who have not previously served on active duty or active duty for training in any of the Armed Forces, or who have served on active duty in any of the Armed forces for a period of less than six months. Aviation Cadets, Officer Candidates, and US Military Academy Preparatory School Cadets are excluded.

24. By Executive Order 11497 and proclamation 3945 of 26 November 1969, the President directed the establishment of a system for a random selection of all registrants who, prior to 1 January 1970, were between 19 and 25 years of age. The sequence would be drawn by lot and in turn would establish the order of susceptibility to induction. The purpose of the program was to reduce the draft vulnerability period of young men to one year and provide for induction by a lottery system.

25. During October 1971, the Chief of Staff of the Army announced a program whereby the draft could be abolished in support of a volunteer Army. This program, known as the Modern Volunteer Army, resulted in the following:

    a. During November 1970, the recruiting force was increased by approximately 500.

    b. In March 1971 approval was given USAREC to procure an additional 3,000 recruiters and establish 500 additional recruiting stations during FY 72.

    c. Additional funds were provided the Advertising and Information Directorate of USAREC to publicize and promote the Army through the mass media. A new recruiting slogan, "Today's Army wants to join you," was adopted to keep in line with the Army's self-examination in light of modern times.

    d. Certain enlistment options were initiated or modified to include:

(1) Choice of enlistment unit within CONUS and certain overseas locations in the Combat Arms Branches.

(2) The Buddy Plan whereby an unlimited number of men from a geographical area could enlist and receive basic training together.

26. Effective for the month of January 1971, the report, Summary of Registrant Examinations for Induction (RCS DCSPER 321), was extracted from mechanically-prepared reports rather than the manually-prepared reports used previously.

27. Beginning in FY 72, the Qualitative Distribution Report of Male Enlistments, Inductions, and Rejections (RCS DD-M(M) 663), the principal source of data for the statistical analyses contained in this document, was extracted from mechanically-prepared reports in lieu of the previously used manually-prepared reports.

28. During March 1972, the US Army Selection Center was established at Fort Jackson, SC, as a test vehicle to determine the impact on Army enlistments of exposing prospective enlistees to a sample of actual Army life before they executed enlistment contracts.

29. In order to encourage a greater number of longer term Army enlistments in the combat arms (Infantry, Armor and Artillery) career fields, Department of the Army implemented an enlistment option plan in June 1972 which pays a cash bonus to any individual who initially enlists for combat arms duty for a period of at least four years. This plan was favorably received on implementation and increases the attractiveness of the "non-glamour" branches of the Army.

Summary

The data presented in this report relate to chargeable accessions procured by the military services through the seventy-four Armed Forces examining and entrance stations (AFEES) from 1 July 1971 through 30 June 1972 (FY 1972). Chargeable accessions, however, do not account for the total manpower procurement of any service.

Of the 397,581 chargeable accessions (enlistments and inductions) procured for the Armed Forces during the fiscal year, 44% were Army, 22% were Navy, 20% were Air Force and 14% were Marine Corps.

While the Army utilized the Selective Service System to supplement intake, the Navy, Marine Corps and Air Force inducted only those reservists who failed to fulfill their reserve commitments. Army inductions represented 14% of the total accessions for that service, as compared to 51% for the previous fiscal year.

Total chargeable accessions for FY 1972 reflect a decrease of 24% from FY 1971. The Navy increased input by 13% and the Marine Corps by less than one percent, whereas the Army and Air Force decreased input by 41% and 11%, respectively.

APPENDIX B

LINEAGE OF THE AFQT AND THE ASVAB

Armed Forces Qualification Test (AFQT)

Form	Dates	Uses	Tests		
			Name	Items	Time
1	Jan 1950-	Induction	Verbal	30	
2	Dec 1952	and	Arithmetic Reasoning	30	
		enlistment	Space Perception	30	
			Total	90	45
3	Jan 1953-	same	Verbal	25	
4	Jul 1956		Arithmetic Reasoning	25	
			Space Perception	25	
			Tool Knowledge	25	
			Total	100	50
5	Aug 1956-	same	Same as forms 3 and 4		
6	Jun 1960				
7	Jul 1960-	same	Same as forms 3 and 4		
8	May 1973		AFQT 7/8 was replaced as follows:		
	May 1973	Army	(then obtain AFQT score from ACB 73)		
	1974	Navy	(then obtain AFQT score from NBTB)		
	Aug 1973	Air Force	(then obtain AFQT score from ASVAB 3)		
	Jun 1974	Marine Corps	(then obtain AFQT score from ASVAB 3)		

(Since January 1976, AFQT scores have been obtained from the ASVAB.)

Armed Forces Qualification Test (AFQT) Using the ASVAB

ASVAB Form	Dates	Uses	Tests		
			Name	Items	Time
5	Jul 1976-	STP and	Word Knowledge	30	10
	Jun 1984	joint-	Arithmetic Reasoning	20	20
		service	Space Perception	20	12
		enlistment	Total	70	42
		(Sum of raw scores converted to percentile scores)			
6	Jan 1976-	Joint-	Same as form 5		
7	Sep 1980	service			
		enlistment			
8	Oct 1980-	Joint-	Word Knowledge	35	11
9	Sep 1984	service	Paragraph Comprehension	15	13
10		enlistment	Arithmetic Reasoning	30	36
			Numerical Operations	50	3
			(1/2 of Numerical Operations score added)		
			Total	130	63
		(Sum of raw scores converted to percentile scores)			
11	Oct 1984-	Joint-	Same as forms 8, 9, and 10		
12	Dec 1988	service			
13		enlistment			
14	Jul 1984-	STP and	Same as forms 8, 9, and 10		
	Dec 1988	joint-			
		service			
		enlistment			
	Jan 1989-	same	Word Knowledge	35	11
	Jun 1992		Paragraph Comprehension	15	13
			Arithmetic Reasoning	30	36
			Mathematics Knowledge	25	24
			Total	105	84

(Sum of standard scores [2 times Word Knowledge and Paragraph Comprehension scores] converted to percentile scores)

ASVAB Form	Dates	Uses	Tests Name	Items	Time
15 16 17	Jan 1989	Joint-service enlistment	Same as form 14 in Jan 1989		
18 19	Jul 1992	STP and joint-service enlistment	Same as form 14 in Jan 1989		

Armed Services Vocational Aptitude Battery (ASVAB)

ASVAB Form	Dates	Uses	Tests Name	Items	Time
1	Sep 1968-Dec 1972	STP only	Word Knowledge	25	10
			Arithmetic Reasoning	25	25
			Tool Knowledge	25	10
			Space Perception	25	15
			Mechanical Comprehension	25	15
			Shop Information	25	10
			Automotive Information	25	10
			Electronics Information	25	10
			Coding Speed	100	7
2	Jan 1973-Jun 1976	STP 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		

ASVAB Form	Dates	Uses	Tests Name	Items	Time
4	not used				
5	Jul 1976-Jun 1984	STP and enlistment	Word Knowledge	30	10
			Arithmetic Reasoning	20	20
			Mathematics Knowledge	20	20
			Numerical Operations	50	3
			Attention to Detail	30	5
			General Science	20	10
			General Information	15	7
			Space Perception	20	12
			Mechanical Comprehension	20	15
			Shop Information	20	8
			Automotive Information	20	10
			Electronics Information	30	15
6 7	Jan 1976-Sep 1980	Joint-service enlistment	Same as form 5, plus: Army Classification Inventory:	20	
			Mechanical Interest	20	
			Electronics Interest	20	
			Combat Interest	27	
			Attentiveness Interest	20	
8 9 10	Oct 1980-Sep 1984	Joint-service enlistment	Word Knowledge	35	11
			Paragraph Comprehension	15	13
			Arithmetic Reasoning	30	36
			Mathematics Knowledge	25	24
			Auto/Shop Information	25	11
			Mechanical Comprehension	25	19
			Electronics Information	20	9
			Numerical Operations	50	3
			Coding Speed	84	7
			General Science	25	11
11 12 13	Oct 1984-Dec 1988	Joint-service enlistment	Same as forms 8, 9, and 10		

ASVAB Form	Dates	Uses	Tests Name
14	Jul 1984-Jun 1992	STP and enlistment	Same as forms 8, 9, and 10
15 16 17	Jan 1989	Joint-service enlistment	Same as forms 8, 9, and 10
18 19	Jul 1992	STP and enlistment	Same as forms 8, 9, and 10

APPENDIX C

DEVELOPMENT OF THE ASVAB 5/6/7

The Services had spent decades developing their own classification batteries, and over the years Service manpower managers had become comfortable with the effectiveness of their tests. By the mid 1970s, when the ASVAB 5/6/7 was formulated, each Service had developed a set of aptitude composites that was uniquely suited to its needs:

- the Marine Corps had used the Army Classification Battery (ACB) until 1974; then it switched to a form of the ASVAB developed for the Student Testing Program (form 3);
- the Air Force had switched in 1973 from its classification battery, The Airman Classification Test (ACT), to the ASVAB 3;
- the Navy continued to use its battery, the Navy Basic Test Battery (NBTB); and
- the Army continued with the ACB.

The ASVAB Working Group spent considerable time in 1974 deciding on the content of the ASVAB 5/6/7. The rule adopted by the Working Group was that the tests in the ASVAB 5/6/7 should allow each Service to compute the aptitude composites it was then using. In the early 1970s, testing time for classification batteries was about three hours, and the ASVAB 5/6/7 was designed to be about this length. These decisions led to test content that was almost identical to the version of the ACB introduced by the Army in May 1973, called ACB 73.

The classification batteries used by the other Services had similar content to each other, whereas the ACB 73 had unique content in addition to that shared with the other Service batteries. The first forms of the ASVAB, introduced in 1968 for testing high school students, contained the test content that was common to the Service batteries then in use. Form 3 of the ASVAB was similar to the first ASVAB, which meant that all Services except the Army could compute their aptitude composites from the common content. Because the ACB 73 had unique content, the content of the ASVAB 5/6/7 had to be expanded over the common content to accommodate the Army aptitude composites. The tests in the ASVAB 5/6/7, the ACB 73, and the ASVAB 3 were as shown on the next page.

The ASVAB 5/6/7 and the ACB 73 contained the same tests, although the labels may have been different. Because the total testing time was constrained to be the same length, about three hours, the two batteries also contained the same number of items in each test. Generally each test had 20 items, except for the two speeded tests: (a) Attention to Detail, which contained 60 items, and (b) Numerical Operations, which had 50 items. The content of the ASVAB 3 was contained in ASVAB 5/6/7 except for Coding Speed, which was replaced by Attention to Detail and Numerical Operations, and Tool Knowledge, which was part of the separate AFQT but was dropped from the AFQT in the early 1970s and from the ASVAB 5/6/7.

The writing of the items for the ASVAB 6/7, the tryout of the items, and the printing of the test booklets was the responsibility of the Air Force as executive agent for research and development of the ASVAB. Rather than use existing items, the Air Force decided to start with completely new items, which meant trying them out and printing test booklets. All this work was done in about one year.

The tryout scores for the items to be scaled were sent to the ASVAB Working Group for review in the summer of 1975. The Army representatives objected to the level of item difficulty in most subtests and insisted on adding more easy items that people in the AFQT categories IV and V would be likely to get right. (This policy of using easy items was consistent with traditional practices for the separate AFQT which contained many easy items.) Because of time pressures there was no time to reprint test booklets, so the easy items provided by the Army were literally pasted onto the end of the existing tests in the booklets. These easy items were retained in the final forms of ASVAB 5/6/7.

For more information, see the *History of the Armed Services Vocational Aptitude Battery (ASVAB) 1974-1980, a Report to the Principal Deputy Assistant Secretary of Defense (FM&P)* (ASVAB Working Group, Mar 1980), a report that reviews the development of the ASVAB from 1974 until 1980.

Tests in the ASVAB 5/6/7, the ACB 73, and the ASVAB 3

Test Name	ASVAB 5/6/7	ACB 73	ASVAB 3
Word Knowledge (WK)	WK	WK	WK
Arithmetic Reasoning (AR)	AR	AR	AR
Space Perception (SP)	SP	SP	SP
Mechanical Comprehension (MC)	MC	MC	MC
Automotive Information (AI)	AI	AI	AI
Shop Information (SI)	SI	SI	SI
Electronics Information (EI)	EI	EI	EI
General Science (GS)	GS	GS	-
General Information (GI)	GI	GI	-
Classification Inventory (CI)	CI	CI	-
Attention to Detail (AD)	AD	AD	-
Mathematics Knowledge (MK)	MK	MK	-
Coding Speed (CS)	-	-	CS
Numerical Operations (NO)	NO	-	-
Tool Knowledge (TK)	-	-	TK

REFERENCES

Army Research Institute. (1942, November). *Report on standardization of WAAC specialist tests.* (ARI Report 392). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1945, August). *Completion of the General Classification Test 1a (Spanish version).* (ARI Report 646). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1947, January). *Classification tests for reception processing (provisional program).* (ARI Report 696). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1949, April). *Comparison of army and navy classification tests.* (ARI Research Report 778). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1949, October). *Classification tests at army training centers.* (ARI Report 798). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1949, October). *Development of aptitude areas for classifying enlisted personnel in the army.* (ARI Report 808). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1951, December). *The detection of malingering* (unclassified title). (ARI Research Report 947). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Army Research Institute. (1952, December). *Test for negatively motivated pre-inductees* (for official use only). (ARI Research Memorandum 53-1). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

ASVAB Working Group. (1980, March). *History of the Armed Services Vocational Aptitude Battery (ASVAB) 1974-1980, a report to the Principal Deputy Assistant Secretary of Defense (FM&P).* Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

ASVAB Working Group. (1986, May). *A review of the development and implementation of the Armed Services Vocational Aptitude Battery, forms 11, 12, 13.* Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Bayroff, A. G. (1963, May). *Successive AFQT forms--comparisons and evaluations.* (ARI Technical Research Note 132). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Bayroff, A. G., & Anderson A. A. (1963, May). *Development of the Armed Forces Qualification Test 7 and 8.* (ARI Research Report 1132). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Bayroff, A. G., & Fuchs, E. F. (1970, February). *The Armed Services Vocational Aptitude Battery.* (ARI Report 1161). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Bock, R. D., & Mislevy, R. J. (1981). *Data quality analysis of the Armed Services Vocational Aptitude Battery.* Washington, DC: National Opinion Research Center.

Boldt, R. F. (1964, April). *Development of an optimum computerized allocation system.* (ARI Report 1135). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Boldt, R. F. (1980, August). *Scaling of the Armed Services Vocational Aptitude Battery form 7 and General Classification Test to the Armed Forces Qualification Test scale.* (OASD-FM&P-AP Technical Memorandum 80-2). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Booth-Kewly, S., Foley, P. P., & Swanson, L. (1984). *Predictive validity of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10 against 100 Navy schools.* (NPRDC-TR-88-15.) San Diego, CA: Navy Personnel Research and Development Center.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37*, 65-76.

Brogden, H. E. (1955). Least squares estimates and optimal classification. *Psychomretrika, 20*, 249-252.

Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19*, 181-190.

Cronbach, L. J. (1979, January). The Armed Services Vocational Aptitude Battery - a test battery in transition. *Personnel and Guidance Journal, 57*, 232-237.

Eitlelberg, M. J., Laurence J. H., & Waters, B. K., & Perelman, L. S. (1984). *Screening for service: aptitude and education criteria for military entry*. Alexandria, VA: Human Resources Research Organization.

Fairbank, B. A., Welsh, J. R., & Sawin, L. L. (1990, September). *Armed Services Vocational Aptitude Battery (ASVAB): validity of ASVAB form 14 for the prediction of high school course grades*. (AFHRL-TR-90-48). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Fischl, M. A., Ross, R. M., & McBride, J. R. (1979, April). *Development of factorially based ASVAB high school composites*. (ARI Technical Paper 360). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Fischl, M. A., & Ross, R. M. (1980, April). *Enhancing quality control in the testing of military applicants*. (ARI Technical Paper 384). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Holland, J. L. (1985). *Self-directed search assessment booklet*. Lutz, FL: Psychological Assessment Resources, Incs.

Holmgren, R.L. & Dalldorf, M.R. (1993). *A validation of the ASVAB against supervisors' ratings in civilian occupations*. Palo Alto, CA: American Institutes for Research.

Hunter, J. E., Crosson, J. J., & Friedman, D. H. (1985). *The validity of the Armed Services Vocational Aptitude Battery (ASVAB) for civilian and military job performance*. Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Jensen, H. E., & Valentine, L. D. (1976, March). *Validation of ASVAB 2 against civilian vocational-technical high school criteria*. (AFHRL TR-76-16). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Jensen, A. R. (1985, April). Test reviews: Armed Services Vocational Aptitude Battery. *Measurement and Evaluation in Counseling and Guidance* (pp. 323-37).

Kettner, N. (1976, October). *Armed Services Vocational Aptitude Battery (ASVAB form 5): comparison with GATB and DAT tests*. (AFHRL TR 76-78). Lackland, AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Laurence, J. H., & Ramsberger, P. F. (1991). *Low aptitude men in the military*. New York: Praeger.

Maier, M. H., & Fuchs, E. F. (1972, September). *Development and evaluation of a new ACB and aptitude area system*. (ARI Technical Research Note 239). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Maier, M. H., & Grafton, F. C. (1980, August). *Renorming ASVAB 6/7 at armed forces examining and entrance stations*. (OASD-FM&P-AP Technical Memorandum 80-1). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Maier, M. H., & Truss, A. R. (1983, March). *Original scaling of ASVAB forms 5/6/7: what went wrong*. (CNA CRC 457). Alexandria, VA: Center for Naval Analyses.

Maier, M. H., & Sims, W. H. (1986, July). *The ASVAB score scales: 1980 and World War II*. (CNA CNR 116). Alexandria, VA: Center for Naval Analyses.

Morton, M. A., Houston, T. J., Mundy, J. P., & Bayroff, A. G. (1957, May). *Mental screening tests for women in the armed forces.* (ARI Report 1103). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62(5),* 529-540.

Seeley, L. C., Fischl, M. A., & Hicks, J. M. (1978, February). *Development of the Armed Services Vocational Aptitude Battery (ASVAB) forms 2 and 3.* (ARI Technical Paper 289). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Sims, W. H. (1978, April). *An analysis of the normalization and verification of the Armed Services Vocational Aptitude Battery (ASVAB) forms 6 and 7.* (CNA Study CNS 1115). Alexandria, VA: Center for Naval Analyses.

Sims, W. H., & Truss, A. R. (1980, April). *A re-examination of the normalization of the Armed Services Vocational Aptitude Battery (ASVAB) forms 6, 7, 6E, and 7E.* (CNA Study 1152). Alexandria, VA: Center for Naval Analyses.

Sims, W. H., & Hiatt, C. M. (1981, February). *Validation of the Armed Services Vocational Aptitude Battery (ASVAB) forms 6 and 7 with applications to ASVAB forms 8, 9, and 10.* (CNA Study 1160). Alexandria, VA: Center for Naval Analyses.

Sims, W. H., & Maier M. H. (1983, June). *The appropriateness for military applications of the ASVAB subtests and score scale in the new 1980 reference population.* (CNA Memorandum 83-3102). Alexandria, VA: Center for Naval Analyses.

Staff. (1945). Personnel research section: the Army General Classification Test. *Psychological Bulletin, 42,* 760-768.

Staff. (1947). Personnel research section: the Army General Classification Test, with special reference to the construction and standardization of forms 1a and 1b. *Journal of Educational Psychology, 38,* 385-419.

Stuit, D. B. (Ed.) (1947). *Personnel research and test development in the bureau of naval personnel.* Princeton, NJ: Princeton University Press.

Uhlaner, J. E. (1952, November). *Development of the Armed Forces Qualification Test and predecessor army screening tests, 1946 - 1950.* (ARI Report 976). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

U.S. Army Recruiting Command. (1972). *Twenty-first annual report; qualitative distribution of military manpower program, fiscal year 1972.* Hampton, VA: Author.

U.S. Department of Defense. (1982, March). *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. (1986). *Military-civilian occupational crosswalk manual.* Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

U.S. Department of Defense. (1992a). *Exploring careers: the ASVAB workbook.* Chicago, IL: Military Entrance Processing Command.

U.S. Department of Defense. (1992b). *Military careers: a guide to military occupations and selected military career paths.* Chicago, IL: Military Entrance Processing Command.

U.S. Department of Labor. (1991). *Dictionary of occupational titles, vols. 1-11* (4th ed.). Washington, DC: Author

Wegner, T. G., & Ree, M. J. (1985, July). *Armed Services Vocational Aptitude Battery: correcting the speeded subtests for the 1980 youth population.* (AFHRL-TR-85-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Weeks, J. L., Mullins, C. J., & Vitola, B.M. (1975). *Airman classification batteries from 1948 to 1975: a review and evaluation.* (AFHRL-TR-75-78, AD-A026 470). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

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

Wigdor, A. K., & Green, B. F. (Eds). (1991). *Performance assessment for the workplace*. Washington DC: National Academy Press.

Willemin, L. P., & Karcher, E. K. (1958, January). *Development of combat aptitude areas*. (ARI TRR 1110). Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Wiskoff, M. F., Zimmerman, R. A., DuBois, D. A., Borman, W. C., & Park, R. K. (1991, December). *A process for revising the ASVAB*. Unpublished manuscript, Defense Manpower Data Center, Monterey, CA.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992, December). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composites*. Monterey, CA: Defense Manpower Data Center.

Zeidner, J., & Drucker, A. J., (1988). *Behavioral science in the army, a corporate history of the army research institute*. Alexandria, VA: Army Research Institute.